IMPROVING EXTREME WIND PREDICTION WITH FREQUENCY-INFORMED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate prediction of extreme wind velocities has substantial significance in industry, particularly for the operation management of wind power plants. Although the state-of-the-art data-driven models perform well for general meteorological forecasting, they may exhibit large errors for extreme weather—for example, systematically underestimating the magnitudes and short-term variation of extreme winds. To address this issue, we conduct a theoretical analysis of how the data frequency spectrum influences errors in extreme wind prediction. Based on these insights, we propose a novel loss function that incorporates a gradient penalty to mitigate the magnitude shrinkage of extreme weather. To capture more precise short-term wind velocity variations, we design a novel structure of physics-embedded machine learning models with frequency reweighting. Experiments demonstrate that, compared to the baseline models, our approach achieves significant improvements in predicting extreme wind velocities while maintaining robust overall performance.

1 INTRODUCTION

Wind velocity field prediction is crucial to both academic research and industrial practice (Masson-Delmotte et al., 2021) (Kunz et al., 2010). For instance, the wind power plants require accurate predictions of the wind speed magnitude to support accurate real-time production estimation and safe operational control, since the output power is approximately proportional to the wind magnitude cubed (v^3), and the wind turbines will cease to work for extremely large wind (Sabzehgar et al., 2020) (Wan et al., 2010). Traditionally, solving dynamic systems by mathematical methods, Numerical Weather Prediction (NWP) has been the workhorse of wind velocity prediction (Coiffier, 2011). However, recent advances in deep learning have revolutionized weather prediction, with models like FourCastNet (Pathak et al., 2022) and Pangu-Weather (Bi et al., 2022) significantly outperforming traditional NWP methods (Coiffier, 2011). Based on an extensive amount of data, these models are specialized in producing accurate overall predictions of the wind velocity field.

One of the key challenges in wind velocity prediction is to accurately predict the amplitude changes of the extreme wind. General data-driven models may be struggling with this challenge if they are trained by common loss functions (like MSE) based on regular wind speed datasets (Olivetti & Messori, 2024). For example, many data-driven forecasters systematically underestimate the amplitudes of extreme winds. This bias persists even when overall (non-extreme) skill is strong, leading to underestimated risks and missed rapid ramps in operational contexts. Therefore, addressing this challenge in extreme wind velocity prediction is the main focus of this paper.

There exist several data-driven models specifically designed for extreme weather predictions. Some models employ classical deep learning models such as RNN (Prasetya & Djamal, 2019), CNN (Zhang et al., 2019), and LSTM (Gao et al., 2018) to capture spatiotemporal dependencies in weather data. Many other models utilize generative data augmentation methods, including variational autoencoder (VAE) (Vega-Bayo et al., 2024) and diffusion models (Zhong et al., 2024), to address data scarcity through weather pattern simulation. Despite these advances, critical gaps remain for extreme wind predictions: (i) most approaches offer little theoretical (or even intuitive) explanation of why errors arise significantly for extreme winds; (ii) many methods implicitly rely on abundant training data that include extreme cases, whereas such cases are intrinsically scarce in real datasets; (iii) to better capture the dynamics of sharply-changing pattern, the data-driven models may re-

quire much more complicated model structure than overall prediction; and (iv) depending mainly on large data, some models may be insufficient for capture the intrinsic dynamics and still suffer from uncontrollable regional errors for extreme weather prediction (Zhou et al., 2024).

To resolve the challenge, we conduct a detailed theoretical analysis of the error behavior in the frequency domain. Based on proper simplification, we separate the traditional mean-squared error (MSE) into three terms: amplitude shrinkage error, pattern translation error, and noise. We show that while training with standard MSE as a loss function, small pattern deviations will lead to significant amplitude shrinkage for the high-frequency wind field components, causing the underestimation of extreme amplitude and blurred short-term variability in prediction. Inspired by the analysis, we propose a gradient-penalized loss function upweighting the amplitude shrinkage error, which encourages the model to capture the change of wind velocity magnitude more accurately. To more effectively reduce the pattern translation error, as well as improve parameter efficiency, we design a physics-embedded structure for the neural network, with a backbone of the Navier-Stokes (NS) equation. The equations reveal how the motion of a fluid, such as the atmosphere, is affected by a combination of external forces, pressures within the fluid, and viscous effects (Marion & Temam, 1998). Moreover, we utilize a frequency separation and reweighting mechanism to coordinate the impact of high- and low-frequency components to the loss function. Based on the above frequency-informed modification of the loss function and neural network structure, our model overcomes the amplitude shrinkage challenge in wind prediction and achieves a significant improvement in extreme wind velocity prediction accuracy.

Our research makes the following significant contributions to the field of extreme wind prediction:

- **Frequency-theoretic explanation of underestimation.** We provide a Fourier-domain analysis showing how small spatial shifts and scaling yield a wavenumber-dependent MSE, theoretically accounting for underestimation of extreme amplitudes and smearing of short-term variability.
- **Gradient-penalized objective for extremes.** We propose a simple, implementation-ready loss that augments MSE with gradient matching, equivalently reweighting high-frequency errors to mitigate spectral shrinkage and recover sharp ramps.
- **Frequency separation & reweighting with a physics-embedded backbone.** We design a spectral pipeline (Fourier masking, band-specific branches, learnable fusion) atop a simplified NS block with continuity regularization, targeting precise short-horizon dynamics while preserving stability and data efficiency.
- **Empirical validation on regional extremes.** Across diverse regions and strong baselines (including classical PINN variants), our method substantially improves extreme-wind prediction while maintaining robust overall performance under normal conditions.

The remainder of this paper is structured as follows: Section 2 includes problem formulation, physical backgrounds, and our theoretical analysis on predictive error. Section 3 displays our methodology, including the novel loss function formulation and the new network design. Section 4 describes our experiments and corresponding results. Section 5 contains the conclusion with limitations and future directions. Detailed specifications are provided in the appendices.

2 Preliminaries and Insights

2.1 Problem Formulation

In this paper, we mainly consider the wind velocity field prediction within certain rectangular regions, which can be discretized into $N \times M$ points. Let $\mathbf{u}(\mathbf{x}, t) = [v(x, y, t), w(x, y, t)]$ denote the wind velocity field in this region, where $\mathbf{x} = (x, y)$ is the two-dimensional spatial domain and t is the temporal domain;v and w represent the velocity components in the longitude and latitude directions, respectively. We denote historical wind data sequences by $\mathbf{u}_{[t_1:t_N]} = \{\mathbf{u}(x, y, t_1), \mathbf{u}(x, y, t_2), \ldots, \mathbf{u}(x, y, t_N)\}$.

Let $\tilde{\mathbf{u}}(\mathbf{x}, t)$ denote the prediction of $\mathbf{u}$ given by a certain model at time t. Then our objective to predict the wind velocity field at the next time can be expressed as follows:

$$\tilde{\mathbf{u}}_{t_{N+1}} = \tilde{\mathbf{u}}(x, y, t_{N+1}) = f_\theta(\mathbf{u}_{[t_1:t_N]}, \text{other data}), \tag{1}$$

where f_θ is the model we intend to train, and "other data" contains other data sequences that might also contribute to wind velocity prediction (like surface pressure, which will be explained later). The error between $\mathbf{u}$ and $\tilde{\mathbf{u}}$ evaluates the performance of the predictive model.

Temporal and Spatial Scale for Extreme Prediction. In atmospheric forecasting, temporal and spatial scales are tightly coupled: short-term predictions are typically associated with short-range dynamics (Jung & Broadwater, 2014) (Zhu et al., 2019). In this paper, we adopt the convention of extreme wind velocity prediction that focuses on short-period and regional prediction, while the temporal and spatial resolutions are also higher compared to global weather forecasting to resolve rapidly evolving, small-scale features.

2.2 Physical Backgrounds

As a fundamental assumption in meteorology, atmospheric systems' dynamics generally satisfy the **Navier-Stokes (NS) equations** (Holton & Hakim, 2013). The Navier-Stokes equations are a set of nonlinear partial differential equations (PDEs) that describe the relationship between the motion of a fluid and the forces acting upon it. For a two-dimensional domain with wind velocity field $\mathbf{u} = [v(x, y, t), w(x, y, t)]$, the NS equation is shown as follows:

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{u} \cdot \nabla \mathbf{u} - \frac{1}{\rho} \nabla P + \nu \nabla^2 \mathbf{u} + \mathbf{F} \tag{2}$$

where $\mathbf{u} \cdot \nabla \mathbf{u}$ is the advective acceleration; $-\frac{1}{\rho}\nabla P$ refers to the pressure gradient force; $\nu \nabla^2 \mathbf{u}$ denotes the viscous friction (ν: kinematic viscosity); and $\mathbf{F}$ is the external body forces. The body force term may vary in different scenarios, with typical examples including gravity and the Coriolis force (Holton & Hakim, 2013).

2.3 Insights from Frequency Domain Analysis

When solving and analyzing PDEs, a standard method is to apply the Fourier transform (often with respect to spatial domains) to convert the PDEs into ODEs. Let $\hat{\cdot}$ denote the Fourier operator, and $\mathbf{k} = (k_x, k_y)$ denote the frequency domain coordinates. For example, applying the Fourier transform to a simplified version of the NS equation equation 2 with advection and diffusion:

$$\partial_t \mathbf{u}(\mathbf{x}, t) + \mathbf{U} \cdot \nabla \mathbf{u}(\mathbf{x}, t) \;=\; \nu \nabla^2 \mathbf{u}(\mathbf{x}, t) + f(\mathbf{x}, t),$$

we will get

$$\partial_t \hat{\mathbf{u}}(\mathbf{k}, t) + i\,(\mathbf{k} \cdot \mathbf{U})\,\hat{\mathbf{u}}(\mathbf{k}, t) \;=\; -\nu \|\mathbf{k}\|^2 \hat{\mathbf{u}}(\mathbf{k}, t) + \hat{f}(\mathbf{k}, t), \tag{3}$$

which is an ODE with respect to $\hat{\mathbf{u}}$. Equation equation 3 shows that the advection of air may appear as a phase shift $e^{-i\,\mathbf{k}\cdot\mathbf{U}\,t}$, corresponding to a spatial translation in physical space. Moreover, diffusion may induce amplitude damping at a rate proportional to $\|\mathbf{k}\|^2$ (stronger for high frequency).

Motivated by this idea, we apply a 2-dimensional Fourier transform to the wind velocity fields and analyze how the frequency spectrum affects the prediction error. By equation equation 3 and statistical convention, we assume that the prediction error is mainly caused by three factors: scaling, translation, and noise. Therefore, the relationship between $\tilde{\mathbf{u}}$ and $\mathbf{u}$ can be illustrated as follows:

$$\tilde{\mathbf{u}}(\mathbf{x}) = a\mathbf{u}(\mathbf{x} + \Delta) + \varepsilon(\mathbf{x}), \tag{4}$$

where a is the scaling magnitude of wind speed amplitude, and Δ corresponds to the deviation amount of the data pattern. We may also assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ to be a Gaussian noise.

Now, let's consider the mean squared error (MSE) of the prediction and the ground-truth. By Rayleigh's energy theorem, we can show that the MSE of the original data in the spatial domain is equivalent to the MSE of the Fourier-transformed data in the (double) frequency domain:

$$\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}}) = \frac{1}{NM} \sum_{\mathbf{x}} |\mathbf{u} - \tilde{\mathbf{u}}|^2 = \frac{1}{(NM)^2} \sum_{k} \left| \hat{\mathbf{u}} - \hat{\tilde{\mathbf{u}}} \right|^2.$$

We denote $\theta_{\mathbf{k}} = 2\pi \left(\frac{k_x \Delta_x}{N} + \frac{k_y \Delta_y}{M} \right)$, and we assume θ_k is sufficiently small. Then the Fourier transform of the prediction is $\hat{\tilde{\mathbf{u}}}(k) = a e^{i\theta_k} \hat{\mathbf{u}}(k) + \varepsilon(k)$, and the expectation of the MSE will be:

$$
\begin{aligned}
\mathbb{E}[\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}})] &= C_1 \sum_{\mathbf{k}} \left(1 - a e^{i\theta_{\mathbf{k}}}\right)^2 ||\hat{\mathbf{u}}(\mathbf{k})||^2 + \sigma^2; \\
&= C_1 \sum_{\mathbf{k}} \left(a^2 + 1 - 2a\mathbb{E}[\cos\theta_{\mathbf{k}}]\right) \cdot ||\hat{\mathbf{u}}(\mathbf{k})||^2 + \sigma^2; \\
&= C_1 \sum_{\mathbf{k}} \underbrace{\{a - \mathbb{E}[\cos\theta_{\mathbf{k}}]\}^2 ||\hat{\mathbf{u}}(\mathbf{k})||^2}_{\text{scaling error}} + \underbrace{\{1 - \mathbb{E}^2[\cos\theta_{\mathbf{k}}]\} ||\hat{\mathbf{u}}(\mathbf{k})||^2}_{\text{translation error}} + \underbrace{\sigma^2}_{\text{noise}},
\end{aligned}
\tag{5}
$$

where C_1 is a constant depend on N and M.

As the last line of equation 5 shows, the MSE has been separated into three components: The *scaling error* reflects the magnitude difference between the prediction and the ground-truth; the *translation error* is caused by the pattern deviation Δ; the *noise* is assumed to be independent of both a and Δ.

The Cause of Amplitude Shrinkage. The scaling error term is highly related to the pattern deviation factor Δ, and the theoretically optimal amplitude scaling will be $a = \mathbb{E}[\cos\theta_{\mathbf{k}}]$. Given the existence of the pattern derivation Δ, we will have $\mathbb{E}[\cos\theta_{\mathbf{k}}] < 1$, causing the shrinkage in the amplitude of the predicted wind speed. Therefore, we also name the *scaling error* as *shrinkage error*.

Moreover, if we assume that Δ is small, then the optimal a will be:

$$
a_{opt} = \mathbb{E}[\cos\theta_{\mathbf{k}}] = 1 - \frac{C_2(\mathbf{k} \cdot \Delta)^2}{2} + o(||\mathbf{k}||^2),
$$

which is decreasing as $\mathbf{k}$ becomes higher and C_2 is a scalar. Therefore, the amplitude shrinkage phenomenon will theoretically tend to be more severe for high-frequency spectrum data.

When trained with mean squared error (MSE), the estimator reduces the squared discrepancy between the prediction and the target, effectively acting on a decomposition of error into translation, scaling, and stochastic noise. If model capacity or optimization is insufficient to avoid the translation component ($\Delta > 0$), gradients can still decrease the objective by attenuating the field's amplitude (i.e., driving $a < 1$). This mechanism explains why general MSE-trained models may underestimate wind-speed amplitudes and dampen short-term variability, thereby degrading performance on extreme-wind prediction. These results yield three practical insights for improving extreme-wind prediction:

- **Upweight scaling error.** Increase the relative weight of the amplitude (scaling) component in the loss to counteract shrinkage.
- **Reduce translation error.** Incorporate mechanisms that explicitly address misalignment Δ so the optimizer need not compensate by damping amplitudes.
- **Frequency-aware weighting.** Reweight residuals by frequency spectrums to mitigate high-frequency attenuation and preserve short-term variability.

3 METHODOLOGY

Guided by the insights from Section 2 on frequency-domain error behavior, we propose a new *gradient-penalized* loss function that mitigates MSE-induced amplitude shrinkage under pattern deviation, and we design a neural framework that combines a *physics-embedded structure* (simplified Navier–Stokes backbone with continuity regularization) and *frequency separation & reweighting* (Fourier masking with band-specific processing and learnable fusion). The model architecture is shown in Figure 1.

3.1 GRADIENT-PENALIZED LOSS FUNCTION

Building on the previous analysis in section 2.3, the amplitude shrinkage phenomena under MSE mainly arise from pattern deviation between the predicted and true wind fields. One idea to solve the problem is to modify MSE by a correction term, which should be insensitive to such deviation, but

capture the field's general spatial change. One of the intuitive approaches is encouraging the norm of the prediction gradient $\|\nabla\tilde{\mathbf{u}}\|$ to match that of the ground-truth $\|\nabla\mathbf{u}\|$. Therefore, we propose our novel *Gradient-Penalized Loss Function* as follows:

$$\mathcal{L}_{\mathrm{gp}}(\tilde{\mathbf{u}}, \mathbf{u}) \;=\; \mathrm{MSE}(\tilde{\mathbf{u}}, \mathbf{u}) \;+\; \lambda \Big|\, \|\nabla\tilde{\mathbf{u}}\|^2 - \|\nabla\mathbf{u}\|^2 \,\Big|. \tag{6}$$

The coefficient $\lambda > 0$ balances pointwise fit against global variation matching: a larger λ more strongly discourages amplitude shrinkage and preserves high-frequency variability; a smaller λ approaches plain MSE.

Connection with error decomposition. Due to the amplitude shrinkage phenomena studied in 2.3, $\|\nabla\tilde{\mathbf{u}}\|^2$ is likely less than $\|\nabla\mathbf{u}\|^2$ in practice. When this happens, minimizing equation equation 6 is equivalent to minimizing the following simplified version:

$$\mathcal{L}_{\mathrm{gp}}(\tilde{\mathbf{u}}, \mathbf{u}) \;=\; \mathrm{MSE}(\tilde{\mathbf{u}}, \mathbf{u}) - \lambda\|\nabla\tilde{\mathbf{u}}\|^2. \tag{7}$$

Applying the Fourier transform to $\nabla\tilde{\mathbf{u}}$ and using the Rayleigh's energy theorem as in 2.3, we get the following correspondence:

$$\|\nabla\tilde{\mathbf{u}}\|^2 \;\propto\; \sum_{\mathbf{k}} \|\mathbf{k}\|^2 \left\|\widehat{\tilde{\mathbf{u}}}(\mathbf{k})\right\|^2. \tag{8}$$

On the other hand, suppose that Δ is sufficiently small, then the decomposition equation equation 5 will also yield

$$\mathbb{E}[\mathrm{MSE}] \;\approx\; C_1 \sum_{\mathbf{k}} (a-1)^2 \|\hat{\mathbf{u}}(\mathbf{k})\|^2 \;+\; C_2^2\, a\, \cos\langle \mathbf{k}, \Delta\rangle \|\Delta\|^2 \|\mathbf{k}\|^2\, \|\hat{\mathbf{u}}(\mathbf{k})\|^2 \;+\; \sigma^2,$$

where the second term is proportional to $\sum_{\mathbf{k}} \cos\langle \mathbf{k}, \Delta\rangle \frac{\|\Delta\|^2}{a} \|\mathbf{k}\|^2\, \|\hat{\tilde{u}}(\mathbf{k})\|^2$, and thus proportional to $\|\nabla\tilde{\mathbf{u}}\|^2$ in equation 8. Therefore, the essential effect of the gradient-penalization can be explained as follows: By tuning λ, the loss $\mathcal{L}_{\mathrm{gp}}$ increases the effective weight on shrinkage error relative to pattern translation error. Consequently, when optimization hits a bottleneck in reducing the translation mismatch, the model will be more likely to optimize on the shrinkage error and thus improve the extreme prediction performance.

3.2 Physics-Embedded Structure

To more effectively reduce the *translation error* highlighted at the end of 2.3, we introduce a physics-embedded structure that leverages the Navier–Stokes (NS) equations as inductive bias. The translation error predominantly stems from uncertainty in the direction and magnitude of the wind-field shift at the next time step. Traditional neural networks do not impose explicit constraints on such pattern transport: they attempt to learn it implicitly via the loss. In contrast, using an (appropriately simplified) physics embedded backbone provides a first-principles estimate of the dominant transport and deformation of the field (e.g., advection and diffusion), yielding a rough but informative pattern forecast. This explicit physical guidance both constrains translation error more directly and reduces the burden on the learnable components, thereby lowering parameter and training costs.

Inspired by equation equation 2, we embed the Navier-Stokes equation into our neural network and name it as *NS Operator*. The operator is decomposed into four components:

1. **Advective Operator**: implements the nonlinear transport term $\mathbf{u} \cdot \nabla\mathbf{u}$.
2. **Viscous Operator**: implements viscous diffusion $\nu\nabla^2\mathbf{u}$ arising from internal friction.
3. **Pressure Operator** : accounts for the pressure-gradient force $\frac{1}{\rho}\nabla P$. Here, we will utilize the pressure data $P[t_1 : t_N]$. However, if the pressure data is not obtainable, we may consider the pressure force as implicit and merge this operator into the Body-Force operator.
4. **Body-Force Operator**(conventional neural networks): because explicit short-term formulations of external forces are often imprecise or unavailable, we model the body force with learnable neural networks that capture dynamics not explained by the above three operators.

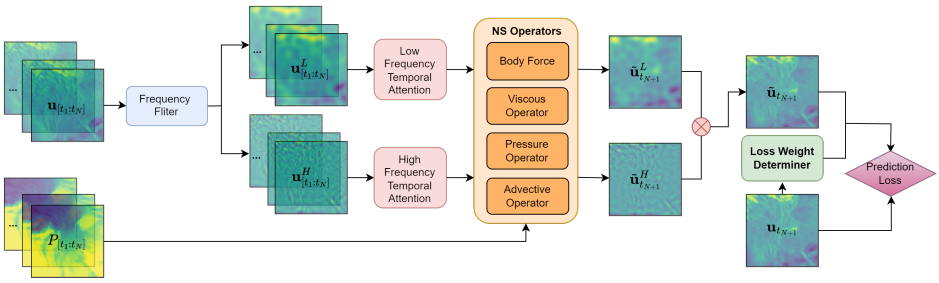

Figure 1: The full architecture of our model. The input data are the wind velocity field $\mathbf{u}$ and the pressure field P. The wind velocity field will be successively processed by the Frequency filter, temporal attention module, and the NS operator to obtain the prediction of high- and low-frequency data. Then the model will combine the two predictions to produce the final prediction results.

Remark that the **body force operator** is equivalent to conventional neural networks and can adopt various structures as other pure data-driven models (but likely contains fewer layers and parameters). During training, the first three operators will learn to generate a rough pre-prediction based on dynamic properties and to ensure the pattern translation magnitude and direction lie in a reasonable range, while the Body-Force operator will learn how to capture the exact wind field dynamics based on the pre-prediction given by the other three operators.

3.3 Frequency Domain Separation and Reweighting

Guided by the insights from section 2.3, we adopt a *frequency-aware weighting* strategy to counter high-frequency attenuation and preserve short-term variability. Concretely, we design a frequency filtering & reweighting scheme that (i) splits the wind field into low- and high-frequency components and (ii) processes and reweights these components respectively so the model can retain rapid, localized dynamics without sacrificing large-scale coherence.

Fourier Filter. We employ a Fourier filter (Alleyne & Cawley, 1991) (Münch et al., 2009) to decompose wind velocity data $\mathbf{u}$ into low-frequency ($\mathbf{u}^L$) and high-frequency ($\mathbf{u}^H$) components. The filter consists of three main steps: 1) Fourier Transform: Converts wind velocity data from the positional domain to the spatial frequency domain. 2) Frequency Masking: Separates high- and low-frequency components using appropriate masks. 3) Inverse Fourier Transform: Transforms the filtered components back to the positional domain. The process is mathematically represented as follows:

$$\hat{\mathbf{u}}(k_x, k_y) = \mathcal{F}(\mathbf{u}) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \mathbf{u}(x, y) \cdot e^{-2\pi i \left(\frac{k_x x}{N} + \frac{k_y y}{M}\right)};$$

$$\hat{\mathbf{u}}_f(k_x, k_y) = \hat{\mathbf{u}}(k_x, k_y) \cdot \mathcal{M}(k_x, k_y);$$

$$\mathbf{u}_f(x, y) = \frac{1}{NM} \sum_{k_x=0}^{N-1} \sum_{k_y=0}^{M-1} \hat{\mathbf{u}}_f(k_x, k_y) \cdot e^{2\pi i \left(\frac{k_x x}{N} + \frac{k_y y}{M}\right)},$$

where $\hat{\mathbf{u}}(k_x, k_y)$ is the Fourier-transformed wind velocity at frequency (k_x, k_y), $M(k_x, k_y)$ denotes the frequency mask (high or low), and $\mathbf{u}_f(x, y)$ refers to the filtered data ($\mathbf{u}_L$ or $\mathbf{u}_H$). This decomposition enables the model to focus on distinct frequency components, enhancing its ability to capture both large-scale trends and rapid, localized variations.

Frequency-Based Temporal Attention To refine the dynamic modeling, we design temporal attention mechanisms for both high- and low-frequency data sequences. Inspired by SENet (Cheng et al., 2016), temporal attention contains two operations: **Squeeze**, which compresses the data of each time slot into a value; and **Excitation**, which produces weight sequences that reflect the relative importance of each time slot for future predictions. The temporal attention is applied at different resolutions for high- and low-frequency components, respectively. Since high-frequency data are more critical for short-term dynamics, they are processed with higher temporal resolution (shorter time

intervals). Conversely, low-frequency sequences, which correspond to long-term trends, are handled at lower temporal resolution. This differentiation ensures that the model effectively captures the unique characteristics of both short-term and long-term dynamics.

4 EXPERIMENTAL RESULTS

We evaluated our approaches through three key experiments: 1. **Effect of Gradient Penalized Loss Function**: Our novel loss function effectively resolves the amplitude shrinkage problem in extreme wind prediction. 2. **Main Prediction Results**: Our model outperformed baselines in both overall accuracy and predictions in extreme wind regions. 3. **Different Frequency Masking Levels**: The results showed that intermediate masking thresholds achieved the best balance between high- and low-frequency information, leading to more accurate predictions.

Data. We evaluate our approaches on meteorological data sampled from the 5th generation of the ECMWF reanalysis (ERA5) database (Hersbach et al., 2020). The dataset includes three key meteorological variables related to wind prediction: the eastward and northward components of 10-meter wind and surface pressure. Guided by the prediction scales stated in Section 2.1, the data of each variable is represented as a time series of two-dimensional latitude–longitude fields over the study region, temporally ordered and co-registered on a common grid. The temporal resolution of the data is 1 hour, while the spatial resolution is 0.25°. For convenience, we define each 24-hour period as a prediction unit, where the first 23 hours are used as inputs to predict the 24th hour.

Baseline Models. To study the effect of the gradient penalized loss function, we utilize the structure of a multivariate meteorological data fusion wind prediction network called MFWPN (Zhang et al., 2025). We further compare our full model with several state-of-the-art regional weather prediction approaches, including CNN, Convolutional LSTM (Tan et al., 2023), and Physics-Informed Neural Network (PINN) (Eivazi et al., 2022). We remark that the PINN model is designed with a revised form of the Navier-Stokes (NS) equations (the Reynolds-averaged Navier-Stokes (RANS) equations) (Alfonsi, 2009).

Evaluation Metrics. We assess the performance of models using *Root Mean Squared Error (RMSE)*, one of the most commonly used metrics for overall predictions. We also evaluate the *Extreme Attentive RMSE (Ex-RMSE)*, which is a modified version of RMSE focusing on regions with extreme wind velocities, calculated as:

$$\text{Ex-RMSE}(\tilde{\mathbf{u}}, \mathbf{u}) = \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot \|\tilde{\mathbf{u}}_i - \mathbf{u}_i\|^2}{\sum_{i=1}^{N} w_i}},$$

where w_i's are determined by the ground-truth data and assign higher weights to extreme wind velocity regions, and $\tilde{u}_i$ and u_i are the predicted and ground truth wind velocities at grid point i. The details of the above metrics can be found in the appendices.

4.1 EFFECT OF GRADIENT PENALIZED LOSS FUNCTION

To quantify the impact of the proposed gradient-penalized objective function, we compare the performance of models trained by equation 6 and MSE over the same baseline structure, and study the impact of different hyperparameter λ on the model performance. The baseline model adopts the same structure as MFWPN (Zhang et al., 2025), which is a machine learning model for short-term wind speed prediction using spatial-temporal fusion and CNN units. We use second-order central differences along both axes to represent spatial gradients. All other settings (optimizer, learning rate, augmentations, and early stopping criteria) are kept identical to the baseline for a fair comparison. The choices of hyperparameters include $\lambda \in \{0, 0.01, 0.02, 0.03, 0.05, 0.07, 0.10, 0.15, 0.20, 0.25\}$.

Results. Figure 2 demonstrates the effectiveness of the gradient penalized loss function and the trade-off between amplitude error and translation error. With a proper λ value, the models trained by gradient-penalized loss outperform the baseline trained by MSE in general. We also observe a

consistent U-shaped curve w.r.t. λ: small positive values markedly reduce extreme attentive error while preserving overall accuracy; too-large values overweight high-frequency residuals and harm stability. In particular, the best performance is achieved at $\lambda = \lambda^{\star}$. Moreover, when $\lambda \geq 0.15$, optimization becomes unstable and the models fail to converge within the prescribed training budget.

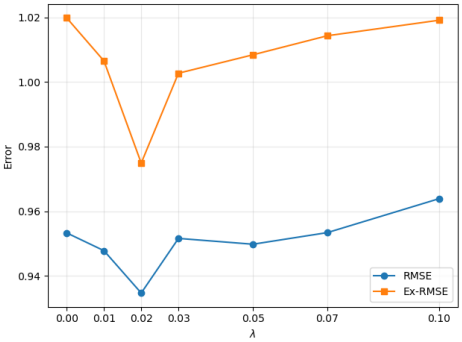

Figure 2: Effect of the gradient-penalized loss across λ values.

The empirical trend aligns with our frequency-domain analysis in Section 2.3 and Section 3.1. The gradient term in equation 6 effectively penalizes the amplitude shrinkage trend and therefore improves the accuracy for extreme wind velocity prediction. However, the penalized term $|\,\|\nabla\tilde{\mathbf{u}}\|^2 - \|\nabla\mathbf{u}\|^2\,|$ itself does not contain any information regarding positional alignment. Therefore, beyond a threshold of λ, the learned model may intend to generate predictions with large spatial fluctuations, regardless of the positional pattern mismatch. This explains the divergence of experimental results for large values of λ.

4.2 Extreme Wind Velocity Prediction

In this section, we evaluate the performance of our final model, which integrates all components in the methodology section with an architecture shown in Figure 1. We compare against several representative baselines on the same regional wind velocity prediction task. The baseline models include: CNN, ConvLSTM, and PINN. We train all models with an initial learning rate of 1×10^{-5} and the SGD optimizer. All approaches consume the same wind-velocity inputs, except that our model additionally utilizes surface pressure P as an auxiliary field.

Results. Table 1 reports both overall prediction errors and extreme wind attentive errors. Our model achieves the best performance on both criteria among all the models tested. Compared to the CNN, ConvLSTM, and PINN baselines, the overall RMSEs of our model decrease by 29.1%, 5.3%, and 16.7%, respectively; while the extreme attentive RMSEs decrease by 41.3%, 18.6%, and 26.5%. These indicating that the gradient-penalized objective and the frequency separation & reweighting are effective for recovering short-horizon, high-wavenumber dynamics.

Model	**RMSE**	**Ex-RMSE**
CNN	0.4639	0.3183
ConvLSTM	0.3471	0.2294
PINN	0.3946	0.2541
NS-Op	0.7061	0.4577
Ours	**0.3287**	**0.1868**

Table 1: Comparative error results across models.

Figure 3 provides a visual comparison of regional wind velocity amplitudes. Compared with the PINN baseline, our predictions exhibit larger and more realistic amplitudes that are closer to the ground truth, particularly in the most extreme zones (highlighted by red boxes). This aligns with our frequency-domain analysis in section 2.3 and the first-stage results in section 4.1.

The above results show our model's better performance for both extreme and overall wind velocity predictions, mitigating the critical amplitude shrinkage problem in extreme weather prediction.

4.3 Different Frequency Masking Level

In this subsection, we investigate how different frequency masking levels affect the model's wind velocity prediction performance using the Fourier Frequency Filter. The results show that excessively high or low masking thresholds degrade accuracy, while optimal performance is achieved at intermediate levels, where a balance between high- and low-frequency information is maintained.

To explore this, we varied the threshold for dividing high- and low-frequency components and analyzed its effect on wind speed prediction accuracy. Experiments were conducted using frequency masking levels of 0.1, 0.3, 0.5, 0.7, and 0.9, which represent the proportion of the highest frequen-

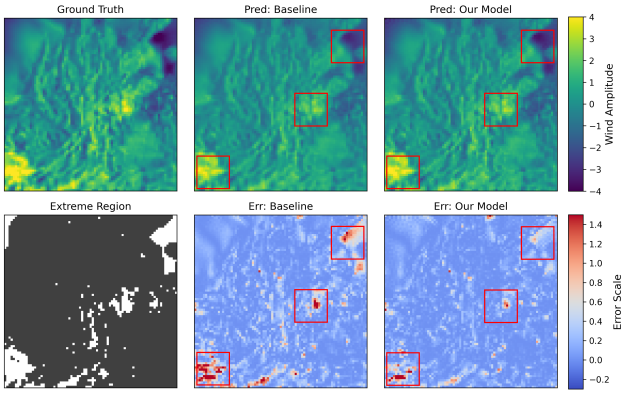

Figure 3: The first line: the ground truth and prediction results of the baseline (PINN) and our models. The first sub-figure in the second line: regions where wind velocity exceeds a specified threshold, highlighted as "Extreme Regions" (in white). The last two sub-figures: comparative prediction errors between our model and the baseline, where bluer indicate lower prediction errors.

cies included in the high-frequency data. These experiments were performed on wind velocity field data from four distinct regions, with the results summarized in 4.

The results demonstrate that both excessively high and excessively low-frequency masking thresholds negatively impact the model's prediction accuracy. When the masking level is too high, critical low-frequency information is excluded, leading to incomplete data representation. Conversely, when the masking level is too low, significant high-frequency details are overlooked, impairing the model's ability to capture rapid variations in wind speed. Optimal prediction performance is achieved when the frequency masking level lies between 0.3 and 0.7, as this range effectively balances the inclusion of high- and low-frequency information, enabling the model to better capture both large-scale and small-scale dynamics.

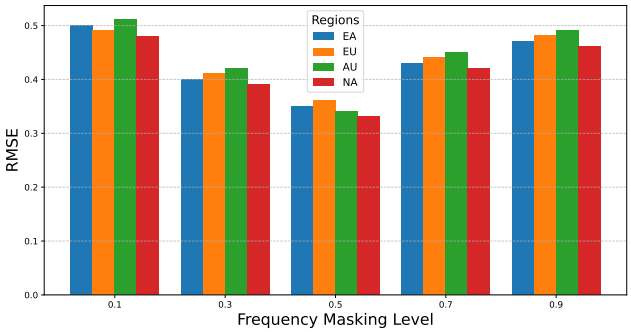

Figure 4: Impact of frequency masking levels on prediction accuracy across regions.

5 Conclusions

We conducted a comprehensive frequency-informed learning to address a key obstacle in wind velocity prediction: the amplitude misalignment (particularly underestimation) of extreme wind velocity prediction. From a frequency-domain perspective, we showed that small spatial pattern deviations combined with standard MSE training induce frequency-dependent amplitude shrinkage, disproportionately suppressing high-frequency components. Guided by this insight, we separated the error in different components and proposed a gradient-penalized loss function that encourages models to emphasize amplitude misalignment. We proposed a frequency separation and reweighting framework with a physics-embedded backbone to further enhance the capture of extreme wind dynamics. Empirically, the proposed methods outperform baselines on regional datasets, significantly improving extreme-wind prediction accuracy while keeping robustness of overall wind prediction.

Limitations Our analysis relies on a simplified assumption on the factors (scaling, shifting, and noise) that cause prediction errors, and a comprehensive study on more complex error-causing factors might be a promising direction. Moreover, generalizations to longer lead times, 3-dimensional scenarios, and cross-region generalization may also be interesting.

REFERENCES

Giancarlo Alfonsi. Reynolds-averaged navier–stokes equations for turbulence modeling. 2009.

David Alleyne and Peter Cawley. A two-dimensional fourier transform method for the measurement of propagating multimode signals. *The Journal of the Acoustical society of America*, 89(3):1159–1168, 1991.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: A review. *Acta Mechanica Sinica*, 37(12): 1727–1738, 2021.

Dongcai Cheng, Gaofeng Meng, Guangliang Cheng, and Chunhong Pan. Senet: Structured edge network for sea–land segmentation. *IEEE Geoscience and Remote Sensing Letters*, 14(2):247–251, 2016.

Jean Coiffier. *Fundamentals of numerical weather prediction*. Cambridge University Press, 2011.

Hamidreza Eivazi, Mojtaba Tahani, Philipp Schlatter, and Ricardo Vinuesa. Physics-informed neural networks for solving reynolds-averaged navier–stokes equations. *Physics of Fluids*, 34(7), 2022.

Song Gao, Peng Zhao, Bin Pan, Yaru Li, Min Zhou, Jiangling Xu, Shan Zhong, and Zhenwei Shi. A nowcasting model for the prediction of typhoon tracks based on a long short term memory neural network. *Acta Oceanologica Sinica*, 37:8–12, 2018.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

James R Holton and Gregory J Hakim. *An introduction to dynamic meteorology*, volume 88. Academic press, 2013.

Jaesung Jung and Robert P Broadwater. Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, 31:762–777, 2014.

Seongchan Kim, Seungkyun Hong, Minsu Joh, and Sa-kwang Song. Deeprain: Convlstm network for precipitation prediction using multichannel radar data. *arXiv preprint arXiv:1711.02316*, 2017.

M. Kunz, S. Mohr, M. Rauthe, R. Lux, and Ch. Kottmeier. Assessment of extreme wind speeds from regional climate models – part 1: Estimation of return values and their evaluation. *Natural Hazards and Earth System Sciences*, 10(4):907–922, 2010. doi: 10.5194/nhess-10-907-2010. URL `https://nhess.copernicus.org/articles/10/907/2010/`.

Wenyuan Li, Zili Liu, Keyan Chen, Hao Chen, Shunlin Liang, Zhengxia Zou, and Zhenwei Shi. Deepphysinet: Bridging deep learning and atmospheric physics for accurate and continuous weather modeling. *arXiv preprint arXiv:2401.04125*, 2024.

Xuelong Li, Kai Kou, and Bin Zhao. Weather gan: Multi-domain weather translation using generative adversarial networks. *arXiv preprint arXiv:2103.05422*, 2021.

Yunjie Liu, Evan Racah, Joaquin Correa, Amir Khosrowshahi, David Lavers, Kenneth Kunkel, Michael Wehner, William Collins, et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*, 2016.

Martine Marion and Roger Temam. Navier-stokes equations: Theory and approximation. *Handbook of numerical analysis*, 6:503–689, 1998.

VP Masson-Delmotte, Panmao Zhai, SL Pirani, C Connors, S Péan, N Berger, Y Caud, L Chen, MI Goldfarb, and Pedro M Scheel Monteiro. Ipcc, 2021: Summary for policymakers. in: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. 2021.

Beat Münch, Pavel Trtik, Federica Marone, and Marco Stampanoni. Stripe and ring artifact removal with combined wavelet—fourier filtering. *Optics express*, 17(10):8567–8591, 2009.

L. Olivetti and G. Messori. Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geoscientific Model Development*, 17(6):2347–2358, 2024. doi: 10.5194/gmd-17-2347-2024. URL https://gmd.copernicus.org/articles/17/2347/2024/.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Elvan P Prasetya and Esmeralda C Djamal. Rainfall forecasting for the natural disasters preparation using recurrent neural networks. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pp. 52–57. IEEE, 2019.

Reza Sabzehgar, Diba Zia Amirhosseini, and Mohammad Rasouli. Solar power forecast for a residential smart microgrid based on numerical weather predictions using artificial intelligence methods. *Journal of Building Engineering*, 32:101629, 2020. ISSN 2352-7102.

Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *Advances in Neural Information Processing Systems*, 36:69819–69831, 2023.

M Vega-Bayo, J Pérez-Aracil, L Prieto-Godino, and S Salcedo-Sanz. Improving the prediction of extreme wind speed events with generative data augmentation techniques. *Renewable Energy*, 221:119769, 2024.

Y H Wan, E Ela, and K Orwig. Development of an equivalent wind plant power-curve: Preprint. 6 2010. URL https://www.osti.gov/biblio/983731.

Yunjun Yu, Junfei Cao, and Jianyong Zhu. An lstm short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access*, 7:145651–145666, 2019.

Rui Zhang, Qingshan Liu, and Renlong Hang. Tropical cyclone intensity estimation using two-branch convolutional neural network from infrared and water vapor images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):586–597, 2019.

Zongwei Zhang, Lianlei Lin, Sheng Gao, Junkai Wang, Hanqing Zhao, and Hangyi Yu. A machine learning model for hub-height short-term wind speed prediction. *Nature Communications*, 16(1): 3195, 2025.

Xiaohui Zhong, Lei Chen, Jun Liu, Chensen Lin, Yuan Qi, and Hao Li. Fuxi-extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Science China Earth Sciences*, pp. 1–13, 2024.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*, 2024.

Qiaomu Zhu, Jinfu Chen, Dongyuan Shi, Lin Zhu, Xiang Bai, Xianzhong Duan, and Yilu Liu. Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. *IEEE Transactions on Sustainable Energy*, 11(1):509–523, 2019.

A Analysis on the Impact of Frequency on Loss function

In this section, we study how the frequency of the Fourier-transformed data affects the MSE loss function. Suppose we discretize the 2-dimensional region by $N \times M$ points (corresponding to longitudinal and latitudinal directions, respectively). Then the discrete Fourier transform of the wind velocity field $\mathbf{u}$ on the region is given by:

$$\hat{\mathbf{u}}(k_x, k_y) = \mathcal{F}(\mathbf{u}) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \mathbf{u}(x, y) \cdot e^{-2\pi i \left(\frac{k_x x}{N} + \frac{k_y y}{M}\right)},$$

where $\hat{\cdot}$ is the Fourier operator, $k = (k_x, k_y)$ are the mode index in the frequency domain. For simplicity, we let $\phi_k(x, y) = e^{-2\pi i \left(\frac{k_x x}{N} + \frac{k_y y}{M}\right)}$ denote the Fourier basis. Then the inverse transform can be written by:

$$\mathbf{u}(x, y) = \frac{1}{NM} \sum_k \hat{\mathbf{u}}(k) \phi_k.$$

Suppose we have a prediction of $\mathbf{u}$ denoted by $\tilde{\mathbf{u}}$. The mean square error (MSE) of the prediction is given by $\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}}) = \frac{1}{NM} \sum_{x,y} |\mathbf{u} - \hat{\mathbf{u}}|^2$ By the discrete orthogonality of the Fourier basis ϕ_k's, we have $\sum_{x,y} \phi_k \overline{\phi_{k'}} = NM \delta_{k,k'}$. Therefore, we can derive the equivalent form of the MSE in the frequency domain:

$$\begin{aligned}
\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}}) =& \frac{1}{NM} \sum_{x,y} |\mathbf{u} - \hat{\mathbf{u}}|^2 \\
=& \frac{1}{NM} \sum_{x,y} \left| \frac{1}{NM} \sum_k \left(\hat{\mathbf{u}}(k) - \hat{\tilde{\mathbf{u}}}(k) \right) \phi_k(x, y) \right|^2 \\
=& \frac{1}{NM} \sum_{x,y} \frac{1}{(NM)^2} \sum_k \sum_{k'} \left(\hat{\mathbf{u}}(k) - \hat{\tilde{\mathbf{u}}}(k) \right) \cdot \overline{\left(\hat{\mathbf{u}}(k') - \hat{\tilde{\mathbf{u}}}(k') \right)} \phi_k(x, y) \, \overline{\phi_{k'}(x, y)} \\
=& \frac{1}{(NM)^3} \sum_k \sum_{k'} \left(\hat{\mathbf{u}}(k) - \hat{\tilde{\mathbf{u}}}(k) \right) \cdot \overline{\left(\hat{\mathbf{u}}(k') - \hat{\tilde{\mathbf{u}}}(k') \right)} \underbrace{\sum_{x,y} \phi_k(x, y) \, \overline{\phi_{k'}(x, y)}}_{NM \, \delta_{k,k'}} \\
=& \frac{1}{(NM)^2} \sum_k \left| \hat{\mathbf{u}} - \hat{\tilde{\mathbf{u}}} \right|^2.
\end{aligned}$$

Now, let's analyze the potential impact of frequency on MSE. Statistically, the prediction of the wind velocity field can be modeled as a translation of the ground truth plus white noise:

$$\tilde{\mathbf{u}}(\mathbf{x}) = \mathbf{u}(\mathbf{x} + \Delta) + \varepsilon(\mathbf{x}),$$

where $\mathbf{x} = (x, y)$; $\Delta = (\Delta_x, \Delta_y)$; $\varepsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2)$ is a Gaussian white noise that may or may not be invariant with respect to $\mathbf{x}$. It should be clarified that the real prediction scenario can be much more complicated than this formula. However, it will be extremely difficult or even impossible to study the real cases in detail. Besides, this simplification is enough to show some insights about the impact of frequency on the loss function. We denote $\theta_k = 2\pi \left(\frac{k_x \Delta_x}{N} + \frac{k_y \Delta_y}{M} \right)$, then the Fourier transform of the prediction is:

$$\hat{\tilde{\mathbf{u}}}(k) = e^{i\theta_k} \hat{\mathbf{u}}(k) + \varepsilon(k).$$

Therefore, the mean square error of the prediction is:

$$\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}}) = \frac{1}{(NM)^2} \sum_k \left| (1 - e^{i\theta_k}) \hat{\mathbf{u}}(k) - \hat{\varepsilon}(k) \right|^2.$$

Since ε is Gaussian white noise, we have $\mathbb{E}[\hat{\varepsilon}] = 0$ and $\mathbb{E}[\hat{\varepsilon}(k) \overline{\hat{\varepsilon}(k')}] = NM\sigma^2 \delta_{k,k'}$. Therefore, the expectation of the MSE will be:

$$\begin{aligned}
\mathbb{E}[\text{MSE}(\mathbf{u}, \tilde{\mathbf{u}})] =& \frac{1}{(NM)^2} \sum_k \left| (1 - e^{i\theta_k}) \right|^2 |\hat{\mathbf{u}}(k)|^2 + \sigma^2. \\
=& \frac{1}{(NM)^2} \sum_k 4 \sin^2(\frac{\theta_k}{2}) \cdot |\hat{\mathbf{u}}(k)|^2 + \sigma^2.
\end{aligned}$$

Suppose the magnitude of the translation $|\Delta|$ is sufficiently small such that $\theta_k \ll 1$. Then we can approximate the $\sin(\theta_k/2)$ term by $\theta/2$:

$$\begin{aligned}\mathbb{E}[\text{MSE}] &\approx \frac{1}{(NM)^2}\sum_k \theta_k^2 \left|\hat{\mathbf{u}}(k)\right|^2 + \sigma^2 \\ &\approx C\sum_k |k|^2 \cdot \left|\hat{\mathbf{u}}\right|^2 + \sigma^2,\end{aligned}$$

where C is constant.

According to the above formula, we can separate the MSE into two parts: one is attributed to the translation of the overall wind velocity field, and the other is attributed to the essential error from the white noise. The error from translation is approximately proportional to the square of the magnitude of the frequency.

The above analysis provides us with insights that the high-frequency data is more likely to affect the MSE.

B Maximum Likelihood Estimation Analysis of Weighted MSE

In this section, we continued to utilize the simplified "translation + noise" model for analysis. As mentioned in Section 2, the variances of the noise are bot likely be constant among the whole region. As before, suppose

$$\tilde{\mathbf{u}}(\mathbf{x}) \;=\; \mathbf{u}(\mathbf{x}+\Delta) + \varepsilon(\mathbf{x}), \qquad \varepsilon(\mathbf{x}) \sim \mathcal{N}(0,\Sigma).$$

We can use a shift operator $S(\Delta)$ to represent the translation: $\tilde{\boldsymbol{u}} \;=\; S(\Delta)\,\boldsymbol{u} \;+\; \boldsymbol{\varepsilon}$. In this case, we consider the $\theta = (\Sigma, \Delta)$ as the parameter, the prediction and the ground-truth as the data $\mathcal{D}$. Since the noise is Gaussian, the likelihood can be written as follows:

$$p(\mathcal{D} \mid \theta) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}\,(\tilde{\boldsymbol{u}} - S(\Delta)\boldsymbol{u})^\top \Sigma^{-1} (\tilde{\boldsymbol{u}} - S(\Delta)\boldsymbol{u})\right).$$

To maximize the above likelihood function is equivalent to minimizing the negative log-likelihood function $NLL(\Sigma, \Delta)$:

$$NLL(\Sigma,\Delta) = -\log p(\mathcal{D} \mid \theta) = \tfrac{1}{2}\,(\tilde{\boldsymbol{u}} - S(\Delta)\boldsymbol{u})^\top \Sigma^{-1} (\tilde{\boldsymbol{u}} - S(\Delta)\boldsymbol{u}) \;+\; \tfrac{1}{2}\log|\Sigma|.$$

The first part is equivalent to the weighted MSE with a weight $W = \Sigma^{-1}$.

Practically, we may have different requirements for the weather system prediction. For example, in our study of extreme wind speed prediction, we required more precision for the regions with higher magnitude of wind speed. Therefore, we hope the variance of predictive noise become smaller, so that our model will produce more concise prediction among these regions. By the relation of the weight and covariance $W = \Sigma^{-1}$, the points with small variance should correspond to heavier weight in general, which is also intuitively true since we need to assigned more weight to the region with larger wind speed.

It should be noticed that the above two analyses are mainly based on a simplified scenario of the prediction. They bring precious insights into our model design, but they are not responsible for the rigorous proof of the real-world prediction scenarios, which is too complicated for theoretical proof.

C Related Works

C.1 Numerical and Data-driven Weather Prediction

Numerical Weather Prediction (NWP) (Bauer et al., 2015) represents traditional physical weather forecasting methods, which rely on computational techniques to solve the physical equations governing atmospheric dynamics. For wind velocity prediction, the Navier-Stokes equations and the continuity equation are pivotal in describing the wind velocity field dynamics within a region. NWP models discretize these equations over a computational grid and solve them numerically using methods such as finite difference, finite volume, or spectral techniques. Despite their widespread use,

NWP models face significant limitations, including a reliance on precise initialization data and high computational costs. These challenges make real-time predictions and extreme weather scenario forecasting particularly difficult.

In contrast, purely data-driven models leverage machine learning algorithms to predict wind speed by identifying patterns in historical data. Examples include CNNs (Liu et al., 2016), LSTMs (Yu et al., 2019), ConvLSTMs (Kim et al., 2017), GANs (Li et al., 2021), and transformers (Bi et al., 2022). These models excel at capturing complex wind patterns and local variations, demonstrating flexibility and adaptability in learning from data. However, they lack the physical constraints required to ensure realistic predictions, which can sometimes result in unreasonable outputs.

C.2 Physics-Informed Data-driven Weather Prediction

Recent advancements in weather prediction have introduced hybrid approaches that integrate physical laws with machine learning. For example, Physics-Informed Neural Networks (PINNs) (Cai et al., 2021) incorporate differential equations into the training process to enforce physical realism. These methods reduce dependency on large datasets and computational resources, ensure predictions adhere to known physical laws, and enhance robustness in complex environments.

A notable example of this approach is DeepPhysiNet (Li et al., 2024), developed by W. Li et al. This model combines physics-guided machine learning with weather prediction by constructing physics networks based on multilayer perceptrons for meteorological variables. Partial Differential Equations are incorporated as part of the loss function, while a hyper-network based on deep learning directly learns weather patterns, contributing to the weights of the physics networks. This hybrid design ensures both physical consistency and the ability to capture intricate weather patterns.

C.3 Extreme Weather Prediction

In this paper, extreme weather refers to the outlying values of specific weather properties. For example, wind power plants require precise predictions of wind speeds at turbine locations, particularly the extreme values of wind speeds (Sabzehgar et al., 2020). According to $\text{osti}_9 83731, when wind speeds exceed the cut-outspeed, wind turbines cease operation to prevent damage. As such, accurate regional wind speed forecasting, especially$

Recent advancements in machine learning have significantly improved predictive capabilities for extreme weather conditions. For instance, Fuxi Extreme, developed by X. Zhong et al. (Zhong et al., 2024), leverages a Denoising Diffusion Probabilistic Model (DDPM) to enhance accuracy and detail in extreme weather predictions. This model combines a base weather prediction framework with DDPM, capturing fine-scale features through a two-step process: adding noise in a forward step and refining details in a reverse denoising step. This innovative approach has demonstrated exceptional accuracy and detail restoration, making it highly effective for forecasting extreme weather conditions.

D ERA5 Dataset

We mainly used the ERA5 datasets for our model training and testing processes. The ERA5 datasets(Hersbach et al., 2020), developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), is a fifth-generation reanalysis of the climate and weather covering data from 1940 to the present. Although the datasets contain detailed reanalysis data globally, it provides flexibility to select and obtain data in rectangular spatial region in different scales and locations. Therefore, the datasets is suitable for studying our work on regional weather prediction. This datasets is created through data assimilation, which combines model data with observations from various sources worldwide, resulting in a globally consistent and comprehensive datasets. ERA5 provides hourly estimates for a wide range of atmospheric, ocean-wave, and land-surface variables, including uncertainty estimates using a 10-member ensemble at three-hour intervals. The data is available on a regular latitude-longitude grid, with a horizontal resolution of $0.25^\circ \times 0.25^\circ$ for atmospheric reanalysis. The temporal resolution of ERA5 is hourly, and the data is accessible in GRIB format, providing high-resolution information for many climate and weather applications.

In this study, we focus on specific variables from the ERA5 dataset relevant to wind speed prediction, namely the 10-meter wind components and surface pressure. The 10-meter u-component of wind represents the eastward component of horizontal wind speed at 10 meters above ground level, while the v-component represents the northward component at the same height. These components are measured in meters per second (m/s) and can be combined to calculate the speed and direction of the horizontal wind. Surface pressure, given in Pascals (Pa), is the atmospheric pressure at the Earth's surface, which reflects the weight of the air column above a specific point. These parameters together provide essential information for modeling and predicting wind dynamics in the atmosphere.

E Impact of Frequency on Temporal Data

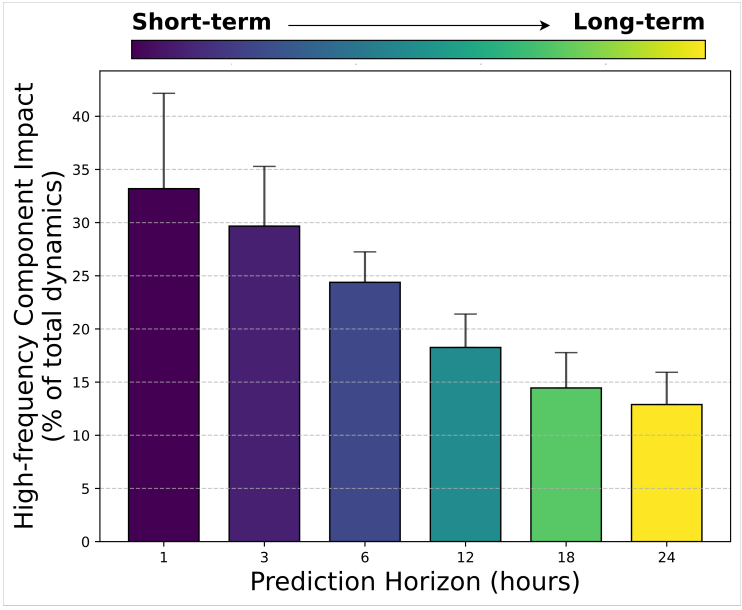

Figure 5: Impact of high-frequency components on wind field dynamics decreases as prediction horizon extends from 1 to 21 hours.

We conducted preliminary experiments to investigate how high- and low-frequency components of wind data contribute to future dynamics across different temporal scales. Using Fourier filtering techniques (detailed in Section 4), we decomposed the two-dimensional wind velocity field time series into their respective frequency components.

Our correlation analysis examined the relationship between these decomposed components and actual future wind patterns across prediction horizons ranging from 1 to 24 hours. The results, illustrated in *Figure 1*, reveal a clear temporal dependency pattern: For longer prediction horizons (approaching 24 hours), low-frequency components demonstrate dominant predictive power in wind speed pattern evolution. Conversely, at shorter intervals (approaching 1 hour), high-frequency components become increasingly significant in determining wind pattern changes.

This observation can be theoretically explained as follows. Suppose $\mathbf{u}(x, y)$ is the wind velocity field. Then the Fourier transform of the spatial gradient of wind velocity $\nabla\mathbf{u}$ is:

$$\widehat{\nabla\mathbf{u}}(\mathbf{k}) = i\mathbf{k} \iint u(x, y) e^{-i2\pi(k_x x + k_y y)}\, dx\, dy,$$

where $k = (k_x, k_y)$ represents frequency domain coordinates and $\hat{\cdot}$ denotes the Fourier transform. This relationship demonstrates that higher frequencies (larger $|k|$) correspond to larger spatial gradients ($\|\nabla u\|$). Consequently, high-frequency components capture small-scale features characterized by sharp gradients and abrupt changes in the wind velocity field—characteristics typically associated with turbulence and extreme weather events.

F Details of Evaluation Metrics

F.0.1 Root Mean Squared Error (RMSE)

The RMSE quantifies the overall accuracy of the predicted wind velocity field by measuring the difference between the predicted and ground truth values. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{u}}_i - \mathbf{u}_i\|^2},$$

where:

- $\hat{\mathbf{u}}_i$ and $\mathbf{u}_i$ are the predicted and ground truth wind velocity vectors at the i-th grid point,
- N is the total number of grid points.

F.0.2 Extreme Region Error (ExtremeErr)

The Extreme Region Error (ExtremeErr) focuses on the model's accuracy in predicting extreme weather regions, characterized by high wind velocities. It assigns larger weights to regions with extreme wind values to emphasize their importance. Mathematically, it is defined as:

$$\text{ExtremeErr} = \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot \|\hat{\mathbf{u}}_i - \mathbf{u}_i\|^2}{\sum_{i=1}^{N} w_i}},$$

where:

- w_i is the weight assigned to the i-th grid point, with higher values for extreme wind velocity regions,
- $\hat{\mathbf{u}}_i$ and $\mathbf{u}_i$ are the predicted and ground truth wind velocity vectors at the i-th grid point.

These metrics collectively assess the model's accuracy, adherence to physical principles, and capability to predict extreme weather conditions effectively.