Multi-Scale Spectral Recurrent Network Based on Random Fourier Features for Wind Speed Forecasting

Eder Arley Leon-Gomez *, Víctor Elvira, Jorge Iván Montes-Monsalve, Andrés Marino Álvarez-Meza *, Alvaro Orozco-Gutierrez, German Castellanos-Dominguez

Posted Date: 19 March 2026

doi: 10.20944/preprints202603.1563.v1

Keywords: power systems; wind speed; neural networks; random fourier features

 doi:10.20944/preprints202603.1563.v1

Article

Multi-Scale Spectral Recurrent Network Based on Random Fourier Features for Wind Speed Forecasting

Eder Arley Leon-Gomez [1,*], Víctor Elvira [2], Jorge Iván Montes-Monsalve [3], Andrés Marino Álvarez-Meza [1,*], Alvaro Orozco-Gutierrez [4] and German Castellanos-Dominguez [1]

[1] AI-Lab, Signal Processing and Recognition Group, Universidad Nacional de Colombia, 170003 Manizales, Colombia
[2] School of Mathematics, University of Edinburgh, EH8 9YL Edinburgh, UK
[3] Dirección Académica, Universidad Nacional de Colombia, sede La Paz, 202010 La Paz, Colombia
[4] Automatics Research Group, Universidad Tecnológica de Pereira, 660003, Pereira
* Correspondence: ealeong@unal.edu.co (E.A.-L.G.); amalvarezme@unal.edu.co (A.M.Á.-M.)

Abstract

Accurate wind speed forecasting is critical for reliable wind-power integration, yet it remains challenging due to the strongly non-stationary and inherently multi-scale nature of atmospheric processes. While deep learning models—such as LSTM, GRU, and Transformer architectures—achieve competitive short- and medium-term performance, they frequently suffer from spectral bias, hyperparameter sensitivity, and reduced generalization under heterogeneous operating regimes. To address these limitations, we propose a multi-scale spectral–recurrent framework, termed RFF-RNN, which integrates multi-band Random Fourier Feature (RFF) encodings with parameterizable recurrent backbones. A key innovation of our approach is the deliberate relaxation of strict shift-invariance constraints; by jointly optimizing spectral frequencies, phase biases, and bandwidth scales alongside the neural weights, the framework dynamically shapes a fully data-driven spectral embedding. To ensure robust adaptation, we employ a two-stage optimization strategy combining gradient-based inner-loop learning with outer-loop Bayesian hyperparameter tuning. Extensive evaluations on a controlled synthetic benchmark and six geographically diverse real-world wind datasets (spanning the USA, China, and the Netherlands) demonstrate the superiority of the proposed framework. Statistical validation via the Friedman test confirms that RFF-enhanced models—particularly RFF-GRU and RFF-LSTM—systematically outperform standard recurrent networks and state-of-the-art Transformer architectures (Autoformer and FEDformer). The proposed approach yields significantly lower error metrics (MAE and RMSE) and higher explained variance (R^2), while exhibiting remarkable resilience against error accumulation at extended forecasting horizons.

Keywords: power systems; wind speed; neural networks; random fourier features

1. Introduction

The accelerating global shift toward renewable energy has positioned wind power as a cornerstone in decarbonization strategies. However, its large-scale integration into power systems is impeded by the inherent intermittency of atmospheric conditions and the nonlinear nature of wind-to-power conversion [1,2]. Accurate wind speed forecasting is thus indispensable for enabling reliable scheduling, grid stability, and cost-efficient market participation, as forecast errors can propagate across unit commitment, reserve allocation, curtailment decisions, and bidding strategies [3,4].

Wind speed prediction poses unique challenges. The underlying time series exhibit pronounced non-stationarity, multi-scale variability, and sensitivity to local topography and weather regimes. This variability stems from interacting mechanisms across diverse temporal scales—ranging from turbulence and boundary-layer instabilities to diurnal cycles and mesoscale phenomena [5]. As a result,

 doi:10.20944/preprints202603.1563.v1

forecast performance tends to degrade over longer horizons, particularly as models trained on short-term fluctuations struggle to generalize under broader climatological shifts and noise contamination [6]. Historically, physics-based numerical weather prediction (NWP) models were the primary tools for wind forecasting. While physically grounded and informative for large-scale systems, these models often require considerable computational resources and fail to capture site-specific variability, especially in complex terrains [7].

Then, the growing availability of high-resolution meteorological measurements and turbine-level operational data has significantly advanced the development of data-driven wind forecasting methodologies, offering computationally efficient and adaptable complements to physics-based models [8]. These approaches can be broadly categorized into statistical time-series models, shallow machine learning algorithms, and deep learning architectures [9]. Classical statistical methods, including autoregressive integrated moving average (ARIMA) models and exponential smoothing techniques, remain effective for capturing linear temporal structures and short-term autocorrelations [10]. However, their underlying linearity assumptions limit their capacity to model nonlinear dynamics, abrupt regime transitions, and multi-scale variability inherent to wind processes [11]. Shallow machine learning techniques—such as support vector regression (SVR), decision trees, and ensemble frameworks including gradient boosting—have demonstrated improved predictive performance under moderately nonlinear regimes. Nevertheless, these methods typically depend on manually engineered features and exhibit limited representational capacity when confronted with complex, long-range temporal dependencies or evolving statistical distributions [12]. As forecasting horizons extend and non-stationary effects intensify, the expressive limitations of these models become increasingly pronounced [13].

Deep learning has emerged as the leading paradigm for short-term wind speed forecasting due to its ability to learn complex nonlinear dynamics directly from data. In particular, recurrent neural networks (RNNs)—especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants—have demonstrated strong effectiveness in modeling temporal dependencies and are now widely adopted in operational forecasting pipelines [14]. To further improve representational capacity, hybrid architectures that integrate convolutional and recurrent components have been proposed, combining local pattern extraction with sequence modeling to better capture spatiotemporal structure [15]. More recently, transformer-based models have shown competitive and often superior performance in long-horizon settings by leveraging self-attention mechanisms that encode global temporal interactions more explicitly [16]. Despite their flexibility, deep models exhibit notable limitations, including sensitivity to hyperparameter selection, limited interpretability, and vulnerability to distributional shifts. Additionally, they are affected by spectral bias, as gradient-based training tends to emphasize low-frequency components, limiting the accurate representation of rapidly varying wind dynamics [17]. These constraints become particularly significant in medium- and long-horizon forecasting, where multi-scale patterns and evolving statistical regimes dominate [18].

To address these limitations, a complementary line of research has revisited kernel-based methods, which offer structured nonlinear representations and strong generalization, particularly in low-data regimes [19]. Kernel machines, including support vector machines and Gaussian processes, capture complex relationships through implicit high-dimensional feature mappings; however, their scalability is constrained by the computational burden associated with kernel matrix evaluations. Random Fourier Features (RFF), introduced by Rahimi and Recht [20], alleviate this limitation by approximating shift-invariant kernels (e.g., Gaussian or Laplacian) through explicit low-dimensional embeddings. This formulation enables frequency-aware representations while preserving computational efficiency, and RFF mappings have been successfully incorporated into neural forecasting frameworks with minimal additional overhead. In wind forecasting applications, RFF-based approaches have demonstrated increasing potential. By projecting input sequences into spectral feature spaces, RFF facilitates a more explicit encoding of oscillatory behavior and localized temporal structures [21]. Rayi et al. [22] proposed a deep mixed-kernel randomized architecture leveraging RFF, reporting improved predictive performance under non-stationary conditions. Furthermore, recent studies indicate that bandwise RFF

 doi:10.20944/preprints202603.1563.v1

decomposition—where signals are encoded across multiple frequency bands—enhances robustness and interpretability by disentangling temporal patterns between scales [23].

Besides, the stability and generalization of forecasting models critically depend on effective hyperparameter optimization. Neural architectures are inherently sensitive to design choices such as hidden dimensionality, dropout rates, and learning schedules, while RFF-based models introduce additional kernel-specific hyperparameters, including bandwidth and the number of Fourier features [24]. Exhaustive strategies, such as grid search, become infeasible in high-dimensional hyperparameter spaces, particularly when evaluating performance across multiple forecasting horizons [25]. Bayesian optimization—through probabilistic surrogate models such as Gaussian processes or Tree-structured Parzen Estimators (TPE)—has demonstrated strong effectiveness in tuning these systems, achieving improved predictive performance with substantially lower computational cost [26,27].

In this work, we introduce a multi-scale RFF recurrent network for wind speed forecasting, termed RFF-RNN, which integrates RFF encodings with parameterizable recurrent backbones. The proposed architecture decomposes wind signals across multiple frequency bands using Gaussian-based RFF mappings, concatenating the resulting spectral representations prior to recurrent processing via GRU, LSTM, or SRNN modules. This design enables explicit modeling of temporal dynamics at different scales, addressing the representational limitations of single-backbone architectures. Also, to ensure robust and adaptive performance, we adopt a two-stage optimization strategy that combines gradient-based training with Bayesian hyperparameter optimization. This joint tuning procedure simultaneously refines spectral configurations (bandwidths and number of features) and neural parameters (hidden dimensionality, dropout rate, and learning schedule), promoting stable generalization across heterogeneous sites and varying forecasting horizons. Hence, the main contributions of this work are summarized as follows:

- Multi-Scale RFF spectral encoding: a structured design that injects frequency-domain priors into forecasting models to enhance robustness under non-stationary conditions.
- Spectral–recurrent hybrid architecture: integration of Multi-Scale RFF encoders with recurrent backbones (SRNN/GRU/LSTM) for flexible modeling.
- Two-stage multiscale parameter optimization: gradient-based learning is coupled with Bayesian probabilistic search to jointly calibrate spectral bandwidths and neural hyperparameters, improving cross-horizon accuracy, robustness, and transferability while reducing manual tuning.

To validate the proposed framework, we designed experiments on both a controlled synthetic benchmark and multiple real-world wind speed datasets, evaluating short- and medium-horizon forecasting under heterogeneous operating conditions. We compared the proposed multi-scale spectral–recurrent variants against representative statistical, machine-learning, recurrent, and transformer-based baselines using error metrics, rank-based analyses, and significance tests. Across datasets and horizons, the MSRFF-based models consistently achieved more competitive performance, particularly in non-stationary regimes. These findings indicate that the proposed multiscale spectral design, combined with joint spectral–neural optimization, can improve forecasting reliability and practical transferability for operational wind-energy decision-making.

The remainder of this paper is organized as follows: Section 2 outlines the materials and methods. Section 3 presents the experimental setup, followed by results in Section 4. Conclusions and future research directions are provided in Section 5.

2. Materials and Methods

2.1. Synthetic Wind Speed Forecasting Datasets

To complement real-world wind observations and enable controlled evaluation under non-stationary conditions, we use a synthetic benchmark termed WindSynth++. The generator produces a

 doi:10.20944/preprints202603.1563.v1

wind-speed series with structured multi-scale seasonality, regime transitions, heteroskedastic noise, and a distribution shift in the test segment. Let $t = \{0, \ldots, T-1\}$, the synthetic series is defined as

$$y(t) = r(t)\, y_{\text{base}}(t) + \epsilon(t), \tag{1}$$

where

$$y_{\text{base}}(t) = y_{\text{diurnal}}(t) + y_{\text{sub}}(t) + y_{\text{weekly}}(t), \tag{2}$$

with

$$y_{\text{diurnal}}(t) = 2.5 + 1.5 \sin\left(\frac{2\pi t}{24} + 0.3 \sin\left(\frac{2\pi t}{168}\right)\right), \tag{3}$$

$$y_{\text{sub}}(t) = 0.6 \sin\left(\frac{2\pi t}{6} + 0.5\right) + 0.4 \sin\left(\frac{2\pi t}{3} + 1.3\right), \tag{4}$$

$$y_{\text{weekly}}(t) = 0.8 \sin\left(\frac{2\pi t}{168} + 0.2\right). \tag{5}$$

This decomposition is intentionally designed to impose multi-scale seasonality in the synthetic wind-speed process. The term $y_{\text{diurnal}}(t)$ captures the dominant 24-hour cycle observed in boundary-layer wind dynamics, while the nested weekly modulation inside its phase introduces slow temporal warping that mimics non-stationary day-to-day shifts. The component $y_{\text{sub}}(t)$ adds 6-hour and 3-hour harmonics to represent faster intra-day fluctuations and local variability. In turn, $y_{\text{weekly}}(t)$ encodes lower-frequency synoptic behavior over a 168-hour period. By superposing these components, the generator produces a signal with coupled short-, medium-, and long-scale temporal structure, which is critical for evaluating forecasting models under realistic multi-scale conditions.

More specifically, the fixed parameters to build the wind speed series can be read directly from the three definitions (see equations 2 to 5). In $y_{\text{diurnal}}(t)$, the constant offset is 2.5, the main amplitude is 1.5, and the base frequency is one cycle every 24 h (angular frequency $2\pi/24$); its phase is not constant, but modulated by $0.3 \sin(2\pi t/168)$ to induce slow weekly phase drift. In $y_{\text{sub}}(t)$, two fixed harmonics are superposed: a 6-hour term with amplitude 0.6 and phase offset 0.5, and a 3-hour term with amplitude 0.4 and phase offset 1.3. In $y_{\text{weekly}}(t)$, the amplitude is fixed at 0.8, the frequency corresponds to a 168-hour cycle (angular frequency $2\pi/168$), and the phase offset is fixed at 0.2. Therefore, amplitudes control oscillation magnitude, frequencies define the temporal scales, and phase terms set cycle alignment, jointly producing the intended multi-scale seasonal wind-speed pattern.

Now, the regime multiplier in Equation (1) is piecewise constant

$$r(t) = \begin{cases} 1.0, & 0 \le t < 0.35T, \\ 1.2, & 0.35T \le t < 0.65T, \\ 0.9, & 0.65T \le t < T, \end{cases} \tag{6}$$

and the heteroskedastic noise is defined as

$$\epsilon(t) = \left(0.15 + 0.15 \left|\sin\left(\frac{2\pi t}{24}\right)\right|\right)\eta(t), \quad \eta(t) \sim \mathcal{N}(0,1). \tag{7}$$

Of note, these terms emulate realistic non-stationary behavior in wind speed. The piecewise multiplier $r(t)$ in Equation (6) creates regime shifts in signal magnitude (normal, intensified, and reduced variability windows), forcing models to adapt across operating conditions instead of learning a single stationary pattern. Besdies, the heteroskedastic noise $\epsilon(t)$ in Equation (7) makes uncertainty time-dependent, with larger variance during phases of stronger diurnal activity, thereby reproducing the fact that forecast difficulty and error dispersion change over time.

 doi:10.20944/preprints202603.1563.v1

Lastly, a domain shift is introduced from $t \geq 0.8T$ by replacing the original diurnal component in Equation (2) with

$$y_{\text{diurnal}}^{\text{shift}}(t) = 2.5 + 1.8\sin\left(\frac{2\pi t}{24} + 0.4\sin\left(\frac{2\pi t}{168}\right)\right), \tag{8}$$

which increases diurnal amplitude and modifies phase modulation in the final segment.

For supervised learning, we construct an input–output dataset based on the simluated series $y(t)$ in Equation (1), yielding: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$, with $\mathbf{x}_n \in \mathbb{R}^{\tau}$ and $\mathbf{y}_n \in \mathbb{R}^{\tau'}$. Each input vector contains the previous τ wind-speed observations, while $\mathbf{y}_n$ stores the next τ' values corresponding to the forecasting horizon. Figure 1 provides an intuitive view of the synthetic sample construction. The component plot decomposes the signal into a dominant diurnal cycle (24 h), higher-frequency sub-daily harmonics (6 h and 3 h), and a slower weekly modulation (168 h), whose sum forms the baseline multi-scale wind pattern. The sample-window plot then shows how a forecasting instance is extracted, with a historical input window $\mathbf{x}_n$ and a future prediction horizon vector $\mathbf{y}_n$.

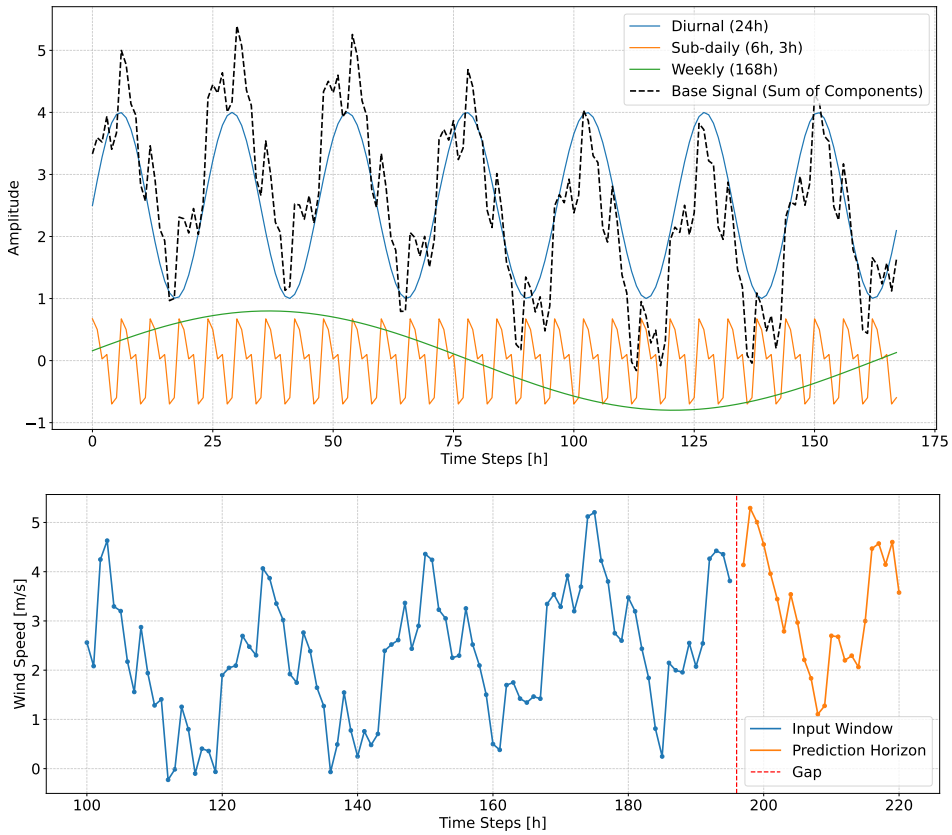


Figure 1. Illustrative example of a sample generated with the *WindSynth++* synthetic dataset for wind speed forecasting. Top: Decomposition of the simulated signal into its main components, including diurnal, sub-daily, weekly, and stochastic variability. Bottom: Sliding-window construction of forecasting samples, where historical observations form the input window used to predict future values within the forecasting horizon.

2.2. Real-World Wind Speed Forecasting Datasets

To assess generalization beyond controlled synthetic scenarios, we evaluate the proposed models on multiple real-world hourly wind speed time series collected across distinct geographic and climatic regimes. The selected datasets cover diverse levels of seasonality, intermittency, and non-stationarity. This diversity enables a rigorous comparison of forecasting performance under heterogeneous atmospheric dynamics and operational conditions. Following, each studied real-world dataset is described.

Argonne Weather Observatory - USA (Argone). The Argonne dataset was obtained from the Argonne Weather Observatory in Illinois, USA (41° N, 87° W). It constitutes one of the most extensive open-access records of hourly wind speed measurements, spanning from 1 January 1998

 doi:10.20944/preprints202603.1563.v1

to 30 August 2005. Observations were collected using high-precision instrumentation, including Met One anemometers and Vaisala HMP45A humidity/temperature sensors (see https://pubs.usgs.gov/wdr/2005/wdr-il-05/data/wind_por/indices0/index.htm, accessed on 1 December 2025). The site represents a continental climate with clear seasonal variability, pronounced diurnal cycles, and occasional extreme wind events. Owing to its long temporal coverage and measurement quality, this dataset serves as a strong reference benchmark for evaluating forecasting under moderate atmospheric variability.

Beijing Capital International Airport - China (Beijing). The Beijing dataset was recorded at a major climatological station near Beijing Capital International Airport (39.9° N, 116.2° E). It spans from 1 August 2011 to 30 December 2018 at hourly resolution (see https://talltowers.bsc.es, accessed on 1 December 2025). The region is characterized by a continental monsoon climate, with marked seasonal contrasts between cold, dry winters and hot, humid summers. These conditions induce substantial interannual and intraseasonal wind variability. In addition, urban effects such as thermal inversions and enhanced surface roughness introduce turbulence and irregular temporal patterns, making this dataset suitable for assessing model robustness under complex urban wind dynamics.

Chengdu Shuangliu International Airport - China (Chengdu). The Chengdu dataset was collected at Shuangliu International Airport in Sichuan Province, China (30.6° N, 104.01° E), covering 1 January 2011 to 30 December 2018 with hourly sampling (see https://mesonet.agron.iastate.edu/request/download.phtml?network=PK__ASOS, accessed on 1 December 2025). The station is located in the Sichuan Basin and is influenced by a subtropical monsoon climate combined with strong orographic effects. Persistent fog, high humidity, and terrain-induced flow disruptions result in comparatively low mean wind speeds, high volatility, and weak temporal regularity. Among the datasets considered, Chengdu exhibits the strongest irregularity and non-stationarity, making it a particularly challenging benchmark for forecasting models.

Royal Netherlands Meteorological Institute KNMI Stations - Netherlands. This study also uses publicly available wind speed records from the Royal Netherlands Meteorological Institute (KNMI), accessed via https://dataplatform.knmi.nl/ (accessed on 1 December 2025). Hourly data were collected from three representative stations: **Schiphol** (52.3081° N, 4.7642° E), **De Bilt** (52.1036° N, 5.1803° E), and **Leeuwarden** (53.2012° N, 5.7999° E). The series span from 1 January 2011 to 29 March 2020, yielding approximately 81,000 hourly samples per station. Because of the country's flat topography and maritime climate, these stations generally exhibit more stable wind regimes, with mean wind speeds typically between 4 and 7 m/s. Seasonal variability remains evident, with stronger winds usually occurring in winter, particularly at northern stations such as Leeuwarden.

Figure 2 illustrates the geographical distribution of the wind speed measurement sites considered in this study across North America, East Asia, and Western Europe. This spatial diversity enables the evaluation of forecasting performance under distinct meteorological regimes, ranging from relatively stable maritime conditions to highly variable continental and monsoon-influenced environments. Table 1 reports key descriptive statistics for each dataset, highlighting the heterogeneity of wind behaviors and motivating the need for forecasting models that generalize robustly across diverse climatic contexts.

Table 1. Statistical summary of wind speed real-world datasets (measured in m/s).

Dataset	Start date	End date	Max	Mean	Min	Median	Standard Deviation
Argonne	01/01/1998	30/08/2005	32.44	7.28	0	6.49	3.83
Chengdu	01/01/2011	30/12/2018	33.53	3.52	0	2.24	2.95
Beijing	01/08/2011	30/12/2018	40.23	6.48	0	4.47	4.85
Schiphol	01/01/2011	29/03/2020	23.00	4.75	0	4.79	2.66
De Bilt	01/01/2011	29/03/2020	14.0	3.21	0	3.21	1.76
Leeuwarden	01/01/2011	29/03/2020	21.0	4.07	0	4.08	2.22

 doi:10.20944/preprints202603.1563.v1

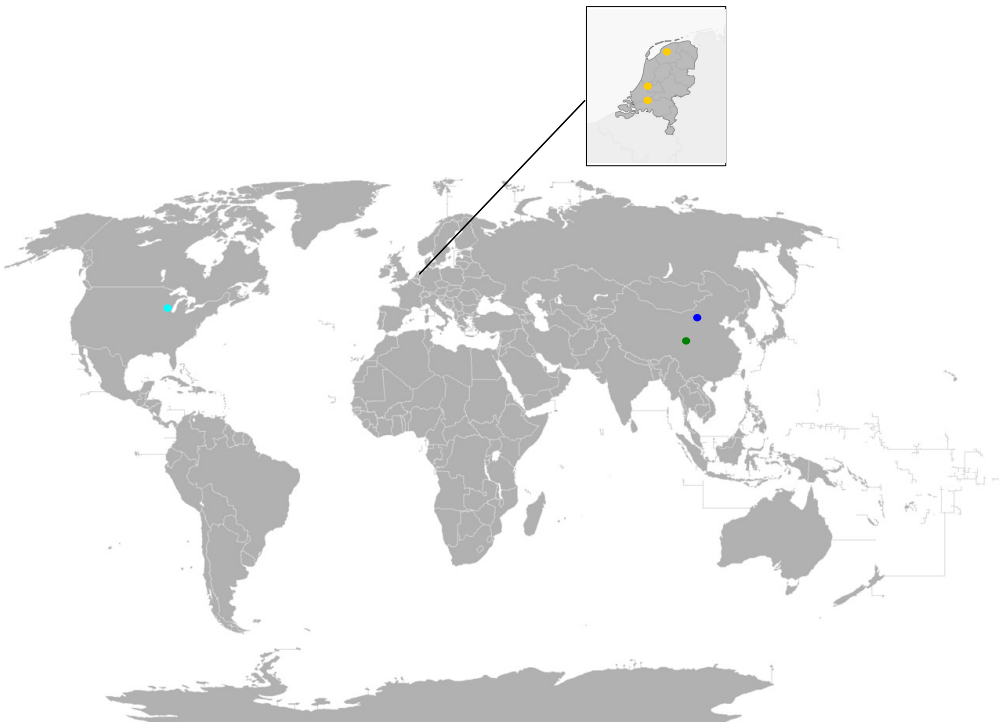


Figure 2. Geographic distribution of the wind speed measurement sites considered in this study. The map indicates the locations of the Argonne Weather Observatory (cyan dot •) in Illinois, USA; Chengdu Airport (green dot •) in Sichuan Province, China; Beijing Capital International Airport (blue dot •) in Beijing, China; the Netherlands stations: Schiphol, De Bilt, and Leeuwarden (yellow dot •) representing a maritime climate regime.

2.3. Benchmarking Frameworks for Wind Speed Forecasting

Let $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$ denote a collection of input–output pairs for wind speed forecasting, where each input $\mathbf{x}_n \in \mathbb{R}^\tau$ is a temporal segment of length τ extracted with a sliding window, and each target $\mathbf{y}_n \in \mathbb{R}^h$ contains the next h values (forecast horizon). A deep learning model aims to learn a forecasting function $\tilde{\varsigma}_\nu : \mathbb{R}^\tau \to \mathbb{R}^h$, parameterized by ν, that maps each historical segment to its future trajectory. Under direct multi-step forecasting, the complete horizon is predicted simultaneously as $\hat{\mathbf{y}} = \tilde{\varsigma}_\nu(\mathbf{x})$.

Here, we consider two representative paradigms to instantiate $\tilde{\varsigma}_\nu$: (i) Recurrent Neural Networks, which model temporal dependencies through evolving hidden states; and (ii) Transformer-based models, which use attention mechanisms to capture long-range dependencies across sequence positions.

Recurrent Neural Networks (RNN) [28]. Given an input sequence $\{\mathbf{x}_t \in \mathbb{R}^d\}_{t=1}^{\tau}$, $d = 1$ in the univariate setting, recurrent models encode temporal information through a hidden state that evolves over time. A simple recurrent update is

$$\mathbf{h}_t = \varphi(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_x \mathbf{x}_t + \mathbf{b}_h), \tag{9}$$

where $\mathbf{h}_t$ denotes the hidden state at time t, φ is a nonlinear activation function, and $\mathbf{W}_h$, $\mathbf{W}_x$, and $\mathbf{b}_h$ are learnable parameters.

Variants such as LSTM and GRU introduce gating mechanisms. For LSTM [29], the update equations are

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i), \tag{10}$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f), \tag{11}$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o), \tag{12}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c), \tag{13}$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{14}$$

 doi:10.20944/preprints202603.1563.v1

where $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ denote the input, forget, and output gates, respectively; $\sigma(\cdot)$ and $\tanh(\cdot)$ represent the sigmoid and hyperbolic tangent activation functions; and $\odot$ denotes the Hadamard product.

For GRU, gating is performed with update and reset mechanisms while using a single hidden state (without an explicit memory cell) [30]. The GRU updates are given by

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z), \tag{15}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r), \tag{16}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h\mathbf{x}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h), \tag{17}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \tag{18}$$

where $\mathbf{z}_t$ and $\mathbf{r}_t$ denote the update and reset gates, respectively, and $\tilde{\mathbf{h}}_t$ is the candidate hidden state. The final hidden state $\mathbf{h}_\tau$ is then mapped to the forecast horizon through a fully connected layer

$$\hat{\mathbf{y}} = \mathbf{W}_{\text{out}}\mathbf{h}_\tau + \mathbf{b}_{\text{out}}. \tag{19}$$

Transformers [31]. They replace recurrent state transitions with self-attention, enabling global context aggregation through pairwise interactions across all sequence positions. Specifically, for an input trajectory matrix $\mathbf{X} \in \mathbb{R}^{\tau \times d}$, the query, key, and value projections are defined as

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{20}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The scaled dot-product attention is then defined as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{21}$$

Among relevant transformer-based time-series forecasting approaches, Autoformer [32] performs an explicit decomposition of each input sequence into trend and seasonal components: $\mathbf{X} = \mathbf{X}_{\text{trend}} + \mathbf{X}_{\text{seasonal}}$. Attention is then applied primarily to $\mathbf{X}_{\text{seasonal}}$ through an autocorrelation-based mechanism, which strengthens periodic pattern extraction while reducing dependence on standard positional encodings.

Besides, Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting (FEDformer) [33] introduces a frequency-enhanced attention block. It first applies the Fourier Transform: $\mathcal{F}(\mathbf{X}) = \mathbf{F}$, where $\mathbf{F}$ captures the frequency-domain representation of the input. A learnable filter $\mathcal{G}$ selects relevant frequency components, and the attention mechanism operates in this transformed space, before applying an inverse Fourier transform $\mathcal{F}^{-1}$ to return to the time domain: $\mathbf{X}' = \mathcal{F}^{-1}(\mathcal{G}(\mathcal{F}(\mathbf{X})))$.

Across these variants, each model encodes the input sequence into a latent representation and applies either frequency-domain filtering or autocorrelation-based attention to infer multi-step dynamics. The forecast is then produced in a direct decoding manner, jointly predicting all h future time steps.

2.4. Multi-Scale Random Fourier Feature Encoder

Kernel methods provide powerful, non-linear representational capabilities by implicitly mapping data into a high-dimensional Reproducing Kernel Hilbert Space (RKHS). However, their exact computation scales poorly with dataset size [34]. RFF bypass this limitation by approximating shift-invariant kernels through explicit, low-dimensional feature maps. The theoretical foundation of RFF rests on Bochner's theorem, which states that any continuous, shift-invariant, and positive-definite kernel $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}') \in \mathbb{R}$ can be represented as the Fourier transform of a non-negative probability measure $p(\mathbf{w}) \in \mathbb{R}^+$ [35]:

$$\kappa(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{X}} p(\mathbf{w}) e^{i\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')} d\mathbf{w}, \tag{22}$$

 doi:10.20944/preprints202603.1563.v1

where $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. By drawing independent samples $\mathbf{w} \in \mathcal{X}$ from $p(\mathbf{w})$, the kernel can be approximated by the inner product of explicit randomized feature vectors.

In this work, we base our approximation on the Gaussian kernel:

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\rho^2}\right), \quad \rho > 0. \tag{23}$$

The Gaussian kernel in Equation (23) is selected because its corresponding probability measure $p(\mathbf{w})$ is also Gaussian, making sampling straightforward ($\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) [36]. More importantly, the Gaussian kernel acts as a universal approximator and produces infinitely smooth functions [37], which are highly appropriate for modeling the continuous, non-linear atmospheric variations characteristic of wind speed dynamics.

Now, to map the raw inputs into a multi-scale spectral domain, each input observation $\mathbf{x}_t \in \mathbb{R}^d$ is transformed through K distinct spectral bands. For a given band k, the RFF mapping is defined as:

$$\boldsymbol{\phi}_k(\mathbf{x}_t) = \sqrt{\frac{2}{N_f}} \cos\left(\text{softplus}(\rho_k)\mathbf{W}_k^\top \mathbf{x}_t + \mathbf{b}_k\right) \tag{24}$$

where $\mathbf{W}_k \in \mathbb{R}^{d \times N_f}$ is the frequency matrix initially sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{b}_k \in \mathbb{R}^{N_f}$ is the phase bias vector initially sampled from $\mathcal{U}(0, 2\pi)$, and $\rho_k \in \mathbb{R}$ is a learnable parameter controlling the spectral scale (effective bandwidth) of the kernel. Crucially, after initialization, $\mathbf{W}_k$ and $\mathbf{b}_k$ are not kept fixed. Instead, they are jointly optimized with the rest of the network through backpropagation. By fine-tuning these frequencies and phases, the model relaxes the strict shift-invariance constraint required by Bochner's theorem. Thus, while the representation is initialized via RFF principles to approximate a Gaussian kernel, it ultimately evolves into a fully data-driven spectral embedding. This relaxation grants the model the flexibility to learn highly localized and non-stationary spectral representations tailored specifically to the complex, evolving temporal patterns of wind speed dynamics.

Furthermore, the use of the softplus function—defined as $\ln(1 + e^{\rho_k})$—is a critical design choice. Because a valid kernel bandwidth must be strictly positive, applying softplus constrains the learnable parameter ρ_k to the positive domain while allowing the underlying weight to be updated freely via gradient descent. Unlike the ReLU function, which can yield "dead" zero gradients for negative inputs, softplus is continuously differentiable, ensuring stable gradient flow during the joint optimization of neural and spectral parameters [38].

Afterward, the full multi-band embedding for a time step t is the concatenation of all K band mappings:

$$\mathbf{z}_t = [\boldsymbol{\phi}_1(\mathbf{x}_t), \ldots, \boldsymbol{\phi}_K(\mathbf{x}_t)] \in \mathbb{R}^F, \qquad F = KN_f. \tag{25}$$

Then, stacking these embeddings across the input window yields the sequence-level representation:

$$\mathbf{Z} = \Phi_{\text{MB}}(\mathbf{X}) = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_\tau]^\top \in \mathbb{R}^{\tau \times F}. \tag{26}$$

Each individual mapping $\boldsymbol{\phi}_k$ approximates a base kernel $\kappa_k(\mathbf{x}, \mathbf{x}') \approx \boldsymbol{\phi}_k(\mathbf{x})^\top \boldsymbol{\phi}_k(\mathbf{x}')$ in N_f dimensions. By concatenating the K embeddings, the encoder implicitly constructs a composite kernel $k_{\text{MB}}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{K} k_k(\mathbf{x}, \mathbf{x}')$. This additive kernel structure yields a direct sum of their respective RKHS spaces ($\mathcal{H}_{\text{MB}} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_K$), granting the model a richer representational capacity. The trainable scales ρ_k allow the network to adaptively emphasize high-frequency components for localized turbulence and low-frequency components for synoptic seasonal patterns concerning the input time-series.

Then, the encoded representation in $\mathbf{Z}$ is further refined using a Temporal Spectral Block (TSB), defined as:

$$\tilde{\mathbf{Z}} = \text{LN}\Big(\mathbf{Z} + \breve{\sigma}\big(\text{Conv}_{\breve{d},\breve{k}_1}\big(\breve{\sigma}(\text{Conv}_{\breve{d},\breve{k}_0}(\mathbf{Z}))\big)\big)\Big) \tag{27}$$

 doi:10.20944/preprints202603.1563.v1

where $\mathrm{Conv}_{\check{d},\check{k}}$ is a dilated 1D convolution with dilation factor d and kernel size k, $\check{\sigma}$ is the GELU activation function, and LN is Layer Normalization. Dilated convolutions expand the receptive field to extract medium-range temporal correlations, serving as a temporal–spectral fusion mechanism without significantly increasing computational overhead.

Moreover, long-range temporal dependencies are subsequently modeled via a recurrent operator:

$$\check{\mathbf{H}} = [\check{\mathbf{h}}_1, \check{\mathbf{h}}_2, \ldots, \check{\mathbf{h}}_\tau] = R(\tilde{\mathbf{Z}}), \tag{28}$$

with hidden-state updates taking the general form:

$$\check{\mathbf{h}}_t = f_{\mathrm{RNN}}(\tilde{\mathbf{z}}_t, \check{\mathbf{h}}_{t-1}; \Theta_r) \tag{29}$$

Depending on the specific architectural variant, f_{RNN} in Equation (29) operates using simple RNN (SRNN), LSTM, or GRU formulations. The RNN module serves as a temporal integrator, resolving the sequential dynamics mapped by the multi-scale spectral decomposition.

Finally, a compact temporal pooling and output projection-based representation is extracted from the sequence of hidden states $\check{\mathbf{H}}$ via a pooling operation:

$$\mathbf{v} = \begin{cases} \check{\mathbf{h}}_\tau, & \text{last-state pooling,} \\ \frac{1}{\tau}\sum_{t=1}^{\tau} \check{\mathbf{h}}_t, & \text{mean pooling.} \end{cases} \tag{30}$$

The multi-horizon wind speed prediction is then generated through a two-layer Multi-Layer Perceptron (MLP):

$$\hat{\mathbf{y}} = \tilde{\mathbf{W}}_2\, \sigma(\tilde{\mathbf{W}}_1 \mathbf{v} + \tilde{\mathbf{b}}_1) + \tilde{\mathbf{b}}_2 \tag{31}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^h$ represents the complete forecasted horizon.

Figure 3 summarizes the proposed multi-scale RFF encoder, termed RFF-RNN.

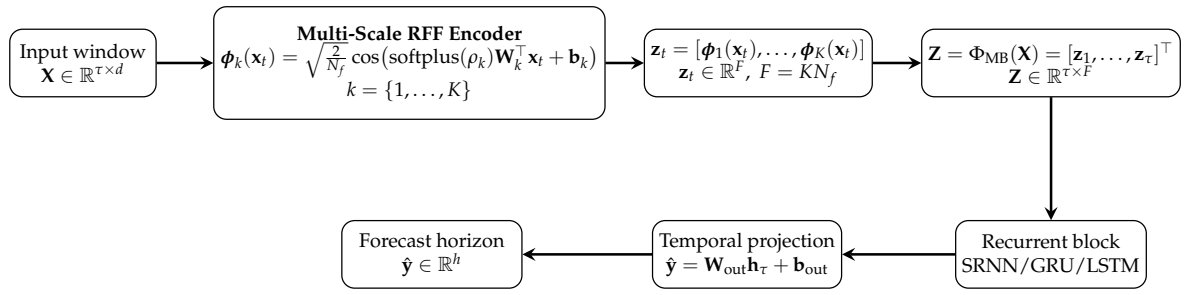


Figure 3. Multi-Scale Random Fourier Feature Encoder with recurrent-based representation (RFF-RNN). The input window $\mathbf{X}$ is mapped by K spectral bands into $\mathbf{z}_t$, stacked into $\mathbf{Z}$, processed by a recurrent backbone, and projected to multi-horizon predictions.

2.5. RNN-RFF Two-Stage Optimization

To ensure robust adaptation to non-stationary wind speed dynamics and to maximize the representational capacity of the spectral-recurrent architecture, the proposed RFF-RNN framework employs a two-stage optimization strategy. This approach decouples the gradient-based learning of the network's internal weights (inner loop) from the probabilistic search of architectural and spectral configurations (outer loop).

Stage 1: Inner-Loop Parameter Optimization. During the first stage, the network's internal parameters, denoted as Θ, are optimized end-to-end via gradient descent. Unlike traditional kernel-based learning where the spectral representation is kept fixed, the proposed framework defines Θ as

 doi:10.20944/preprints202603.1563.v1

the union of the recurrent backbone parameters, the projection head weights, and the Multi-Scale RFF parameters:

$$\Theta = \Theta_{\text{RNN}} \cup \{\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2\} \cup \{\mathbf{W}_k, \mathbf{b}_k, \rho_k\}_{k=1}^{K}. \tag{32}$$

By actively updating the frequency matrices $\mathbf{W}_k$, phase biases $\mathbf{b}_k$, and spectral bandwidths ρ_k alongside the neural weights, the model dynamically shapes its multi-scale embedding. This joint optimization allows the RFF encoder to shift from a strict mathematical approximation of a shift-invariant Gaussian kernel toward a highly adaptive, data-driven feature space.

In particular, the parameter set Θ is trained to minimize the Mean Squared Error (MSE) between the multi-step predictions $\hat{\mathbf{y}}$ and the ground truth horizon $\mathbf{y}$. For a batch of $\tilde{N}$ training samples and a prediction horizon of h steps, the loss function is defined as:

$$\mathcal{L}_{\text{MSE}}(\Theta) = \frac{1}{\tilde{N} \cdot h} \sum_{n=1}^{\tilde{N}} \sum_{j=1}^{h} \left(\hat{y}_{n,j}(\Theta) - y_{n,j}\right)^2. \tag{33}$$

Through backpropagation [28], the gradients flow seamlessly from the temporal projection head, through the recurrent module, and into the parameterized RFF encoder. The continuous differentiability of the softplus function ensures stable bandwidth (ρ_k) updates throughout the minimization of $\mathcal{L}_{\text{MSE}}(\Theta)$.

Stage 2: Outer-Loop Bayesian Hyperparameter Optimization. The generalization capability of the RFF-RNN heavily depends on a set of higher-level hyperparameter configurations, $\lambda \in \Lambda$, which govern both the neural topology and the spectral boundaries. This search space includes the number of Fourier features per band (N_f), multi-band frequency presets, recurrent hidden size (d_h), number of recurrent layers (L), spectral dropout rate (p), and specific learning-rate multipliers. Manual tuning of such a high-dimensional space Λ is computationally prohibitive and prone to sub-optimal local minima.

To address this, the second stage employs Bayesian Optimization (BO) via the Tree-structured Parzen Estimator (TPE) algorithm [39]. The objective is to find the optimal hyperparameter configuration λ^* that minimizes the validation loss over E training epochs:

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \min_{e \in [1,E]} \mathcal{L}_{\text{MSE}}^{\text{valid}}(\lambda, e). \tag{34}$$

The TPE algorithm probabilistically guides the search by modeling the conditional distribution $p(\lambda \mid y)$ based on past trial evaluations. It partitions evaluated configurations into *good* ($\mathcal{L} < \gamma$) and *bad* ($\mathcal{L} \geq \gamma$) groups, constructing separate kernel density estimators:

$$p(\lambda \mid y) = \begin{cases} \ell(\lambda) & \text{if } y < \gamma, \\ g(\lambda) & \text{if } y \geq \gamma. \end{cases} \tag{35}$$

New candidate configurations are sampled by maximizing the Expected Improvement (EI), defined proportionally as $\text{EI}(\lambda) \propto \ell(\lambda)/g(\lambda)$. This inherently balances exploration (sampling from regions with high $\ell(\lambda)$ density) and exploitation (avoiding regions with high $g(\lambda)$ density).

To further enhance computational efficiency during this outer loop, we incorporate a *Median Pruner*. A specific training trial operating under configuration λ is halted early at epoch e if its intermediate validation loss exceeds the median validation loss of all previously completed trials at that same epoch. This mechanism prevents the allocation of gradient-descent resources to unpromising hyperparameter configurations, enabling an exhaustive yet computationally tractable optimization of the RFF-RNN framework.

 doi:10.20944/preprints202603.1563.v1

3. Experimental Set-Up

To assess the predictive performance and generalization capability of the proposed RFF-RNN framework, this section details the experimental protocol adopted for both the synthetic and real-world wind speed datasets. The experimental design emphasizes temporal integrity, architectural fairness, and reproducible hyperparameter optimization across all model variants.

All datasets are chronologically partitioned into training (70%), validation (10%), and testing (20%) subsets to prevent information leakage and preserve the causal structure of the time series. The input window length τ and forecast horizon h are selected according to the statistical characteristics of each site. Stations such as Argonne, Beijing, and Chengdu, which exhibit highly non-stationary and intermittent wind dynamics, employ input windows of $\tau = 20$ hours with prediction horizons $h \in \{1, 2, \ldots, 7\}$. In contrast, the Dutch stations—characterized by smoother maritime wind regimes and stronger temporal coherence—use shorter input windows ($\tau = 10$) and extended horizons up to $h = 10$ hours. Prior to model training, all time series are normalized to the $[0, 1]$ range using Min-Max scaling to enhance numerical stability and convergence. Supervised samples are generated using a sliding-window strategy with unit stride, which maximizes data utilization while preserving strict temporal ordering. No explicit data augmentation is applied, ensuring that any performance gains are strictly attributable to the proposed spectral-temporal modeling.

Further, our RFF-RNN-based models are benchmarked against two distinct families of deep learning architectures. The first group consists of standard recurrent networks, including the Simple Recurrent Neural Network (SRNN), GRU, and LSTM approaches. Evaluating these baselines directly isolates the performance contribution of the Multi-Scale RFF encoder when paired with identical recurrent backbones. The second group comprises state-of-the-art Transformer-based architectures designed for time-series forecasting, specifically the standard Transformer, Autoformer, and FEDformer. These models rely on sophisticated self-attention and frequency-enhanced decomposition mechanisms, serving as robust, high-capacity benchmarks to evaluate whether the proposed RFF-RNN framework can competitively model long-range temporal dependencies and spectral biases without the massive computational overhead typical of explicit attention mechanisms.

To ensure a fair comparison and adequate architectural capacity, hyperparameter tuning is performed using Bayesian Optimization via the TPE algorithm. The objective function minimized during this search is the validation-set MSE. The hyperparameter search space Λ for the RFF-based models is defined as follows:

- Fourier features per band: $N_f \in \{16, 24, 32, 48, 56, 88, 104, 112\}$.
- Recurrent hidden size: $d_h \in \{32, 48, 64, 96, 104, 112, 144, 176, 240, 256\}$.
- Number of recurrent layers: $L \in \{1, 2, 3\}$.
- Spectral dropout rate: $p_{\text{spec}} \in [0.1, 0.5]$.
- Spectral learning-rate multiplier: $\eta_\rho \in [0.5, 4.0]$.

The search space bounds are established to balance representational capacity with computational efficiency. Specifically, the limits on N_f and d_h are constrained to prevent severe overfitting and unbounded memory scaling. Additionally, the multi-scale receptive field presets $\mathcal{B}$ are selected from the discrete set $\mathcal{B} \in \{(4, 24, 168), (6, 24, 72), (12, 24, 168)\}$; these are deterministically chosen to explicitly capture the known physical periodicities inherent to wind dynamics, such as sub-daily, 24-hour diurnal, and 168-hour weekly cycles. Each model configuration is trained for 100 epochs using the Adam optimizer with a batch size of 256 and an initial learning rate of 10^{-3}.

To keep a consistent and fair experimental setup, Transformer baselines (Transformer, Autoformer, and FEDformer) are tuned with Bayesian optimization using TPE in Optuna [40]. The search space is designed to balance expressiveness, computational efficiency, and robustness in real-world forecasting: $d_{\text{model}} \in \{64, 96, 128, 192, 256\}$ and $n_{\text{heads}} \in \{1, 2, 4, 8\}$ follow standard Transformer design under the constraint $d_{\text{model}} \bmod n_{\text{heads}} = 0$ [41]; Autoformer and FEDformer depths are limited to shallow ranges (1–3 layers), as deeper stacks are often unnecessary for structured temporal dependencies [32,33]. The corresponding search space Λ is defined as follows:

 doi:10.20944/preprints202603.1563.v1

- Model dimension: $d_{\text{model}} \in \{64, 96, 128, 192, 256\}$.
- Attention heads: $n_{\text{heads}} \in \{1, 2, 4, 8\}$, restricted by d_{model} mod nhead $= 0$.
- Dropout rate: $p_{\text{drop}} \in [0.05, 0.3]$.
- Transformer: encoder layers $L_e \in \{1, 4\}$ and decoder layers $L_d \in \{1, 2\}$.
- Autoformer: number of layers $L_A \in \{1, 3\}$ and kernel size $Q \in \{13, 25, 51\}$.
- FEDformer: number of layers $L_F \in \{1, 3\}$ and frequency modes $n_{\text{modes}} \in \{8, 32\}$.

The best configurations obtained from this process are reported in Table 2 for the RFF-recurrent models and Table 3 for the Transformer baselines. In comparative terms, the RFF-enhanced variants (RFF-SRNN, RFF-GRU, and RFF-LSTM) are selected with architecture-specific spectral settings ($\eta_\rho, \mathcal{B}, N_f, d_h, L, p_{\text{spec}}$), whereas the standard recurrent baselines (SRNN, GRU, and LSTM) are tuned only through backbone capacity (hidden size and number of layers). This distinction is important because the RFF models optimize both temporal memory and spectral representation, while the non-RFF baselines optimize only temporal capacity. Moreover, we search three bandwidth parameters (ρ_1, ρ_2, ρ_3) to explicitly model the multi-scale nature of wind speed, which is known to contain short-term turbulent fluctuations, medium-scale diurnal variability, and longer synoptic-scale components [1,42,43]. This multi-band design is aligned with the spectral-gap perspective in atmospheric wind dynamics and motivates the three-scale RFF parameterization.

Table 2. Merged configuration table for recurrent baselines and spectral-recurrent variants across all datasets. For RFF-based models, the table includes spectral hyperparameters ($\eta_\rho, \mathcal{B}, N_f, d_h, L, p_{\text{spec}}$). For standard recurrent baselines, spectral fields are not applicable and are reported as "–".

Dataset	Model	η_ρ	$\mathcal{B}$	N_f	d_h	L	p_{spec}
Synthetic	RFF-SRNN	0.950	12-24-168	24	256	3	0.1
	RFF-GRU	1.660	4-24-168	16	176	1	0.3
	RFF-LSTM	1.180	4-24-168	32	48	1	0.5
	SRNN	–	–	–	96	3	–
	GRU	–	–	–	80	2	–
	LSTM	–	–	–	128	3	–
Argonne	RFF-SRNN	0.580	6-24-72	48	32	3	0.2
	RFF-GRU	0.950	12-24-168	24	256	3	0.1
	RFF-LSTM	3.850	12-24-168	104	144	3	0.2
	SRNN	–	–	–	96	3	–
	GRU	–	–	–	80	2	–
	LSTM	–	–	–	128	3	–
Beijing	RFF-SRNN	0.580	6-24-72	48	32	3	0.2
	RFF-GRU	0.950	12-24-168	24	256	3	0.1
	RFF-LSTM	3.850	12-24-168	104	144	3	0.2
	SRNN	–	–	–	160	3	–
	GRU	–	–	–	80	2	–
	LSTM	–	–	–	128	3	–
Chengdu	RFF-SRNN	1.180	4-24-168	48	104	3	0.3
	RFF-GRU	0.950	12-24-168	24	256	3	0.1
	RFF-LSTM	3.640	4-24-168	48	96	3	0.1
	SRNN	–	–	–	96	2	–
	GRU	–	–	–	224	3	–
	LSTM	–	–	–	128	3	–
Schiphol	RFF-SRNN	1.995	12-24-168	112	240	3	0.3
	RFF-GRU	0.613	12-24-168	56	64	3	0.2
	RFF-LSTM	3.648	4-24-168	48	112	3	0.3
	SRNN	–	–	–	160	2	–
	GRU	–	–	–	80	3	–
	LSTM	–	–	–	128	3	–
De Bilt	RFF-SRNN	1.007	4-24-168	88	48	1	0.2
	RFF-GRU	0.662	12-24-168	40	176	1	0.1
	RFF-LSTM	3.648	4-24-168	48	96	3	0.1
	SRNN	–	–	–	48	2	–
	GRU	–	–	–	80	1	–
	LSTM	–	–	–	160	1	–
Leeuwarden	RFF-SRNN	0.580	6-24-72	48	32	3	0.2
	RFF-GRU	1.706	6-24-72	12	96	3	0.1
	RFF-LSTM	3.648	4-24-168	48	240	3	0.3
	SRNN	–	–	–	48	2	–
	GRU	–	–	–	80	3	–
	LSTM	–	–	–	128	1	–

 doi:10.20944/preprints202603.1563.v1

Table 3. Hyperparameter configurations identified via Bayesian Optimization for the Transformer-based baselines. To ensure fair structural comparison, the table reports the model dimensions (d_{model}), number of attention heads (n_{heads}), dropout rates (p_{drop}), and architecture-specific topology constraints for the standard Transformer, Autoformer, and FEDformer.

Dataset	Model	d_{model}	n_{heads}	p_{drop}	L_e	L_d	$L_{A/F}$	Q	n_{modes}
Synthetic	Autoformer	96	4	0.134	–	–	2	13	–
	FEDformer	192	1	0.052	–	–	3	–	23
	Transformer	64	1	0.103	4	2	–	–	–
Argonne	Autoformer	96	8	0.131	–	–	2	13	–
	FEDformer	256	8	0.147	–	–	1	–	25
	Transformer	64	1	0.103	4	2	–	–	–
Beijing	Autoformer	64	8	0.097	–	–	3	25	–
	FEDformer	192	2	0.262	–	–	3	–	24
	Transformer	64	1	0.103	4	2	–	–	–
Chengdu	Autoformer	96	4	0.134	–	–	2	13	–
	FEDformer	128	1	0.184	–	–	2	–	23
	Transformer	256	8	0.179	4	1	–	–	–
Schiphol	Autoformer	64	2	0.209	–	–	2	25	–
	FEDformer	256	8	0.147	–	–	1	–	25
	Transformer	64	1	0.103	4	2	–	–	–
De Bilt	Autoformer	96	8	0.131	–	–	2	13	–
	FEDformer	192	2	0.262	–	–	3	–	24
	Transformer	64	1	0.103	4	2	–	–	–
Leeuwarden	Autoformer	64	2	0.209	–	–	2	25	–
	FEDformer	256	8	0.147	–	–	1	–	25
	Transformer	64	1	0.103	4	2	–	–	–

To ensure full transparency and reproducibility, the source code, preprocessing scripts, and datasets utilized in this study are publicly available via the GitHub repository (https://github.com/ealeongomez/MSRRFF-Wind-Forecast). All experiments and hyperparameter optimization routines were executed in a standard Kaggle Notebook environment. The computational setup utilized an NVIDIA Tesla T4 GPU with 16 GB of VRAM, supported by 4 vCPUs and 16 GB of system RAM. The experimental framework was implemented in Python 3.14, utilizing PyTorch 2.6 for tensor computations and neural network training, and Optuna 4.7 for orchestrating the Bayesian hyperparameter optimization.

4. Results and Discussion

4.1. Two-Stage RFF-RNN Training Results

Figure 4 illustrates the diverse statistical and temporal characteristics of the evaluated wind speed datasets, underscoring the complexity of the forecasting task. The left column, presenting the Kernel Density Estimates (KDE), reveals significant distributional heterogeneity across the geographic locations. While the Argonne dataset exhibits a relatively smooth, right-skewed profile typical of standard wind speed distributions (resembling a Weibull distribution), datasets such as Chengdu, De Bilt, Leeuwarden, and Schiphol display highly multi-modal and spiky distributions, indicative of complex localized atmospheric constraints, topography, or measurement discretization. Furthermore, the right column highlights the inherent non-stationarity of the time series across the chronological training (blue), validation (orange), and testing (green) partitions. Pronounced variations in amplitude, volatility, and distinct regime shifts are clearly visible, particularly within the Beijing dataset and the controlled Synthetic benchmark. Together, these plots provide a strong empirical justification for the proposed RFF-RNN framework. The extreme variations in both marginal distributions and temporal dynamics demonstrate that traditional models with fixed functional forms or static hyperparameters would struggle to generalize. Instead, these conditions necessitate the proposed multi-scale spectral encoder, which leverages learnable bandwidths to dynamically adapt to the specific frequency components and structural shifts unique to each wind regime.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 March 2026 doi:10.20944/preprints202603.1563.v1

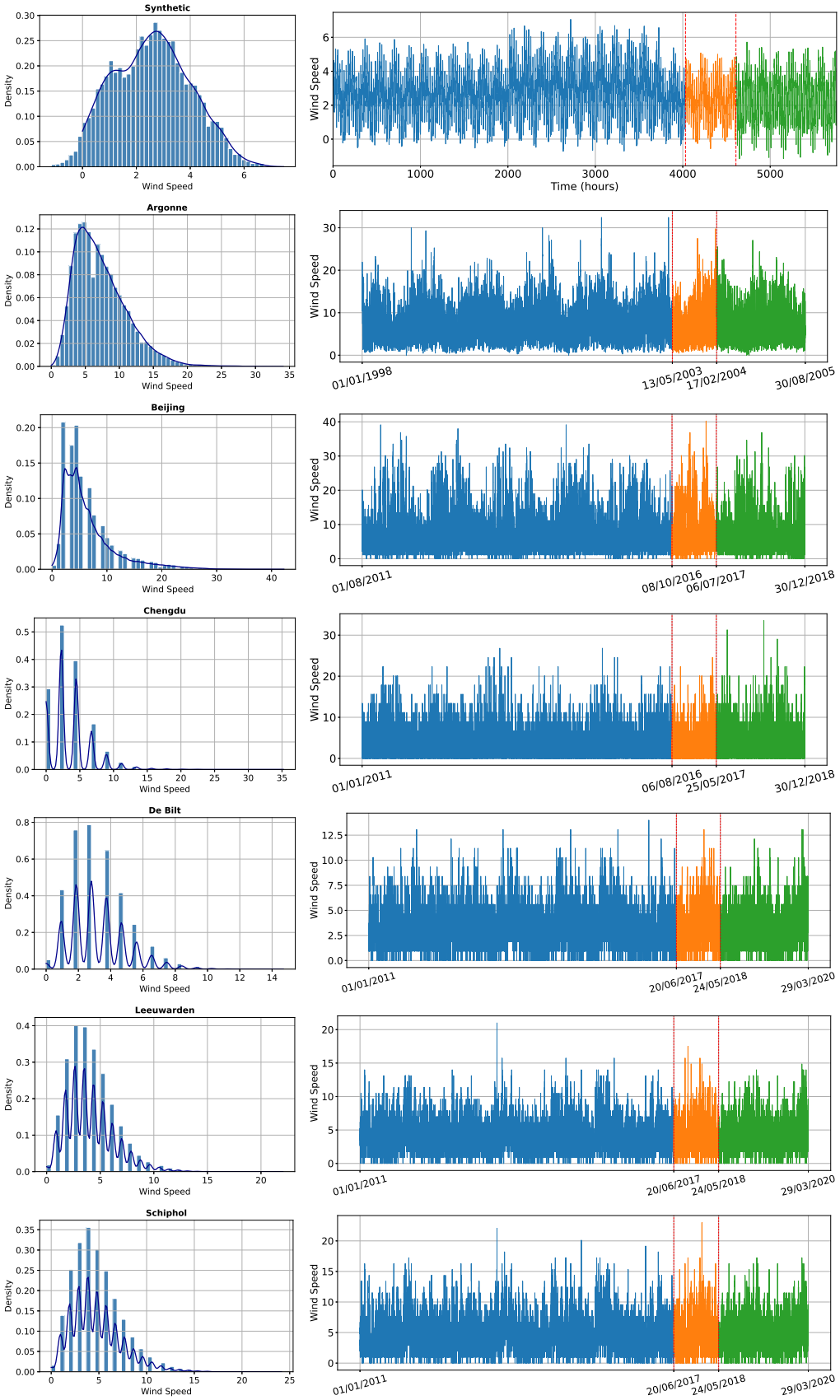


Figure 4. Wind speed forecasting datasets. The left column presents the kernel density estimate (KDE) of the time series, while the right column shows the corresponding amplitude data (blue: training, orange: validation, and green: testing).

 doi:10.20944/preprints202603.1563.v1

Figure 5 illustrates the backpropagation-based convergence of the RFF-RNN spectral scales alongside a normalized relevance analysis computed via the integrated gradients algorithm. The architecture leverages three distinct bandwidth parameters (ρ_1, ρ_2, ρ_3) to explicitly capture the multi-scale nature of wind speed, which inherently consists of short-term turbulent fluctuations, medium-scale diurnal variability, and longer synoptic-scale components [5]. This multi-band design aligns precisely with the spectral-gap perspective in atmospheric wind dynamics and justifies the three-scale RFF parameterization. As observed in the convergence plots (left column) for both the Synthetic and Argonne datasets, the bandwidth parameters exhibit a consistent decreasing trend during the initial training epochs before gradually stabilizing. This behavior indicates that the model progressively refines its spectral resolution, transitioning from broader kernel representations toward more localized, high-frequency responses. Furthermore, the integrated gradients analysis (right column) provides critical insight into the relative contribution of each scale to the model's predictions. Across both datasets, the small-scale component (ρ_1) demonstrates the highest predictive importance (approaching 50%), underscoring the model's heavy reliance on capturing immediate, high-frequency wind volatility for accurate forecasting. The large-scale component (ρ_3) follows as the second most relevant feature, while the medium-scale (ρ_2) contributes the least. This hierarchical convergence and distinct importance distribution confirm that the multi-scale RFF encoder effectively and automatically allocates spectral capacity across different temporal resolutions. Hence, these results demonstrate that the proposed architecture learns scale-specific spectral representations in a data-driven manner rather than relying on fixed kernel hyperparameters, validating bandwidth adaptation as a core mechanism underlying the framework's multi-scale modeling capability.

To further substantiate the scale tuning and relevance analysis, Figure 6 provides a direct visual representation of the learned multi-scale embeddings alongside their corresponding forecasting outputs. The heatmaps (right panels) display the extracted RFF feature maps, structurally ordered from the lowest to the highest learned bandwidth (ρ_k). It is visually apparent that the lowest bandwidths generate smooth, temporally persistent feature tracks, which effectively encode low-frequency, synoptic wind trends. Conversely, as the bandwidth increases, the feature maps become markedly more granular, successfully capturing the rapid, high-frequency turbulent fluctuations that the integrated gradients analysis identified as highly predictive. By explicitly disentangling these temporal scales into distinct spectral feature spaces, the network avoids spectral bias and preserves complex oscillatory dynamics. This structured decomposition directly translates into the robust predictive performance observed in the left panels, where the model maintains tight alignment with the ground-truth horizon across both the highly periodic Synthetic benchmark and the non-stationary, noisy Argonne dataset.

In turn, Figure 7 summarizes the relative importance of spectral and recurrent hyperparameters obtained from the Bayesian Optimization process on the synthetic and real-world datasets. The analysis highlights how different architectural components contribute to model performance across recurrent backbones. For the RFF-SRNN model, importance is distributed between spectral configuration hyperparameters and network depth, indicating a balanced sensitivity to both spectral encoding and recurrent structure. In contrast, the RFF-GRU model places greater emphasis on recurrent capacity parameters, particularly the hidden dimension and spectral scaling factor, suggesting stronger reliance on gated temporal dynamics. The RFF-LSTM model shows dominant sensitivity to the number of layers and the number of Fourier features per band, reflecting its higher capacity to exploit richer spectral representations. Overall, these results confirm that hyperparameter relevance varies across recurrent backbones, while consistently highlighting the importance of spectral design choices. This supports the role of the proposed multi-scale RFF encoder as a key contributor to model performance, even in controlled synthetic settings.

 doi:10.20944/preprints202603.1563.v1

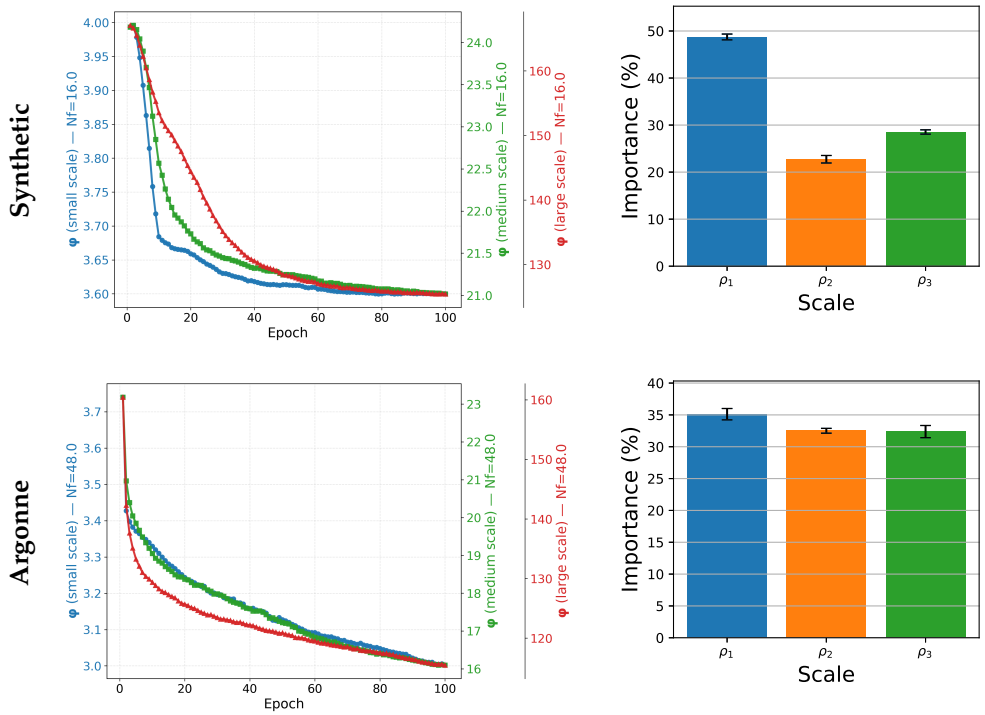


Figure 5. Evolution of the learnable RFF-RNN kernel-bandwidth parameters (ρ_k) across training epochs for the Synthetic benchmark and the Argonne real-world dataset. Left column: the curves show how gradient descent progressively tunes the spectral scales of the RFF encoder across small-, medium-, and large-scale components. Right column: integrated gradient-based scale relevance along the testing set.

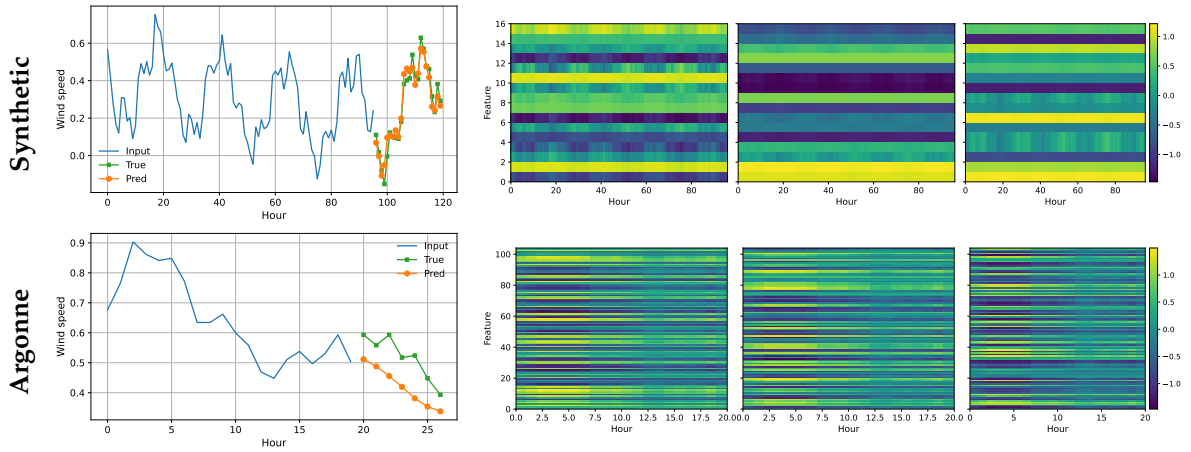


Figure 6. Illustrative examples of multi-scale RFF encoder feature maps, structurally ordered from lower to higher learned bandwidth (ρ_k), and the corresponding wind speed forecasting results (input window versus forecast horizon) for the Synthetic and Argonne datasets.

 doi:10.20944/preprints202603.1563.v1

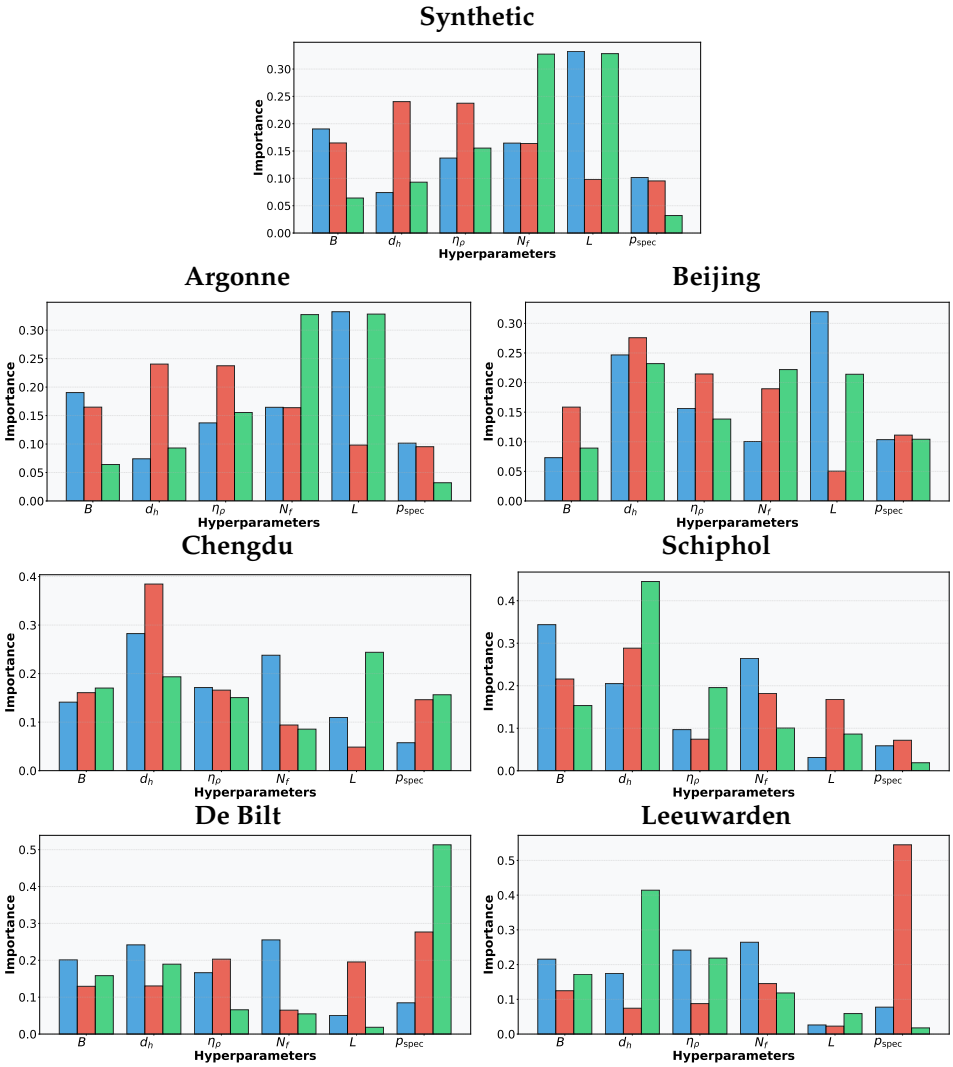


Figure 7. Hyperparameter importance for the spectral–recurrent models across the Synthetic benchmark and six real-world datasets. Bar colors indicate model variants: blue for RFF-SRNN, red for RFF-GRU, and green for RFF-LSTM. The plots quantify the contribution of each hyperparameter to the validation loss, as inferred from Bayesian optimization.

4.2. Horizon-Wise Wind Speed Forecasting Results

Figure 8 presents the horizon-wise forecasting performance on the synthetic time series using the R^2, MAE, and RMSE metrics. As a controlled benchmark, this dataset enables an isolated analysis of error propagation across prediction horizons, providing a clear view of model behavior without the influence of real-world noise or non-stationarity. Across all horizons, the RFF-enhanced models consistently outperform their baseline recurrent counterparts, confirming the benefit of incorporating explicit spectral representations. Among all evaluated approaches, the RFF-GRU model achieves the best overall performance, exhibiting the highest R^2 values and the lowest MAE and RMSE across nearly all prediction horizons. This superior behavior remains stable as the horizon increases, indicating strong robustness to error accumulation. In contrast, baseline recurrent architectures (SRNN, GRU, and LSTM) display a more pronounced degradation in performance as the forecast horizon grows, while transformer-based models, although competitive at short horizons, are consistently outperformed at medium and long horizons.

 doi:10.20944/preprints202603.1563.v1

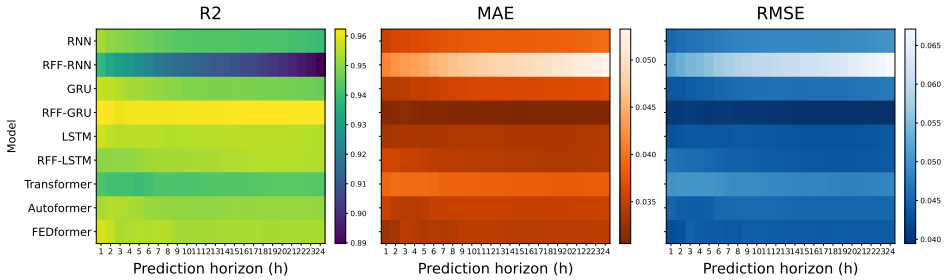


Figure 8. Horizon-wise performance heatmaps on the synthetic time series for all evaluated models. The plots report R^2, MAE, and RMSE across increasing prediction horizons (24 h).

Having established the effectiveness of spectral–recurrent modeling in a controlled synthetic setting, we next examine whether these gains translate to real-world wind speed forecasting, where non-stationarity, regional variability, and measurement noise pose additional challenges. Figure 9 summarizes the horizon-wise forecasting performance across the six real-world wind speed datasets using the same evaluation metrics. Despite the increased complexity of real atmospheric dynamics, the performance trends observed in the synthetic benchmark are consistently reproduced, indicating strong generalization capability of the proposed approach. Across all datasets and prediction horizons, the RFF-augmented models systematically outperform their baseline recurrent counterparts. These improvements are reflected in higher R^2 values and lower MAE and RMSE, together with smoother degradation patterns as the forecast horizon increases, suggesting that spectral–recurrent models are more robust to error accumulation in medium- and long-range forecasting tasks. Among all evaluated approaches, the RFF-LSTM model achieves the strongest and most consistent overall performance across datasets, followed by the RFF-GRU. This shift—from RFF-GRU being optimal in the synthetic case to RFF-LSTM in real-world scenarios—highlights the role of architectural complexity when modeling richer, non-stationary temporal dynamics. Performance gains are particularly pronounced in datasets characterized by stronger short-term variability, such as Beijing and Schiphol, where explicit spectral encoding provides advantage in capturing high-frequency temporal components. In smoother datasets, such as Chengdu, the improvements are smaller yet consistent across horizons. Overall, this horizon-wise analysis demonstrates that the proposed multi-scale RFF framework not only improves forecasting accuracy in controlled settings but also generalizes effectively to complex real-world wind speed data.

Figure 10 presents a qualitative comparison of temporal reconstructions on the synthetic time series across increasing prediction horizons. This controlled setting allows for a clear visual assessment of how different model families preserve temporal structure as forecasting difficulty increases. The figure contrasts baseline recurrent models (dashed lines with circles), RFF-enhanced architectures (solid lines with squares), and transformer-based models (dash-dot lines with diamonds), with the black dashed line indicating the ground truth. At short horizons ($h = 1$), all models closely follow the reference signal, showing minimal differences. As the horizon increases ($h = 5$ and $h = 10$), baseline recurrent models progressively smooth high-frequency oscillations and exhibit slight phase shifts. In contrast, the RFF-enhanced models—RFF-RNN, RFF-GRU, and RFF-LSTM—preserve sharper dynamics and remain better aligned with the ground truth. At longer horizons ($h \geq 15$), the degradation of baseline models becomes more evident, while the RFF-based architectures retain more coherent oscillatory patterns. These observations confirm that explicit spectral encoding improves temporal stability as the prediction horizon increases under idealized conditions. In turn, Figure 11 provides a qualitative comparison of multi-horizon wind speed reconstructions across all real-world datasets. It includes hourly datasets (Argonne, Beijing, and Chengdu) and higher-frequency Netherlands datasets (Schiphol, De Bilt, and Leeuwarden), allowing direct visual assessment of temporal fidelity as the forecast horizon increases. At short horizons ($h = 1$), all models track the ground-truth signal closely, reproducing the main oscillatory behavior and local variations, consistent with the synthetic benchmark. At intermediate horizons ($h = 4$–5 and $h = 6$–7), baseline recurrent models begin to smooth

 doi:10.20944/preprints202603.1563.v1

sharp peaks and show mild phase shifts, particularly in more variable series such as Beijing and Schiphol. By contrast, RFF-enhanced models preserve higher-frequency fluctuations and remain more aligned with the reference signal. At longer horizons ($h \geq 10$–12), baseline degradation becomes more evident, with attenuated amplitudes and delayed responses, whereas RFF-based architectures—especially RFF-LSTM—maintain more coherent temporal dynamics and better recover cyclical patterns. Overall, these visual findings are consistent with the synthetic-case trends and support the conclusion that spectral–recurrent modeling improves temporal robustness in complex real-world wind speed forecasting.

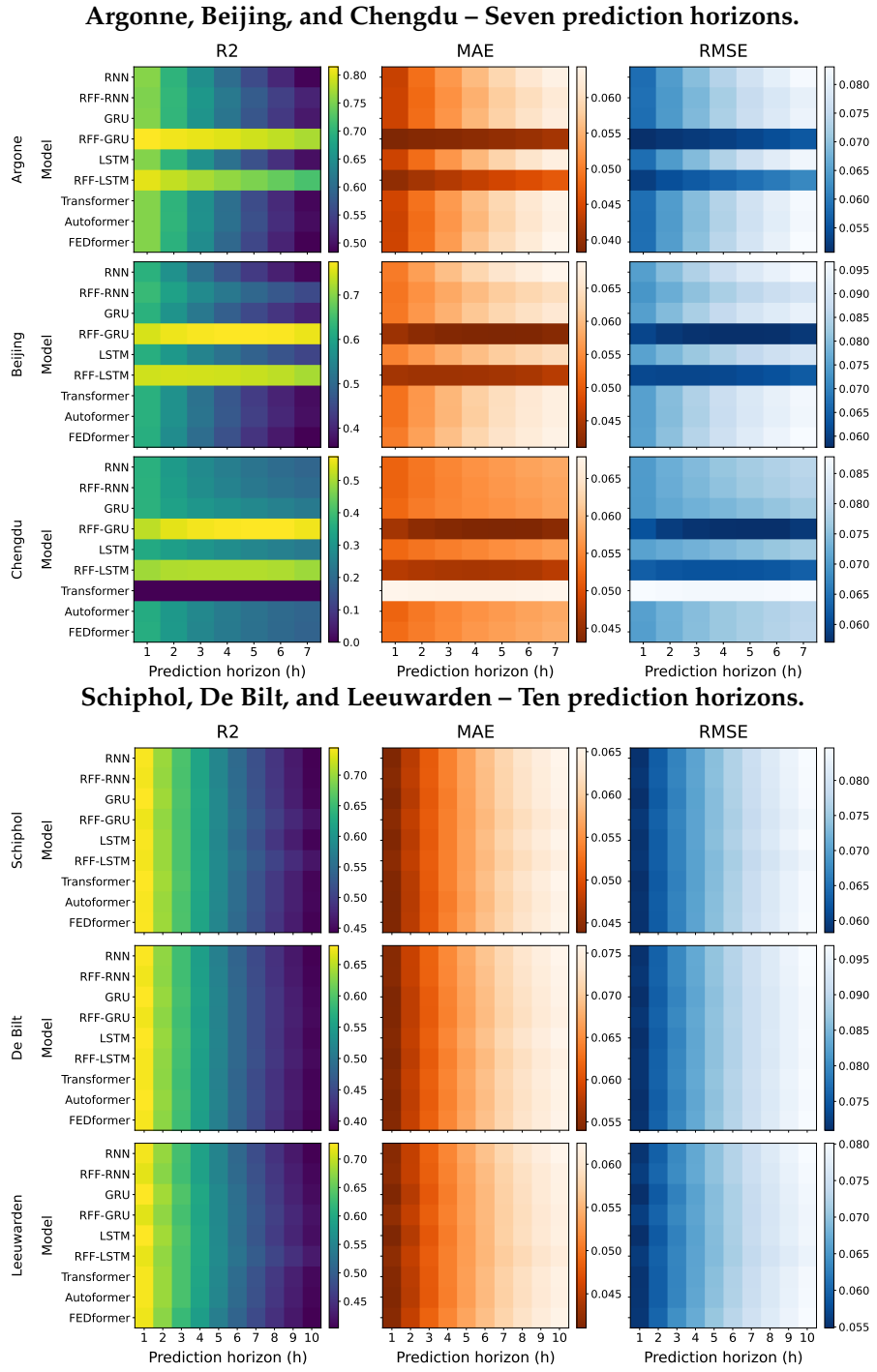


Figure 9. Horizon-wise performance heatmaps for the real-world wind speed datasets. The plots report R^2, MAE, and RMSE across increasing prediction horizons for all evaluated models.

 doi:10.20944/preprints202603.1563.v1

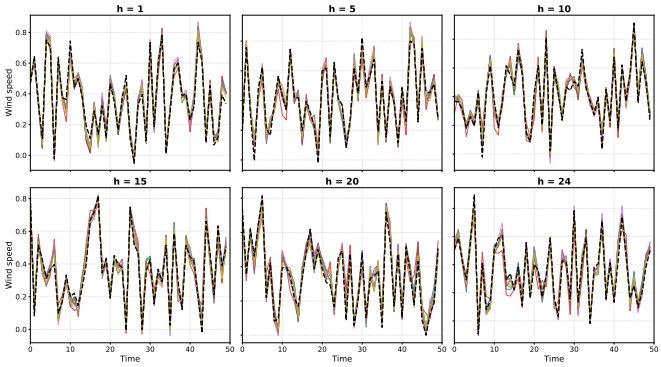
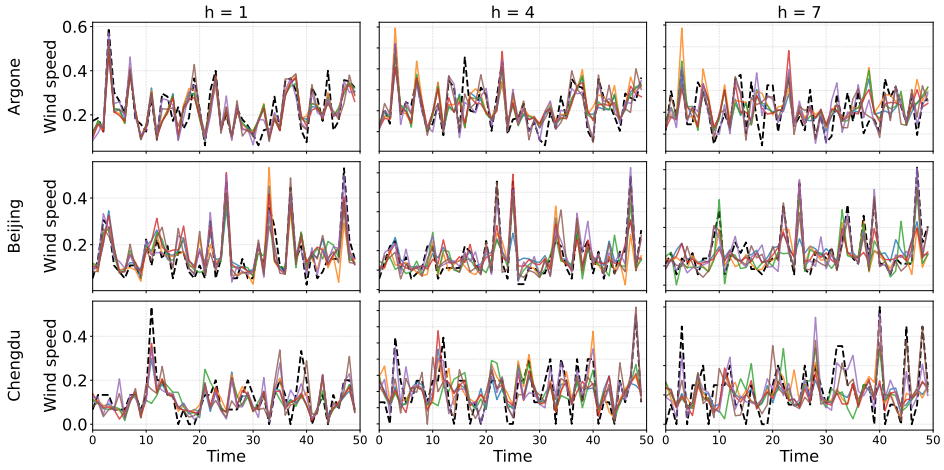


Figure 10. Multi-horizon temporal reconstruction on the synthetic time series. Forecasts are shown for increasing prediction horizons ($h = 1, 5, 10, 15, 20, 24$), each displaying 50 consecutive time steps. The black dashed line **- - -** represents the ground truth signal. Baseline recurrent models: — RNN, — GRU, — LSTM. RFF-enhanced models: — RFF-RNN, — RFF-GRU, — RFF-LSTM. Transformer-based models: — Transformer, — Autoformer, — FEDformer.

Hourly frequency datasets: Argonne, Beijing, and Chengdu, shown at early ($h = 1$), intermediate ($h = 4$–5), and longer ($h = 7$–10) forecast horizons.



Higher-frequency Netherlands datasets: Schiphol, De Bilt, and Leeuwarden, shown at $h = 1$, $h = 6$, and $h = 12$.

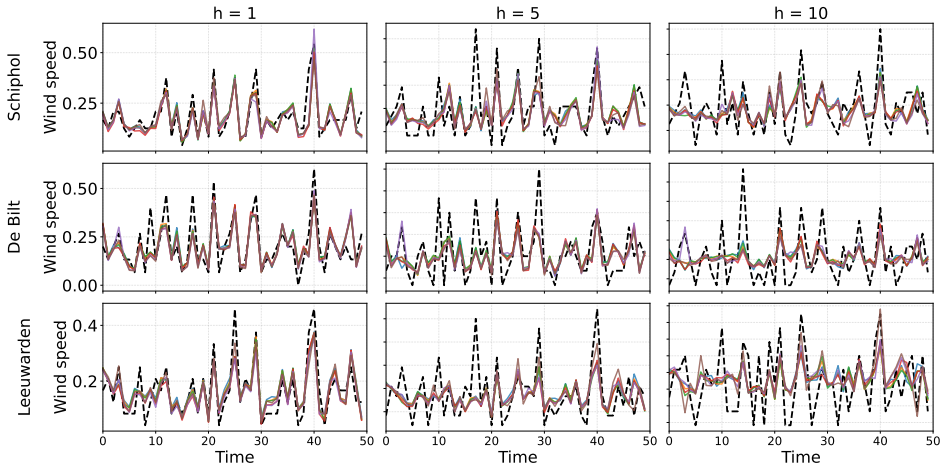


Figure 11. Multi-horizon wind speed time series reconstruction for real-world datasets (Argonne, Beijing, Chengdu, and three sites in the Netherlands). Forecasts are shown for three representative prediction horizons per location, displaying 50 consecutive time steps each. The black dashed line **- - -** indicates ground truth observations. Baseline recurrent models: — RNN, — GRU, — LSTM. RFF-enhanced models: — RFF-RNN, — RFF-GRU, — RFF-LSTM. Transformer-based models: — Transformer, — Autoformer, — FEDformer.

 doi:10.20944/preprints202603.1563.v1

4.3. Statistical Comparison Results

To complement the horizon-wise heatmap analysis, we performed a non-parametric Friedman chi-square test across all models for each dataset and metric [44]. Table 4 reports the average ranks (mean ± standard deviation) computed over all horizons for MAE, R^2, and RMSE, together with the corresponding p-values and χ^2 statistics. For all metrics, *lower ranks indicate better performance* (including R^2, after ranking models in descending order of R^2). Across datasets, a consistent pattern emerges: incorporating multi-band RFF improves the relative standing of recurrent backbones. In most cases, RFF-based variants achieve the top ranks and clearly outperform their baseline counterparts, indicating that explicit spectral encoding yields more reliable multi-horizon forecasts. This trend is particularly strong in MAE, where RFF-GRU is ranked first in Argonne, Beijing, and Chengdu, while RFF-LSTM and RFF-GRU remain among the best performers in the Netherlands stations. A similar behavior is observed for R^2 and RMSE. In the hourly datasets (Argonne, Beijing, Chengdu), RFF-GRU obtains the best ranks, followed by RFF-LSTM. In the Netherlands datasets, RFF-LSTM becomes more competitive and frequently attains the top position (e.g., Leeuwarden and Schiphol for R^2/RMSE), suggesting that the gating mechanism of LSTM combined with spectral features is especially robust under higher-frequency regimes.

The Friedman test strongly supports the statistical relevance of these differences. For all datasets and metrics, the p-values are far below 0.05 (often orders of magnitude smaller), with large χ^2 values, confirming that the ranking shifts are unlikely to be due to chance. Notably, the synthetic benchmark exhibits the strongest separation (very small p-values and the largest χ^2), which is consistent with the controlled setting where the effect of spectral encoding can be observed with minimal confounding noise. Overall, the statistical comparison confirms that the proposed spectral–recurrent strategy systematically improves performance across horizons.

Table 4. Friedman chi-squared statistical test and average ranking analysis across all forecasting models. Rankings are reported as mean ± standard deviation across prediction horizons. RMSE and R^2 provide inversely symmetric rankings. Best-ranked models per dataset are highlighted in bold.

Dataset	Model	MAE Rank	R^2 & RMSE Rank
Synthetic	RNN	7.62 ± 0.63	7.62 ± 0.63
	RFF-RNN	9.00 ± 0.00	9.00 ± 0.00
	GRU	6.04 ± 0.89	6.04 ± 1.10
	RFF-GRU	**1.00 ± 0.00**	**1.00 ± 0.00**
	LSTM	3.12 ± 0.88	3.33 ± 0.99
	RFF-LSTM	3.17 ± 1.37	3.25 ± 1.36
	Transformer	6.83 ± 0.90	6.75 ± 0.92
	Autoformer	4.58 ± 1.04	4.54 ± 1.04
	FEDformer	3.62 ± 1.41	3.46 ± 1.47
	χ^2 (p-value)	*167.29 (4.77e-32)*	*164.69 (1.67e-31)*
Argonne	RNN	7.00 ± 1.69	7.57 ± 1.92
	RFF-RNN	3.57 ± 1.40	3.86 ± 2.10
	GRU	4.29 ± 0.70	4.00 ± 0.00
	RFF-GRU	**1.00 ± 0.00**	**1.00 ± 0.00**
	LSTM	5.14 ± 0.64	5.14 ± 0.35
	RFF-LSTM	2.00 ± 0.00	2.00 ± 0.00
	Transformer	7.71 ± 1.03	7.14 ± 0.35
	Autoformer	5.71 ± 0.70	5.86 ± 0.35
	FEDformer	8.57 ± 0.73	8.43 ± 0.73
	χ^2 (p-value)	*48.72 (7.18e-08)*	*47.62 (1.17e-07)*

 doi:10.20944/preprints202603.1563.v1

Table 4. *Cont.*

Dataset	Model	MAE Rank	R^2 & RMSE Rank
Beijing	RNN	8.57 ± 0.73	7.43 ± 1.40
	RFF-RNN	4.00 ± 0.53	3.29 ± 0.45
	GRU	5.14 ± 0.35	5.00 ± 0.00
	RFF-GRU	**1.00 ± 0.00**	**1.00 ± 0.00**
	LSTM	4.00 ± 2.07	4.29 ± 1.58
	RFF-LSTM	2.00 ± 0.00	2.00 ± 0.00
	Transformer	6.14 ± 0.99	7.00 ± 0.00
	Autoformer	6.29 ± 1.58	6.00 ± 0.00
	FEDformer	7.86 ± 0.64	9.00 ± 0.00
	χ^2 *(p-value)*	*47.50 (1.23e-07)*	*51.66 (1.96e-08)*
Chengdu	RNN	6.14 ± 1.46	6.29 ± 0.70
	RFF-RNN	4.43 ± 0.73	4.86 ± 0.35
	GRU	5.14 ± 1.55	3.43 ± 0.49
	RFF-GRU	**1.00 ± 0.00**	**1.00 ± 0.00**
	LSTM	3.57 ± 1.05	4.14 ± 1.64
	RFF-LSTM	2.00 ± 0.00	2.00 ± 0.00
	Transformer	9.00 ± 0.00	9.00 ± 0.00
	Autoformer	6.14 ± 0.99	6.57 ± 0.49
	FEDformer	7.57 ± 0.49	7.71 ± 0.70
	χ^2 *(p-value)*	*49.10 (6.07e-08)*	*52.00 (1.68e-08)*
De Bilt	RNN	8.60 ± 0.49	8.20 ± 1.47
	RFF-RNN	6.80 ± 1.08	6.00 ± 1.26
	GRU	3.80 ± 1.47	**2.50 ± 1.12**
	RFF-GRU	1.90 ± 1.14	3.10 ± 2.70
	LSTM	7.50 ± 1.50	3.90 ± 1.70
	RFF-LSTM	**1.70 ± 0.64**	2.60 ± 1.43
	Transformer	5.10 ± 1.51	6.40 ± 1.11
	Autoformer	3.30 ± 0.46	4.50 ± 1.43
	FEDformer	6.30 ± 1.00	7.80 ± 1.17
	χ^2 *(p-value)*	*65.31 (4.20e-11)*	*50.83 (2.83e-08)*
Leeuwarden	RNN	8.50 ± 0.92	8.10 ± 1.30
	RFF-RNN	6.00 ± 1.48	5.90 ± 1.58
	GRU	5.50 ± 1.86	3.80 ± 1.25
	RFF-GRU	3.00 ± 2.19	3.10 ± 2.30
	LSTM	3.40 ± 1.56	2.70 ± 0.64
	RFF-LSTM	**2.20 ± 2.44**	**1.90 ± 1.92**
	Transformer	5.10 ± 1.45	5.30 ± 0.90
	Autoformer	4.20 ± 1.66	6.70 ± 1.68
	FEDformer	7.10 ± 1.58	7.50 ± 1.28
	χ^2 *(p-value)*	*43.95 (5.82e-07)*	*52.80 (1.18e-08)*
Schiphol	RNN	7.80 ± 1.40	7.80 ± 1.60
	RFF-RNN	5.40 ± 1.11	7.20 ± 0.75
	GRU	4.80 ± 1.89	4.30 ± 0.90
	RFF-GRU	3.00 ± 2.79	3.40 ± 2.80
	LSTM	5.30 ± 0.78	6.80 ± 0.98
	RFF-LSTM	3.30 ± 3.13	**1.90 ± 1.81**
	Transformer	7.30 ± 1.85	3.80 ± 2.23
	Autoformer	**2.50 ± 0.81**	3.70 ± 0.90
	FEDformer	5.60 ± 2.01	6.10 ± 2.30
	χ^2 *(p-value)*	*35.89 (1.84e-05)*	*43.89 (5.96e-07)*

 doi:10.20944/preprints202603.1563.v1

4.4. Limitations

Despite the consistent performance improvements achieved by the proposed RFF-RNN framework across heterogeneous wind regimes, several limitations must be acknowledged. Primarily, the integration of a multi-band Random Fourier Feature encoder inherently increases the computational complexity and memory footprint of the model. By projecting low-dimensional temporal inputs into an expanded, high-dimensional spectral space ($F = K \times N_f$), the parameter count and the computational cost of both the forward and backward passes grow significantly compared to baseline recurrent architectures. Furthermore, the framework's success is heavily contingent on the rigorous two-stage optimization process. While the Bayesian hyperparameter search via TPE effectively navigates the complex interactions between spectral bandwidths and neural capacities, it introduces a substantial computational burden during the training phase. This reliance on extensive iterative search procedures may limit the framework's rapid deployment or scalability in highly resource-constrained operational environments.

Secondly, the architectural decision to jointly optimize the frequency matrices ($\mathbf{W}_k$) and phase biases ($\mathbf{b}_k$) alongside the network weights introduces specific theoretical trade-offs. Although fine-tuning these parameters enables the model to learn highly adaptive, data-driven spectral embeddings, it explicitly relaxes the shift-invariance constraints of Bochner's theorem, thereby forfeiting the strict mathematical guarantees of approximating a pure Gaussian kernel. Additionally, the current formulation and evaluation of the RFF-RNN are restricted to a univariate forecasting paradigm, relying exclusively on historical wind speed data. In practical wind energy systems, atmospheric dynamics and turbine power generation are influenced by a complex interplay of multivariate meteorological factors, including wind direction, temperature, and atmospheric pressure. Extending the multi-scale spectral encoding to capture cross-variable spatio-temporal interactions in a multivariate context remains a critical avenue for future research to fully align the model with operational grid management requirements.

5. Conclusions

In this study, we introduced a novel multi-scale spectral–recurrent framework, termed RFF-RNN, designed to address the inherent non-stationarity and complex multi-scale variability of wind speed time series. By integrating a multi-band Random Fourier Feature (RFF) encoder with recurrent neural network backbones (SRNN, GRU, and LSTM), the proposed architecture explicitly maps raw temporal sequences into a rich, frequency-aware representation space. A key innovation of this approach is the deliberate relaxation of strict shift-invariance constraints; by jointly optimizing the spectral frequencies, phase biases, and bandwidth scales alongside the neural weights, the framework dynamically shapes a fully data-driven spectral embedding. This inner-loop learning is further supported by a robust outer-loop optimization strategy that utilizes Bayesian probabilistic search via Tree-structured Parzen Estimators, ensuring an optimal balance between architectural capacity and spectral resolution without exhaustive manual tuning.

Extensive evaluations conducted on a controlled synthetic benchmark and six heterogeneous real-world wind datasets—spanning diverse climatic regimes in the USA, China, and the Netherlands—demonstrated the consistent superiority of the spectral–recurrent approach. The RFF-enhanced models systematically outperformed baseline recurrent networks and state-of-the-art transformer architectures (such as Autoformer and FEDformer) across short- and medium-term forecasting horizons, exhibiting higher explained variance (R^2) and significantly lower error metrics (MAE and RMSE). Statistical validation through the non-parametric Friedman test confirmed that the integration of multi-band RFF mappings yields significant, systematic performance gains across all evaluated metrics. Notably, the RFF-GRU and RFF-LSTM variants showed remarkable resilience against error accumulation at extended horizons, proving their capacity to simultaneously capture high-frequency turbulent fluctuations and low-frequency synoptic patterns without succumbing to the spectral bias typical of standard deep learning models.

 doi:10.20944/preprints202603.1563.v1

Expanding our RFF-RNN approach to accommodate multivariate atmospheric inputs—such as wind direction, temperature, pressure, and localized topographical features—represents a critical next step to fully align the model with the complex, multidimensional demands of operational wind energy systems [45]. Additionally, integrating the adaptive multi-scale RFF encoder with modern attention-driven or state-space backbones could further enhance long-range dependency modeling [46]. Finally, exploring sparse or more computationally efficient kernel approximations will be vital to reduce the memory footprint of the joint optimization process, thereby facilitating the scalable deployment of spectral–recurrent models in real-time power grid management and renewable energy forecasting systems [47].

Author Contributions: Conceptualization, E.L.-G., V.-E., and A.A.-M.; data curation, E.L.-G.; methodology, E.L.-G., A.A.-M., J.M.-M., and G.C.-D.; project administration, A.A.-M., A.O-G., and G.C.-D.; supervision, A.A.-M., V.E., J.M.-M., and G.C.-D.; resources, E.L.-G., A.A.-M., A.O.-G., and G.C.-D. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The publicly available dataset and Python codes employed in this study can be found at https://github.com/ealeongomez/MSRRFF-Wind-Forecast (accessed on 1 December 2025).

Funding: This research was funded by the project: "Estimación de la capacidad de generación de energía eléctrica del agua coproducida en campos de petróleo y gas en Colombia a partir de técnicas de aprendizaje automático informado por la física", 951-2024 CONVOCATORIA FORTALECIMIENTO DEL CONOCIMIENTO GEOCIENTÍFICO Y TECNOLÓGICO DE LAS FUENTES, funded by Minciencias. Also, E. Leon-Gomez thanks to the program “Beca de Excelencia Doctoral del Bicentenario-2019-Minciencias”

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, Y.; Lou, H.; Wu, J.; Zhang, S.; Gao, S. A survey on wind power forecasting with machine learning approaches. *Neural Computing and Applications* **2024**. https://doi.org/10.1007/s00521-024-09923-4.
2. Kahveci, U.B.; Barutçu, B. Recent Hybrid Machine Learning Approaches in Wind Speed Forecasting—A Review. *Journal of Economic Surveys* **2026**.
3. Konstantinou, T.; Hatziargyriou, N. Regional wind power forecasting based on Bayesian feature selection. *IEEE Transactions on Power Systems* **2024**.
4. Guo, X.; Zhu, C.; Hao, J.; Zhang, S.; Zhu, L. A hybrid method for short-term wind speed forecasting based on Bayesian optimization and error correction. *Journal of Renewable and Sustainable Energy* **2021**, *13*, 036101.
5. Liu, H.; Yang, R. Artificial intelligence for wind speed forecasting: A review on multi-scale decomposition and intelligent fusion strategies. *Advances in Wind Engineering* **2025**, p. 100055. https://doi.org/10.1016/j.awe.2025.100055.
6. Habtemariam, E.T.; Kekeba, K.; Martinez-Ballesteros, M.; Martinez-Alvarez, F. A Bayesian optimization-based LSTM model for wind power forecasting in the Adama district, Ethiopia. *Energies* **2023**, *16*, 2317.
7. Kareem, A. Emerging frontiers in wind engineering: computing, stochastics, machine learning and beyond. *Journal of Wind Engineering and Industrial Aerodynamics* **2020**, *203*, 104230.
8. Guo, X.; Zeng, P.; Xiong, X.; Wang, G.; Cui, Y. Short-term wind power forecasting methods based on machine learning: A review and case study. *Energy Reports* **2025**, *14*, 3753–3782.
9. Wang, Y.; Zou, R.; Liu, F.; Zhang, L.; Liu, Q. A review of wind speed and wind power forecasting with deep neural networks. *Applied energy* **2021**, *304*, 117766.
10. Zhang, Z.; Lin, L.; Gao, S.; Wang, J.; Zhao, H.; Yu, H. A machine learning model for hub-height short-term wind speed prediction. *Nature Communications* **2025**, *16*, 3195.
11. Foley, A.M.; Leahy, P.G.; Marvuglia, A.; McKeogh, E.J. Current methods and advances in forecasting of wind power generation. *Renewable Energy* **2012**, *37*, 1–8.
12. Khosravi, A.; Nahavandi, S.; Creighton, D.; Atiya, A. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *29*, 3567–3581.
13. Houssein, E.H.; Mohamed, M.; Younis, E.M.; Mohamed, W.M. Artificial intelligence and classical statistical models for time series forecasting: a comprehensive review. *Journal of Big Data* **2025**.

 doi:10.20944/preprints202603.1563.v1

14. Lu, G.; Ou, Y.; Wang, Z.; Qu, Y.; Xia, Y.; Tang, D.; Kotenko, I.; Li, W. A Survey of Deep Learning for Time Series Forecasting: Theories, Datasets, and State-of-the-Art Techniques. *Computers, Materials & Continua* **2025**. https://doi.org/10.32604/cmc.2025.068024.
15. Zhang, Y.M.; Wang, H. Multi-head attention-based probabilistic CNN-BiLSTM for day-ahead wind speed forecasting. *Energy* **2023**, *278*, 127865.
16. Nie, Y.; et al. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv preprint arXiv:2211.14730* **2022**.
17. Rahaman, N.; Baratin, A.; Arpit, D.e.a. On the spectral bias of neural networks. *International Conference on Machine Learning (ICML)* **2019**.
18. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In Proceedings of the The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 95–104.
19. Wang, W.; Khalil, M.M.Y.; Bayisa, L.Y. Online variational Gaussian process for time series data. *Journal of Big Data* **2024**, *11*, 174.
20. Rahimi, A.; Recht, B. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing Systems* **2008**, *20*.
21. Kiessling, J.; Ström, E. Wind field reconstruction with adaptive random Fourier features. *Proceedings of the Royal Society A* **2021**, *477*, 20210236.
22. Rayi, V.K.; Bisoi, R.; Mishra, S.P.; Dash, P.K. Improved deep mixed kernel randomized network for wind speed prediction. *Clean Energy* **2023**, *7*, 1006–1025.
23. Seman, L.O.; Klaar, A.C.R.; Ribeiro, M.H.D.M. Enhanced Random Vector Functional Link Networks With Bayesian-Based Hyperparameter Optimization for Wind Speed Forecasting. *IEEE Access* **2025**.
24. Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated machine learning: methods, systems, challenges*; Springer, 2019.
25. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research* **2012**, *13*.
26. Quan, J.; Shang, L. Short-Term Wind Speed Forecasting Based on Ensemble Online Sequential Extreme Learning Machine and Bayesian Optimization. *Mathematical Problems in Engineering* **2020**.
27. Chaibi, M.; Tarik, L.; Berrada, M. Machine learning models based on random forest feature selection and Bayesian optimization for predicting daily global solar radiation. *International Journal of Renewable Energy Research* **2022**.
28. Murphy, K.P. *Probabilistic machine learning: an introduction*; MIT press, 2022.
29. Al-Selwi, S.M.; Hassan, M.F.; Abdulkadir, S.J.; Muneer, A.; Sumiea, E.H.; Alqushaibi, A.; Ragab, M.G. RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *Journal of King Saud University-Computer and Information Sciences* **2024**, *36*, 102068.
30. Ławryńczuk, M.; Zarzycki, K. LSTM and GRU type recurrent neural networks in model predictive control: A Review. *Neurocomputing* **2025**, *632*, 129712.
31. Saravana, M.; Roopa, M.; Arunalatha, J.; Venugopal, K. Transformers for Multivariate Time Series Forecasting: Comprehensive Analysis, Challenges, Research Opportunities and Future Prospects. *IEEE Access* **2026**.
32. Wu, H.; Xu, Y.; Wang, J.; Long, G. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, Vol. 34, pp. 22419–22430.
33. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2022, pp. 27268–27285.
34. Scholkopf, B.; Smola, A.J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*; MIT press, 2018.
35. Brault, R.; Heinonen, M.; Buc, F. Random fourier features for operator-valued kernels. In Proceedings of the Asian Conference on Machine Learning. PMLR, 2016, pp. 110–125.
36. Avron, H.; Theodoropoulos, L.; Sindhwani, V.; et al. Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees. *arXiv preprint arXiv:1804.09893* **2018**.
37. Williams, C.K.; Rasmussen, C.E. *Gaussian processes for machine learning*; Vol. 2, MIT press Cambridge, MA, 2006.
38. Sun, K.; Yu, J.; Zhang, L.; Dong, Z. A convolutional neural network model based on improved softplus activation function. In Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence. Springer, 2019, pp. 1326–1335.

 doi:10.20944/preprints202603.1563.v1

39. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* **2011**, *24*.
40. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; et al.. Attention Is All You Need. *Advances in Neural Information Processing Systems* **2017**.
42. Van der Hoven, I. Power Spectrum of Horizontal Wind Speed in the Frequency Range from 0.0007 to 900 Cycles per Hour. *Journal of Meteorology* **1957**, *14*, 160–164. https://doi.org/10.1175/1520-0469(1957)014<0160:PSOHWS>2.0.CO;2.
43. Vincent, C.L.; Dowdy, A.J. Multi-scale variability of southeastern Australian wind resources. *Atmospheric Chemistry and Physics* **2024**, *24*, 10209–10223. https://doi.org/10.5194/acp-24-10209-2024.
44. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* **2006**, *7*, 1–30.
45. Sørensen, M.L.; Nystrup, P.; Bjerregård, M.B.; Møller, J.K.; Bacher, P.; Madsen, H. Recent developments in multivariate wind and solar power forecasting. *Wiley Interdisciplinary Reviews: Energy and Environment* **2023**, *12*, e465.
46. Haruna, Y.; Lawan, A. vgamba: Attentive state space bottleneck for efficient long-range dependencies in visual recognition. *arXiv preprint arXiv:2503.21262* **2025**.
47. Peng, H.; Pappas, N.; Yogatama, D.; Schwartz, R.; Smith, N.A.; Kong, L. Random feature attention. *arXiv preprint arXiv:2103.02143* **2021**.