W2P-FM: LEARNING TRANSFERABLE WEATHER-TO-POWER REPRESENTATIONS FOR WIND POWER FORECASTING ACROSS SITES AND TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

As wind power expands globally, accurate forecasting is critical for managing weather-driven generation variability and enabling its reliable integration into modern energy systems. Yet, continued wind-farm deployment and diverse forecasting demand across horizons and temporal granularities expose a fundamental limitation of conventional data-driven models, which are typically optimized on specific sites and tasks. Meeting these requirements calls for transferable modeling of weather-driven power dynamics while addressing three key challenges: aligning multi-source weather and power data, capturing cross-site heterogeneity, and supporting diverse forecasting tasks within a unified framework. To tackle these challenges, we propose W2P-FM, a foundation model for wind power forecasting pretrained on more than 20 million time points. W2P-FM constructs a uniform temporal-variable representation by using MeteoRouter to adaptively fuse numerical weather predictions and granularity-aware patching to align power and meteorological sequences. It further incorporates rich static site context and recent response observations to characterize site-specific patterns. Finally, weather-to-power representations are learned through intra-patch cross-variable interactions and local multi-scale inter-patch aggregation, with task-specific heads supporting different forecasting settings. Experiments on 33 real-world datasets comprising over 1,000 site-level wind power series across five continents show that W2P-FM consistently outperforms state-of-the-art forecasting models, reducing MSE by 12.3% in full-shot and 48.9% in zero-shot settings. Frozen-backbone adaptation reduces MSE by 17.2% in cross-granularity forecasting and CRPS by 15.8% in probabilistic forecasting relative to the strongest task-specific baselines.

1 INTRODUCTION

The sustained worldwide deployment of wind power generation is reshaping electricity supply and advancing the transition toward low-carbon power systems (Jakhmola et al., 2026). As wind power penetration increases, the strong dependence of power output on evolving atmospheric conditions introduces substantial variability and limited controllability, complicating system dispatch and electricity-market operation (Grams et al., 2017; Gonzalez et al., 2023). Accurate wind power forecasting is consequently essential for coordinating generation schedules and reserves, improving market decisions, and maintaining secure power-system operation (Terrén-Serrano et al., 2026).

Meanwhile, expanding wind deployment has created sites with markedly different data availability, from established farms with extensive operational records to newly commissioned sites lacking sufficient training data (Meng et al., 2026). Forecasting methods must therefore exploit site-specific supervision where available while generalizing zero-shot to unseen sites (Wang et al., 2026; Zhang et al., 2026). Moreover, as shown in Figure 1, downstream tasks differ in horizon, temporal granularity and output formulation (Bazionis et al., 2022). Forecasts with a temporal resolution of 10 or 15 minutes support real-time dispatch and intraday trading, whereas hourly day-ahead forecasts inform unit commitment and market bidding (Dai & Qiao, 2013; Li et al., 2025). Beyond point estimates, probabilistic forecasts quantify predictive uncertainty to support risk-aware reserve allocation and decision-making (Matos & Bessa, 2011). Collectively, wind power forecasting increasingly calls for a unified framework capable of serving diverse deployment scenarios and task requirements.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Wind power forecasting has advanced substantially in both research and operational deployment, evolving from statistical and feature-engineered machine-learning methods to deep models that learn weather-power dependencies and temporal dynamics directly from data (Wang et al., 2021; Yang et al., 2024). Although effective in specific forecasting settings, these models often transfer poorly across sites and downstream tasks (Tasnim et al., 2018; Dong et al., 2023). Time-series foundation models, including Chronos (Ansari et al., 2024) and Moirai (Woo et al., 2024), address this limitation by pretraining on large-scale temporal corpora to learn transferable representations for zero-shot forecasting across domains, datasets, and tasks (Liang et al., 2024; Ye et al., 2027). This paradigm has inspired wind power forecasting foundation models, including WindFM (Fan et al., 2025) and Tyan-WP (Huang et al., 2026). Nevertheless, their reliance on simulated data (Draxl et al., 2015) and omission of future NWP inputs may limit generalization across real-world sites.

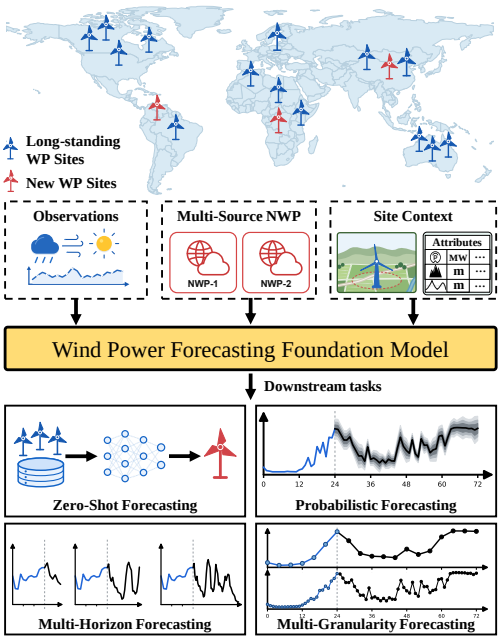

Figure 1: Diverse deployment scenarios and task requirements in wind power forecasting. A unified framework should enable robust zero-shot generalization to unseen sites and efficient adaptation to downstream tasks.

Motivated by these observations, we identify and address three key challenges in developing a foundation model for wind power forecasting: **Challenge 1: Unified Representation Learning from Heterogeneous Multi-Source Wind-Power Data.** Wind power forecasting combines multi-resolution power and meteorological observations with multi-source NWP whose reliability and availability vary across variables, regions, and lead times (Yan et al., 2022). Existing methods rarely reconcile this heterogeneity, causing resolution-dependent distortion and underused NWP guidance. The challenge is to align these inputs within a temporally unified representation while preserving their distinct predictive roles. **Challenge 2: Cross-Site Generalization under Zero-Shot Settings.** Identical meteorological forcing can induce markedly different power responses across sites due to variations in terrain, capacity, and local environmental conditions (Prakash et al., 2024). This heterogeneity hinders zero-shot transfer to data-scarce or poorly represented sites. The challenge is to integrate static context and recent weather-power observations into site-conditioned representations that generalize across regions and to unseen sites. **Challenge 3: Task Adaptability Across Diverse Forecasting Requirements.** Forecasting applications differ in horizon, temporal granularity, and output formulation. Although these tasks share underlying knowledge, their input-output structures and objectives typically require separate training or redesign. The challenge is to decouple transferable response representations from task-specific interfaces, enabling lightweight adaptation without sacrificing accuracy.

Our Approach. To address these challenges, we propose **W2P-FM**, a weather-to-power foundation model for global wind power forecasting. Specifically, W2P-FM first produces robust meteorological guidance through availability-aware fusion of multi-source NWP and encodes multi-resolution power and meteorological sequences into a unified temporal-variable token grid. On this basis, the resulting patch tokens are conditioned on static site context and augmented with site-specific response patterns distilled from recent observations. Within the Variable-wise Multi-scale Transformer, weather-to-power representation learning is factorized into intra-patch cross-variable interaction and local multi-scale inter-patch aggregation. Finally, task-specific heads decode future power tokens into deterministic or probabilistic forecasts across horizons and temporal granularities. Our main contributions are summarized as follows:

- We propose W2P-FM, a domain-specific foundation model for generalizable wind power forecasting across sites, horizons, temporal granularities, and output formulations.
- W2P-FM introduces a dedicated architecture that combines availability-aware multi-source NWP fusion and granularity-aware tokenization with static context conditioning and historical response

encoding. Furthermore, a Variable-wise Multi-scale Transformer is designed to efficiently learn site-conditioned weather-to-power representations by integrating intra-patch cross-variable interactions with local multi-scale inter-patch aggregation.

- Extensive experiments on 1,012 real-world sites show that W2P-FM consistently outperforms representative baselines in both full-shot and zero-shot settings. Strong adaptation to fine-grained and probabilistic forecasting further validates the transferability of its learned representations.

2 Related work

Wind Power Forecasting. Accurate wind power forecasting is essential for power-system scheduling and electricity-market operation (Dai & Qiao, 2013; Saint-Drenan et al., 2020). Traditional statistical and classical machine-learning methods remain widely used for their computational efficiency and strong local performance, but often rely on hand-engineered covariates and manually selected weather products (Wang et al., 2021; Yang et al., 2024; Bazionis et al., 2022; Ke et al., 2017). Recent models based on RNNs, CNNs, MLPs, and Transformers have advanced the modeling of weather-power dependencies and temporal patterns, while NWP-driven and probabilistic approaches further exploit meteorological guidance and quantify predictive uncertainty (Xu et al., 2025; Wang et al., 2025; Matos & Bessa, 2011; Terrén-Serrano et al., 2026). Nevertheless, most existing methods remain specialized to particular sites, tasks, or weather products, limiting their scalability, transferability, and robustness across diverse operational settings (Dong et al., 2023).

Time-Series Foundation Models. Recent research has moved from task-specific models toward time-series foundation models (TSFMs) pretrained on large-scale heterogeneous corpora to improve scalability and cross-domain generalization (Liang et al., 2024; Ye et al., 2027). Early TSFMs, including Timer (Liu et al., 2024d), TimesFM (Das et al., 2023), and Chronos (Ansari et al., 2024), focused primarily on univariate autoregressive forecasting over patched or discretized sequences, whereas subsequent models such as Chronos-2 (Ansari et al., 2025), TiRex-2 (Podest et al., 2026), Time-MoE (Shi et al., 2025), and Moirai-MoE (Liu et al., 2024b) extended this paradigm to multivariate and covariate-informed forecasting, expert specialization, in-context adaptation, and probabilistic generation. Inspired by these advances, wind power forecasting foundation models have begun to emerge. WindFM (Fan et al., 2025) and Tyan-WP (Huang et al., 2026) explore domain-specific pretraining and probabilistic forecasting, but their reliance on simulated WIND Toolkit data and omission of future NWP may limit predictive fidelity and real-world applicability.

3 Preliminaries

Wind Power Series. Wind power series represent electrical generation from turbines or wind farms. For sample i from site g_i, let H_{in} and H_{out} denote the look-back and forecast horizons in hours. For power measurements at different temporal granularities, n_i denotes the number of observations per hour. The historical and future power sequences are defined as $\boldsymbol{x}_i \in \mathbb{R}^{n_i H_{\text{in}}}$ and $\boldsymbol{y}_i \in \mathbb{R}^{n_i H_{\text{out}}}$.

Historical Meteorological Observations. Let C denote the number of meteorological variables. For sample i, the hourly historical meteorological data is denoted by $\boldsymbol{X}_i^{\text{met}} \in \mathbb{R}^{H_{\text{in}} \times C}$.

Multi-Source NWP. Suppose there are S NWP sources. For source $s \in \{1, \ldots, S\}$, the hourly meteorological forecast for sample i is denoted by $\mathbf{F}_i^{(s)} \in \mathbb{R}^{H_{\text{out}} \times C}$. The collection of NWP inputs is denoted by $\mathcal{F}_i = \{\mathbf{F}_i^{(s)}\}_{s=1}^{S}$. As some NWP products may be unavailable, we define a vector $\boldsymbol{m}_i^{\text{nwp}} \in \{0,1\}^S$, where $m_{i,s}^{\text{nwp}} = 1$ indicates that source s is available. This study uses $S = 3$ NWP sources, namely GFS, EC, and Tianji (see Appendix C.2 for details and references).

Static Site Information. Static site information describes persistent factors affecting wind generation, including equipment, terrain, and the surrounding environment. We represent it using tabular features $\boldsymbol{s}_{g_i}^{\text{tab}} \in \mathbb{R}^{d_{\text{tab}}}$, satellite imagery $\mathcal{I}_{g_i} \in \mathbb{R}^{K \times 3 \times H_{\text{img}} \times W_{\text{img}}}$, and an AlphaEarth embedding $\boldsymbol{a}_{g_i} \in \mathbb{R}^{64}$ (Brown et al., 2025). These modalities are collected as $\mathcal{S}_i = \{\boldsymbol{s}_{g_i}^{\text{tab}}, \mathcal{I}_{g_i}, \boldsymbol{a}_{g_i}\}$.

Problem Statement. Given the inputs defined above, our goal is to learn a unified foundation model that forecasts future wind generation across sites, horizons, and temporal granularities. For deterministic forecasting, the forecast function is formulated as

$$\hat{\boldsymbol{y}}_i = f_\theta(\boldsymbol{x}_i, \mathbf{X}_i^{\text{met}}, \mathcal{F}_i, \boldsymbol{m}_i^{\text{nwp}}, \mathcal{S}_i), \tag{1}$$

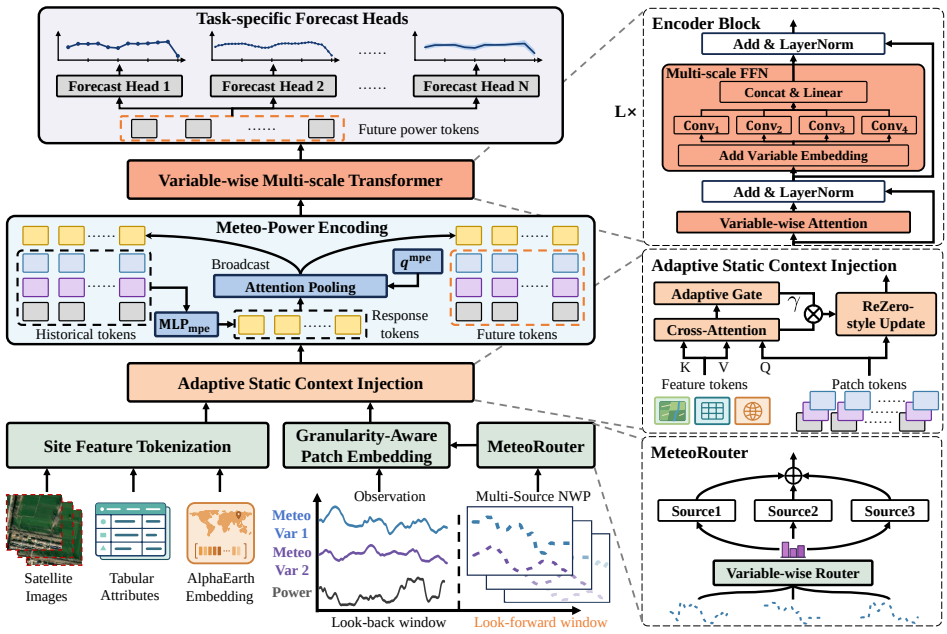

Figure 2: Overall architecture of W2P-FM. Heterogeneous inputs are first unified by MeteoRouter and Granularity-Aware Patch Embedding. Adaptive Static Context Injection and Meteo-Power Encoding further integrate site-specific context. Finally, a Variable-wise Multi-scale Transformer backbone is employed to learn weather-to-power representations for task-specific forecasting heads.

where $\hat{\boldsymbol{y}}_i \in \mathbb{R}^{n_i H_{\text{out}}}$ denotes the predicted future power sequence. For probabilistic forecasting, the model estimates a predictive distribution at each forecast step,

$$\widehat{\mathcal{P}}_i = f_\theta^{\text{prob}}(\boldsymbol{x}_i, \mathbf{X}_i^{\text{met}}, \mathcal{F}_i, \boldsymbol{m}_i^{\text{nwp}}, \mathcal{S}_i) = \{\widehat{P}_{i,\tau}\}_{\tau=1}^{n_i H_{\text{out}}}, \tag{2}$$

where $\widehat{P}_{i,\tau}$ denotes the predictive distribution of wind power at future target step τ.

4 Methodology

As illustrated in Figure 2, W2P-FM uses Granularity-Aware Patch Embedding to encode multi-resolution power and meteorological sequences into a unified temporal-variable token grid, with future NWP trajectories adaptively fused by MeteoRouter. The patch tokens are then conditioned on static site context and further augmented with recent site-specific response patterns distilled by Meteo-Power Encoding (MPE). The Variable-wise Multi-scale Transformer subsequently learns weather-to-power representations through intra-patch cross-variable interaction and local multi-scale inter-patch aggregation. Finally, task-specific forecast heads decode the future power tokens into point forecasts or conditional quantiles at the target horizon and temporal granularity. To simplify notation, the sample index i is suppressed when it is unambiguous.

4.1 Heterogeneous Input Unification

MeteoRouter. Multi-source NWP varies in availability and reliability across sites, variables, and instances. W2P-FM therefore introduces MeteoRouter, a variable-wise routing module that adaptively fuses available sources into future weather trajectories tailored to wind power forecasting.

Given the NWP matrix $\mathbf{F}_i^{(s)}$ from source s, we denote the future trajectory of variable c as $\boldsymbol{f}_c^{(s)} \in \mathbb{R}^{H_{\text{out}}}$. We first project each source-variable trajectory into the hidden space: $\boldsymbol{u}_{s,c} = \phi_{\text{met}}(\boldsymbol{f}_c^{(s)}) \in \mathbb{R}^D$, where D is the hidden dimension, and $\phi_{\text{met}} : \mathbb{R}^{H_{\text{out}}} \to \mathbb{R}^D$ is a linear projection shared across sources and variables. Given a learnable routing query $\boldsymbol{q}_c^{\text{met}} \in \mathbb{R}^D$ specific to meteorological variable c, MeteoRouter computes the availability-aware routing weights and fused trajectory as

$$\alpha_{s,c} = \text{Softmax}_{s:\, m_{i,s}^{\text{nwp}}=1}\left(\frac{\langle \boldsymbol{q}_c^{\text{met}}, \boldsymbol{u}_{s,c}\rangle}{\sqrt{D}}\right), \quad \widetilde{\boldsymbol{f}}_c = \sum_{s:\, m_{i,s}^{\text{nwp}}=1} \alpha_{s,c} \boldsymbol{f}_c^{(s)}. \tag{3}$$

The trajectories are assembled into the fused NWP matrix $\widetilde{\mathbf{F}} = [\widetilde{\boldsymbol{f}}_1, \ldots, \widetilde{\boldsymbol{f}}_C] \in \mathbb{R}^{H_{\text{out}} \times C}$. This variable-wise routing adaptively weights reliable sources without manually specified priorities.

Granularity-Aware Patch Embedding. Conventional patching uses a fixed number of observations per patch (Nie et al., 2022). W2P-FM adjusts this number according to the sampling resolution while keeping the patch duration fixed, thereby producing temporally aligned token representations for multi-resolution power and meteorological sequences. Let L_{p} denote the patch duration in hours. The numbers of historical and future patches are $T_{\text{in}} = \left\lceil \frac{H_{\text{in}}}{L_{\text{p}}} \right\rceil$, $T_{\text{out}} = \left\lceil \frac{H_{\text{out}}}{L_{\text{p}}} \right\rceil$, and $T = T_{\text{in}} + T_{\text{out}}$.

For each time step and variable, we augment the raw value with two auxiliary indicators: a time marker and a class marker. The time marker encodes the relative position with respect to the forecast origin, where historical positions take negative normalized values and future positions take positive normalized values. The class marker distinguishes historical observations, future NWP inputs, future power placeholders, and invalid padded positions.

Since future power is unavailable, we append zero placeholders to the historical sequence: $\bar{\boldsymbol{x}} = [\boldsymbol{x}_i; \mathbf{0}_{n_i H_{\text{out}}}] \in \mathbb{R}^{n_i(H_{\text{in}}+H_{\text{out}})}$. The power sequence is stacked with the corresponding markers and then mapped to patch tokens using a 1D convolution along the temporal dimension:

$$\mathbf{Z}^{\text{pow}} = \text{Reshape}(\text{Conv}_{\text{pow}}^{(n_i)}(\text{Stack}(\bar{\boldsymbol{x}}, \boldsymbol{r}^{\text{time}}, \boldsymbol{r}^{\text{class}}))) \in \mathbb{R}^{T \times 1 \times D}, \tag{4}$$

where $\boldsymbol{r}^{\text{time}}$ and $\boldsymbol{r}^{\text{class}}$ are the time and class markers for the power variable, and $\text{Conv}_{\text{pow}}^{(n_i)}$ denotes the power patch projection associated with sampling rate n_i. The projection output is reshaped to introduce a singleton variable dimension for concatenation with meteorological tokens.

For meteorological variables, all historical observations and future NWP inputs are represented at hourly resolution. Separate 1D convolutional patch projections are used for historical meteorological observations and future NWP forecasts. The corresponding computations are as follows:

$$\mathbf{Z}^{\text{hist}} = \text{Conv}_{\text{hist}}(\text{Stack}(\mathbf{X}_i^{\text{met}}, \mathbf{R}_{\text{hist}}^{\text{time}}, \mathbf{R}_{\text{hist}}^{\text{class}})) \in \mathbb{R}^{T_{\text{in}} \times C \times D}, \tag{5}$$

$$\mathbf{Z}^{\text{nwp}} = \text{Conv}_{\text{nwp}}(\text{Stack}(\widetilde{\mathbf{F}}, \mathbf{R}_{\text{nwp}}^{\text{time}}, \mathbf{R}_{\text{nwp}}^{\text{class}})) \in \mathbb{R}^{T_{\text{out}} \times C \times D}, \tag{6}$$

where $\mathbf{R}_{\text{hist}}^{\text{time}} \in \mathbb{R}^{H_{\text{in}} \times C}$ and $\mathbf{R}_{\text{nwp}}^{\text{time}} \in \mathbb{R}^{H_{\text{out}} \times C}$ are time marker matrices; $\mathbf{R}_{\text{hist}}^{\text{class}} \in \mathbb{R}^{H_{\text{in}} \times C}$ and $\mathbf{R}_{\text{nwp}}^{\text{class}} \in \mathbb{R}^{H_{\text{out}} \times C}$ are class marker matrices. The three patched tensors are assembled into a unified temporal-variable token grid by concatenating historical and future meteorological tokens along time and then combining them with power tokens across variables: $\mathbf{Z} = \text{Concat}_v(\mathbf{Z}^{\text{pow}}, \text{Concat}_t(\mathbf{Z}^{\text{hist}}, \mathbf{Z}^{\text{nwp}})) \in \mathbb{R}^{T \times (C+1) \times D}$.

4.2 Site-Specific Context Integration

Site Feature Tokenization. The same meteorological trajectory can lead to varying power responses across wind farms due to differences in site characteristics, including installed capacity, terrain, and land cover. To capture this site-level heterogeneity, Site Feature Tokenization maps the three modalities of static information into three modality-specific tokens in a shared latent space.

Tabular attributes and AlphaEarth embeddings are encoded through modality-specific linear projections and layer normalization, yielding feature tokens $\boldsymbol{c}^{\text{tab}}, \boldsymbol{c}^{\text{geo}} \in \mathbb{R}^D$, respectively. For satellite imagery, we employ a pretrained GeoRSCLIP model (Zhang et al., 2024) for compact tokenization. Given $\mathcal{I}_{g_i}$, GeoRSCLIP assigns each of the K surrounding tiles to one of G predefined geospatial classes, yielding class IDs $\boldsymbol{\ell}_{g_i} \in \{1, \ldots, G\}^K$. The resulting tile-level class embeddings are flattened and similarly mapped to the image feature token $\boldsymbol{c}^{\text{img}} \in \mathbb{R}^D$.

The three feature tokens are stacked into $\mathbf{C} \in \mathbb{R}^{3 \times D}$. To prevent over-reliance on static site context, we apply structured instance-level dropout during training: $\widetilde{\mathbf{C}} = \text{Dropout}(\mathbf{C}; p_{\text{feat}})$, where p_{feat} is the drop probability with one decision shared across the three site feature tokens.

Adaptive Static Context Injection. Given the patch tokens $\mathbf{Z}$, we flatten the temporal and variable dimensions to obtain $\bar{\mathbf{Z}} \in \mathbb{R}^{T(C+1) \times D}$. $\bar{\mathbf{Z}}$ then attends to the site feature tokens through multi-head attention (Vaswani et al., 2017): $\Delta\bar{\mathbf{Z}} = \text{MHA}(\bar{\mathbf{Z}}, \widetilde{\mathbf{C}}, \widetilde{\mathbf{C}})$. An instance-adaptive gate is subsequently computed from the pooled cross-attention output and a shared learnable gate vector:

$$\gamma = \sigma(\text{MLP}([\text{Mean}(\text{LN}(\Delta\bar{\mathbf{Z}})); \boldsymbol{q}^{\text{gate}}])), \tag{7}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, $\gamma \in (0,1)$ controls the static context contribution, and $\boldsymbol{q}^{\mathrm{gate}} \in \mathbb{R}^D$ is a gate vector. The gated cross-attention output is then incorporated into the patch tokens through a dropout-regularized ReZero-style update (Bachlechner et al., 2021):

$$\bar{\mathbf{Z}}^{\mathrm{ctx}} = \bar{\mathbf{Z}} + \alpha_{\mathrm{rz}}(\mathrm{LN}(\bar{\mathbf{Z}} + \mathrm{Dropout}_{\mathrm{res}}(\gamma \Delta \bar{\mathbf{Z}})) - \bar{\mathbf{Z}}), \tag{8}$$

where α_{rz} is a learnable ReZero coefficient. Finally, $\bar{\mathbf{Z}}^{\mathrm{ctx}}$ is reshaped back to the temporal-variable grid $\mathbf{Z}^{\mathrm{ctx}} \in \mathbb{R}^{T \times (C+1) \times D}$, where static site context has been incorporated into the patch tokens.

Meteo-Power Encoding. Historical power and meteorological observations provide direct evidence of how the target site responds to past atmospheric conditions. Meteo-Power Encoding summarizes the historical tokens into a compact response token and broadcasts it across all patches.

For each historical patch, MPE concatenates all variable tokens within the patch and maps them into a local response token: $\boldsymbol{r}_t = \mathrm{MLP}_{\mathrm{mpe}}(\mathrm{vec}(\mathbf{Z}_t^{\mathrm{ctx}})) \in \mathbb{R}^D$, $t = 1, \ldots, T_{\mathrm{in}}$, where $\mathbf{Z}_t^{\mathrm{ctx}} \in \mathbb{R}^{(C+1) \times D}$ contains all variable tokens of historical patch t, and $\mathrm{MLP}_{\mathrm{mpe}} : \mathbb{R}^{(C+1)D} \to \mathbb{R}^D$ denotes the patch-wise response extractor. Using a learnable query $\boldsymbol{q}^{\mathrm{mpe}} \in \mathbb{R}^D$, multi-head attention pooling aggregates the local tokens into a compact representation of the target site's weather-to-power response patterns: $\boldsymbol{\rho} = \mathrm{MHA}(\boldsymbol{q}^{\mathrm{mpe}}, [\boldsymbol{r}_1; \ldots; \boldsymbol{r}_{T_{\mathrm{in}}}], [\boldsymbol{r}_1; \ldots; \boldsymbol{r}_{T_{\mathrm{in}}}]) \in \mathbb{R}^D$. Finally, $\boldsymbol{\rho}$ is appended to every patch as an auxiliary response token: $\mathbf{H}^0 = \mathrm{Concat}_v(\mathbf{Z}^{\mathrm{ctx}}, \mathrm{Broadcast}_T(\boldsymbol{\rho})) \in \mathbb{R}^{T \times V_{\mathrm{enc}} \times D}$, where $V_{\mathrm{enc}} = C + 2$ and $\mathbf{H}^0$ denotes the input to the first encoder block.

4.3 Weather-to-Power Representation Learning

W2P-FM employs a Variable-wise Multi-scale Transformer backbone comprising L encoder blocks. Each block combines Variable-wise Attention with a Multi-scale FFN.

Variable-wise Attention. Given input $\mathbf{H}^\ell$ at layer ℓ, variable-wise attention is applied independently within each patch: $\mathbf{A}_t^\ell = \mathrm{MHA}_{\mathrm{var}}^\ell(\mathbf{H}_t^\ell, \mathbf{H}_t^\ell, \mathbf{H}_t^\ell) \in \mathbb{R}^{V_{\mathrm{enc}} \times D}, t = 1, \ldots, T$. Stacking the patch-wise outputs along the temporal dimension yields $\mathbf{A}^\ell$. The resulting representation is obtained by a residual update: $\widetilde{\mathbf{H}}^\ell = \mathrm{LN}(\mathbf{H}^\ell + \mathrm{Dropout}(\mathbf{A}^\ell))$. This variable-wise attention models intra-patch weather-to-power interactions, allowing contemporaneous meteorological forcing to be interpreted in light of the site's recent generation behavior from the MPE tokens.

Multi-scale FFN. The Multi-scale FFN (MSFFN) aggregates temporal context across neighboring patches independently for each variable channel. Before applying multiple temporal convolution branches, a learnable variable-specific embedding is added to channel v: $\mathbf{G}_{:,v}^{\ell,b} = \mathrm{Conv}_b^\ell(\widetilde{\mathbf{H}}_{:,v}^\ell + \boldsymbol{e}_v), b = 1, \ldots, B$, where $\boldsymbol{e}_v \in \mathbb{R}^D$ is the learnable embedding of variable v, and B is the number of active temporal branches. These branches use different local receptive fields over the patch sequence, ranging from the current patch to progressively wider neighboring contexts, allowing the FFN to integrate short-term fluctuations and broader local temporal patterns across multiple scales. The branch outputs are concatenated and mapped back to the hidden dimension by a linear layer: $\mathbf{M}_{:,v}^\ell = \mathrm{GELU}(\mathrm{Linear}^\ell(\mathrm{Concat}_{b=1}^B(\mathbf{G}_{:,v}^{\ell,b}))) \in \mathbb{R}^{T \times D}$, where $\mathrm{GELU}(\cdot)$ denotes the Gaussian error linear unit. Applying this computation independently for each v yields $\mathbf{M}^\ell$. The block output is obtained as $\mathbf{H}^{\ell+1} = \mathrm{LN}(\widetilde{\mathbf{H}}^\ell + \mathrm{Dropout}(\mathbf{M}^\ell))$.

MPE provides sequence-wide historical response context, while MSFFN performs local multi-scale temporal aggregation. This complementary design captures the temporal dependencies required for wind power forecasting without global temporal self-attention, supporting efficient modeling of site-conditioned weather-to-power dynamics over long forecast horizons.

4.4 Forecast Heads

After the final encoder layer, the future power tokens are selected and mapped to forecasting targets. For deterministic forecasting, a granularity-specific head outputs the point forecasts within each future patch: $\hat{\boldsymbol{y}}_i = \mathrm{Flatten}(\mathrm{Head}_{\mathrm{det}}^{(n_i)}(\mathbf{H}_{T_{\mathrm{in}}+1:T,\mathrm{pow}}^L)) \in \mathbb{R}^{n_i H_{\mathrm{out}}}$, where $\mathbf{H}_{T_{\mathrm{in}}+1:T,\mathrm{pow}}^L \in \mathbb{R}^{T_{\mathrm{out}} \times D}$ denotes the final-layer power tokens of future patches, and $\mathrm{Head}_{\mathrm{det}}^{(n_i)} : \mathbb{R}^D \to \mathbb{R}^{n_i L_{\mathrm{p}}}$ is the deterministic forecast head for sampling density n_i. For probabilistic forecasting, we instead use a quantile head: $\widehat{\mathbf{Q}}_i = \mathrm{Flatten}(\mathrm{Head}_{\mathrm{quan}}^{(n_i)}(\mathbf{H}_{T_{\mathrm{in}}+1:T,\mathrm{pow}}^L)) \in \mathbb{R}^{n_i H_{\mathrm{out}} \times |\mathcal{Q}|}$, where $\mathcal{Q}$ denotes the set of quantile levels, and $\mathrm{Head}_{\mathrm{quan}}^{(n_i)} : \mathbb{R}^D \to \mathbb{R}^{n_i L_{\mathrm{p}} \times |\mathcal{Q}|}$ predicts patch-wise quantiles.

Table 1: Full-shot forecasting performance. P-GRU denotes PatchTST-GRU. Imp. (%) reports the reduction over the best baseline. Lower values are better. **Bold**: best; underlined: second best.

Continent	Horizon	Imp. (%)		W2P-FM		W2P-FM-NS		Crossformer		DUET		TimeBridge		TimeXer		XLinear		P-GRU		C-LSTM		iTransformer		LightGBM	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Asia	6	20.68%	30.15%	**.0928**	**.0190**	.1030	.0230	.1219	.0281	.1288	.0333	.1194	.0297	.1205	.0298	.1225	.0311	.1211	.0295	.1369	.0353	.1197	.0280	.1170	.0272
	24	21.80%	30.67%	**.1112**	**.0260**	.1235	.0318	.1444	.0384	.1735	.0537	.1622	.0482	.1661	.0503	.1701	.0521	.1469	.0402	.1602	.0441	.1546	.0439	.1422	.0375
	168	16.36%	24.37%	**.1467**	**.0422**	.1548	.0468	.1837	.0566	.2052	.0728	.2005	.0698	.2041	.0713	.2030	.0707	.1754	.0561	.1902	.0618	.1946	.0629	.1839	.0558
Europe	6	12.21%	13.17%	**.0870**	**.0178**	.0922	.0194	.0991	.0205	.1311	.0349	.1221	.0314	.1265	.0332	.1256	.0332	.1127	.0254	.1279	.0326	.1120	.0266	.1029	.0230
	24	14.98%	19.73%	**.0993**	**.0240**	.1076	.0270	.1205	.0300	.1919	.0659	.1723	.0565	.1840	.0626	.1805	.0605	.1241	.0311	.1641	.0549	.1576	.0490	.1168	.0299
	168	15.26%	15.65%	**.1516**	**.0496**	.1588	.0532	.1981	.0699	.2313	.0926	.2195	.0836	.2286	.0911	.2279	.0871	.1849	.0622	.2152	.0798	.2196	.0821	.1789	.0588
North America	6	5.44%	5.63%	**.1008**	**.0218**	.1032	.0224	.1076	.0233	.1564	.0449	.1349	.0354	.1444	.0394	.1491	.0423	.1119	.0246	.1229	.0287	.1133	.0261	.1066	.0231
	24	3.81%	1.77%	**.1137**	**.0277**	.1158	.0281	.1233	.0296	.2038	.0674	.1641	.0473	.1824	.0569	.1949	.0628	.1259	.0303	.1337	.0335	.1325	.0338	.1182	.0282
	168	5.78%	7.23%	**.1679**	**.0513**	.1691	.0521	.1840	.0579	.2353	.0866	.2081	.0697	.2275	.0836	.2269	.0809	.1827	.0572	.1917	.0615	.1988	.0652	.1782	.0553
South America	6	4.30%	6.45%	**.1024**	**.0203**	.1044	.0214	.1070	.0217	.1260	.0298	.1165	.0263	.1180	.0267	.1189	.0271	.1146	.0241	.1182	.0256	.1126	.0244	.1076	.0233
	24	5.03%	6.77%	**.1151**	**.0248**	.1153	.0251	.1212	.0266	.1494	.0391	.1353	.0328	.1368	.0337	.1391	.0347	.1287	.0297	.1374	.0332	.1321	.0315	.1252	.0298
	168	4.46%	5.07%	**.1479**	**.0393**	.1487	.0398	.1548	.0414	.1780	.0528	.1626	.0456	.1707	.0496	.1732	.0505	.1569	.0426	.1697	.0486	.1649	.0466	.1641	.0474
Oceania	6	3.14%	4.26%	**.0924**	**.0180**	.0945	.0187	.0954	.0188	.1356	.0344	.1232	.0296	.1243	.0297	.1346	.0350	.0991	.0199	.1071	.0223	.1038	.0221	.0984	.0204
	24	4.65%	5.24%	**.1045**	.0235	.1057	**.0232**	.1096	.0248	.1784	.0539	.1504	.0404	.1590	.0451	.1816	.0552	.1135	.0259	.1226	.0297	.1215	.0293	.1122	.0262
	168	5.97%	8.61%	**.1371**	**.0361**	.1384	**.0361**	.1458	.0395	.2153	.0731	.1878	.0580	.2042	.0685	.2189	.0735	.1487	.0401	.1594	.0449	.1690	.0487	.1504	.0418

5 EXPERIMENTS

W2P-FM is extensively evaluated in four aspects: (1) deterministic forecasting accuracy across multiple horizons in the full-shot setting; (2) cross-site and cross-region zero-shot forecasting on unseen sites and regional datasets; (3) transferability of the pretrained representations to probabilistic objectives and alternative temporal granularities through lightweight adaptation; and (4) the effectiveness of key design choices through ablation studies and empirical analyses. Dataset details, experimental settings and protocols, and complete results are provided in Appendices B, C, and E.

5.1 EXPERIMENTAL SETUP

Data and Training. A large-scale corpus of 1,012 site-level wind power series from 33 real-world datasets spanning five continents is constructed. Among these, 74 geographically and environmentally diverse sites are held out exclusively for cross-site zero-shot evaluation. Each site's samples are chronologically split into training, validation, and test sets at a $7:1:2$ ratio. W2P-FM is jointly pretrained on the training splits of the remaining 938 sites at an hourly resolution across three context–horizon configurations, $(H_{\text{in}}, H_{\text{out}}) \in \{(24, 6), (72, 24), (72, 168)\}$ hours. The resulting pretrained checkpoint is reused for adaptation to cross-granularity and probabilistic forecasting.

Baselines. We compare W2P-FM with representative methods from several paradigms: the industrial machine learning model LightGBM (Ke et al., 2017); recent advanced time-series forecasting models, including Crossformer (Zhang & Yan, 2023), DUET (Qiu et al., 2025), iTransformer (Liu et al., 2024c), TimeBridge (Liu et al., 2024a), XLinear (Chen et al., 2026), and TimeXer (Wang et al., 2024); recent domain-specific models, including PatchTST-GRU (Xu et al., 2025) and C-LSTM (Wang et al., 2025); time-series foundation models, including Chronos-2 (Ansari et al., 2025), TiRex-2 (Podest et al., 2026), CITRAS-FM (Yamaguchi et al., 2026), and Moirai 2.0 (Liu et al., 2025); the wind-power foundation model WindFM (Fan et al., 2025); and probabilistic forecasting models, including DeepAR (Salinas et al., 2020), TMDM (Li et al., 2024), and NsDiff (Ye et al., 2025). Because most baselines lack native support for static site information, we also include **W2P-FM-NS**, an otherwise identical variant without static context conditioning, to facilitate fair comparison. Results for all trainable methods are averaged over three random seeds.

5.2 MAIN RESULTS

Full-shot. Table 1 reports full-shot forecasting results on the test splits of sites included in pretraining. W2P-FM and W2P-FM-NS outperform baselines trained separately for each dataset or horizon in all 15 settings, although each uses a single set of parameters across sites and horizons. Relative to the strongest baseline in each setting, W2P-FM reduces MSE by 12.32% on average. The static context conditioning further improves performance in most settings, with a minor exception in Oceania. Several datasets in this region aggregate generation from geographically dispersed turbines, making static context extracted at a single coordinate an imperfect representation of the target's spatial support. This mismatch may explain the observed decline in forecasting accuracy.

Zero-shot. As shown in Table 2, W2P-FM consistently achieves the lowest zero-shot MAE and MSE in the cross-site setting. Its advantage over W2P-FM-NS across horizons indicates that static

Table 2: Zero-shot forecasting performance. In addition to the cross-site results, cross-region results are reported on 9 entirely held-out regional datasets, using the same W2P-FM architecture separately pretrained on the remaining regions. Imp. (%) reports the reduction over the best zero-shot baseline. Best and second-best zero-shot results are **bold** and underlined, respectively.

Protocol	Horizon	Zero-shot																Full-shot			
		Imp. (%)		W2P-FM		W2P-FM-NS		WindFM		Chronos-2		TiRex-2		CITRAS-FM		Moirai 2.0		LightGBM		Crossformer	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Cross-Site	6	28.75%	49.54%	**0.1031**	**0.0219**	0.1054	0.0228	0.1601	0.0474	0.1466	0.0441	0.1474	0.0458	0.1447	0.0434	0.1570	0.0517	0.1085	0.0233	0.1048	0.0217
	24	33.09%	52.88%	**0.1197**	**0.0294**	0.1213	0.0298	0.2050	0.0709	0.1851	0.0668	0.1789	0.0624	0.1928	0.0718	0.1952	0.0692	0.1235	0.0298	0.1228	0.0292
	168	30.05%	49.21%	**0.1539**	**0.0450**	0.1551	0.0454	0.2615	0.1187	0.2391	0.1090	0.2200	0.0886	0.2286	0.0949	0.2365	0.0922	0.1675	0.0490	0.1648	0.0485
Cross-Region	6	20.03%	40.35%	**0.1046**	**0.0238**	0.1061	0.0242	0.1570	0.0524	0.1338	0.0401	0.1308	0.0399	0.1395	0.0436	0.1452	0.0472	0.0910	0.0172	0.0952	0.0182
	24	31.67%	53.66%	**0.1249**	**0.0323**	0.1252	0.0329	0.2300	0.1018	0.1950	0.0782	0.1828	0.0697	0.2047	0.0832	0.1962	0.0745	0.1039	0.0220	0.1110	0.0237
	168	27.59%	47.89%	**0.1693**	**0.0531**	0.1711	0.0533	0.2795	0.1413	0.2508	0.1217	0.2338	0.1040	0.2468	0.1111	0.2451	0.1019	0.1633	0.0482	0.1647	0.0479

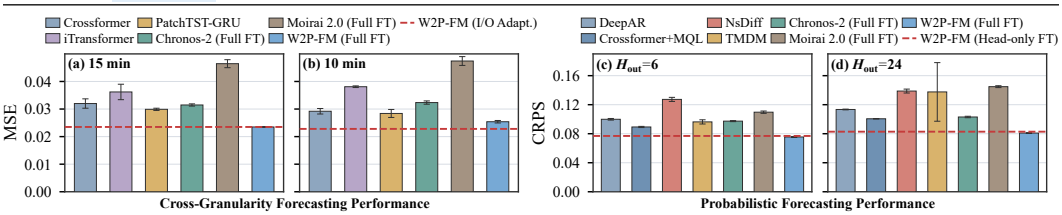

Figure 3: Cross-granularity and probabilistic forecasting performance. (a), (b) MSE at resolutions of 15 and 10 minutes; (c), (d) CRPS for horizons of 6 and 24 hours. Lower values are better.

context also benefits transfer to unseen sites. Generic time-series foundation models and WindFM remain less competitive. The former are less suited to site-conditioned weather-to-power modeling, whereas the latter relies on simulated pretraining and omits future NWP guidance. Moreover, W2P-FM performs comparably to the strongest full-shot baselines at shorter horizons and surpasses them at the longest horizon. To further assess W2P-FM's generalization to unseen regions, we conduct an additional cross-region zero-shot evaluation with the target regions entirely excluded from pretraining (Appendix C.7). The results, also reported in Table 2, demonstrate that W2P-FM retains a substantial lead in zero-shot forecasting, although a gap to full-shot references remains.

Cross-Granularity and Probabilistic Forecasting. W2P-FM's transferability to cross-granularity and probabilistic forecasting is evaluated through lightweight adaptation and full fine-tuning (Full FT) with all parameters updated. For forecasting at resolutions of 15 and 10 minutes, I/O Adapt. updates only the granularity-specific power projection in Patch Embedding and the forecast head. As shown in Figure 3(a) and (b), both protocols rank in the top two overall at finer temporal granularities, with I/O Adapt. generally matching or surpassing Full FT and achieving the best average MAE and MSE. For probabilistic forecasting, head-only FT also delivers better results than all baselines including Crossformer+MQL, while Full FT yields only modest additional gains (Figure 3(c) and (d)). Collectively, these results demonstrate effective transfer of W2P-FM's pretrained weather-to-power representations to downstream tasks through lightweight adaptation with a frozen backbone.

5.3 Model Analysis

To better understand W2P-FM's representation learning and generalization, we first assess the effect of joint training across horizons, then analyze multi-source NWP fusion, static context conditioning, and temporal interaction modeling. See Appendix C.10 for details.

Joint Multi-Horizon Training. In W2P-FM, temporally aligned tokens allow supervision from different horizons to jointly shape shared parameters, encouraging representations useful across horizons. To evaluate the benefit of joint training, the jointly trained model is compared with independently trained horizon-specific counterparts under otherwise identical settings. Table 3 indicates that joint training improves average accuracy across all horizons, suggesting positive cross-horizon transfer. A natural explanation is that all tasks share underlying weather-to-power relationships despite differences in their prediction targets.

Table 3: Comparison of joint multi-horizon pretraining with independent horizon-specific training. **Bold**: best.

Horizon	MSE			MAE		
	Single	Joint	Imp. (%)	Single	Joint	Imp. (%)
6	0.0194	**0.0189**	**2.58%**	0.0946	**0.0928**	**1.90%**
24	0.0257	**0.0250**	**2.72%**	0.1102	**0.1072**	**2.72%**
168	0.0456	**0.0441**	**3.29%**	0.1540	**0.1486**	**3.51%**

Availability-Aware Multi-Source NWP Fusion. To assess MeteoRouter under varying NWP availability, we evaluate all seven nonempty combinations of GFS, EC, and Tianji on the Shanxi and

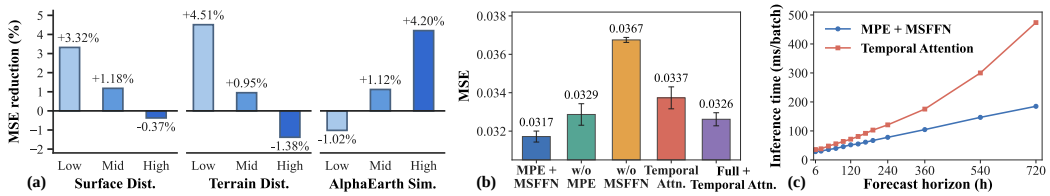

Figure 5: (a) Gains from static context conditioning across unseen site groups defined by surface-cover distance, terrain-attribute distance, and AlphaEarth similarity to pretraining sites. (b) MSE averaged across three horizons for five configurations (left to right): MPE + MSFFN, MSFFN alone, MPE + standard FFN, temporal attention + standard FFN, and MPE + MSFFN + temporal attention. (c) Inference time of MPE + MSFFN and temporal attention as the forecast horizon increases.

Anhui datasets using a fixed pretrained checkpoint. The three-source configuration achieves the lowest MSE in both regions (Table 4), suggesting that MeteoRouter effectively exploits complementary information across NWP sources. We further examine how these gains relate to NWP fusion by comparing the three-source fused NWP trajectory with each single-source trajectory at 50 sites per region. For each site and comparison, instances are divided into 8 equal-sized groups (G1–G8) according to the RMS difference between normalized trajectories. Figure 4 shows that larger adjustments generally correspond to greater MSE reductions across regions and horizons, further supporting the predictive value of the adjustments made by MeteoRouter.

Table 4: MSE of a fixed pretrained W2P-FM with different NWP source combinations. TJ denotes Tianji. **Bold**: best.

Dataset	GFS+EC +TJ	w/o GFS	w/o EC	w/o TJ	GFS only	EC only	TJ only
Shanxi	**0.0276**	0.0283	0.0288	0.0280	0.0326	0.0285	0.0310
Anhui	**0.0168**	0.0172	0.0172	0.0170	0.0188	0.0178	0.0177

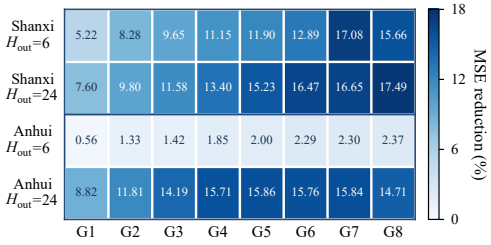

Figure 4: Grouped analysis of NWP adjustments induced by MeteoRouter. Instances are grouped by difference between three-source fused and single-source normalized NWP trajectories, ordered from smallest to largest.

Static Context Conditioning. For a more detailed analysis of how static site context affects generalization, sites from the cross-site zero-shot evaluation are grouped separately according to their surface-cover distance, terrain-attribute distance, and AlphaEarth similarity relative to the pretraining corpus. We then compare the MSE of W2P-FM and W2P-FM-NS within each group. As demonstrated in Figure 5(a), the MSE reduction generally increases with static-context proximity, while the most dissimilar groups show slight degradation. This observation suggests that static information helps the model learn site-specific weather-to-power representations when target contexts are adequately represented during pretraining, whereas larger static distribution shifts reduce its informativeness and may induce mild negative transfer.

Temporal Interaction Modeling. MPE and MSFFN provide complementary temporal interaction pathways. We evaluate their contributions through controlled variants that remove either component or introduce temporal attention as an alternative or additional pathway. Figure 5(b) shows that removing either component or introducing temporal attention increases average MSE. Moreover, the temporal-attention variant incurs higher inference latency, as presented in Figure 5(c). These results suggest that unrestricted long-range interactions may hinder the capture of local dependencies while incurring quadratic temporal complexity. In contrast, MPE and MSFFN support weather-to-power representation learning through more accurate and computationally efficient temporal modeling.

6 Conclusion

We presented W2P-FM, a domain-specific foundation model for wind power forecasting across sites, horizons, temporal granularities, and predictive formulations. W2P-FM unifies heterogeneous weather-power inputs and jointly exploits static context and recent observations, enabling its Variable-wise Multi-scale Transformer to learn site-conditioned weather-to-power representations. Experiments on 33 real-world datasets across five continents demonstrate consistent improvements over representative baselines in both full-shot and zero-shot settings, while strong frozen-backbone adaptation to downstream tasks further supports the transferability of the learned representations.

AI USE STATEMENT

An LLM was used to assist with language polishing and literature search. Model development, experiments, and data analyses were conducted by the authors. The authors carefully reviewed all AI-assisted content, verified the search results and bibliographic details, and take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This study uses non-personal wind power, meteorological, remote-sensing, and geospatial data from existing sources and involves no human participants or personally identifiable information. The use and redistribution of all datasets remain subject to their respective licenses. W2P-FM should undergo target-site validation before being used to support operational decisions.

REPRODUCIBILITY STATEMENT

The Appendix documents the datasets, preprocessing procedures, data partitions, model and baseline configurations, training objectives, and evaluation protocols, and provides complete experimental results and instructions for training and evaluation. An anonymized implementation and configuration files are available at `https://anonymous.4open.science/r/W2P-FM-E381` to facilitate reproducibility.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Ioannis K Bazionis, Panagiotis A Karafotis, and Pavlos S Georgilakis. A review of short-term wind power probabilistic forecasting and a taxonomy focused on input data. *IET Renewable Power Generation*, 16(1):77–91, 2022.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Raymond Byrne and Paul MacArtain. Multi-annual time series operational data for a vestas v52 wind turbine located in a coastal peri-urban environment in ireland. *Data in Brief*, 48:109227, 2023.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. Xlinear: A lightweight and accurate mlp-based model for long-term time series forecasting with exogenous inputs. *arXiv preprint arXiv:2601.09237*, 2026.

Alex Clerc and Elizabeth Lingkan. Hill of towie wind farm open dataset, March 2025. URL `https://doi.org/10.5281/zenodo.14870023`.

Ting Dai and Wei Qiao. Trading wind power in a competitive electricity market using stochastic programing and game theory. *IEEE Transactions on Sustainable Energy*, 4(3):805–815, 2013.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Xiaochong Dong, Yingyun Sun, Lei Dong, Jian Li, Yan Li, and Lei Di. Transferable wind power probabilistic forecasting based on multi-domain adversarial networks. *Energy*, 285:129496, 2023.

Caroline Draxl, Andrew Clifton, Bri-Mathias Hodge, and Jim McCaa. The wind integration national dataset (wind) toolkit. *Applied Energy*, 151:355–366, 2015.

ECMWF. Open data. 2026. doi: 10.21957/OPEN-DATA. URL https://www.ecmwf.int/en/forecasts/datasets/open-data.

Nina Effenberger and Nicole Ludwig. A collection and categorization of open-source wind and wind power datasets. *Wind Energy*, 25(10):1659–1683, 2022.

Hang Fan, Yu Shi, Zongliang Fu, Shuo Chen, Wei Wei, Wei Xu, and Jian Li. Windfm: An open-source foundation model for zero-shot wind power forecasting. *arXiv preprint arXiv:2509.06311*, 2025.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Jose M Gonzalez, James E Tomlinson, Eduardo A Martínez Ceseña, Mohammed Basheer, Emmanuel Obuobie, Philip T Padi, Salifu Addo, Rasheed Baisie, Mikiyas Etichia, Anthony Hurford, et al. Designing diversified renewable energy systems to balance multisector performance. *Nature Sustainability*, 6(4):415–427, 2023.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.

Christian M Grams, Remo Beerli, Stefan Pfenninger, Iain Staffell, and Heini Wernli. Balancing europe's wind-power output through spatial deployment informed by weather regimes. *Nature climate change*, 7(8):557–562, 2017.

Kurt Schaldemose Hansen, Steen Arne Sørensen, and Nikola Vasiljevic. Turbulence and turbine measurements from the Norrekaer Enge windfarm. 4 2022. doi: 10.11583/DTU.19336667.v1. URL https://data.dtu.dk/articles/dataset/Turbulence_and_turbine_measurements_from_the_Norrekaer_Enge_windfarm_/19336667.

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.

Jiahui Huang, Ao Luo, Lei Liu, Hongwei Zhao, Tengyuan Liu, Ruibo Guo, Bo Wang, Zhao Wang, and Bin Li. Tyan-wp: A wind power foundation model for ultra-short-term probabilistic forecasting. *arXiv preprint arXiv:2606.08630*, 2026.

Avi Jakhmola, Jessica Jewell, Vadim Vinichenko, and Aleh Cherp. Probabilistic projections of global wind and solar power growth based on historical national experience. *Nature Energy*, 11: 743–755, 2026.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Yuhao Li, Han Wang, Jie Yan, Chang Ge, Shuang Han, and Yongqian Liu. Ultra-short-term wind power forecasting based on the strategy of "dynamic matching and online modeling". *IEEE Transactions on Sustainable Energy*, 16(1):107–123, 2025.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 18604–18622, 2024.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *arXiv preprint arXiv:2410.04442*, 2024a.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024c.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024d.

Manuel A Matos and Ricardo J Bessa. Setting the operating reserve using probabilistic wind power forecasts. *IEEE transactions on power systems*, 26(2):594–603, 2011.

Qinglin Meng, Yuan Gao, Ying He, Sheharyar Hussain, Jinghang Lu, Yahui Chen, Chao Qin, Fengzhang Luo, and Josep M Guerrero. Powermistral: A data-efficient wind power forecasting framework leveraging pre-trained large language models. *Applied Energy*, 411:127625, 2026.

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, 2015. URL `https://doi.org/10.5065/D65D8PWK`.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026.

Daniel Vazquez Pombo, Oliver Gehrke, and Henrik W Bindner. Solete, a 15-month long holistic dataset including: Meteorology, co-located wind and solar pv power from denmark with various resolutions. *Data in Brief*, 42:108046, 2022.

Abhinav Prakash, Se Yoon Lee, Xin Liu, Lei Liu, Bani K Mallick, and Yu Ding. A bayesian hierarchical model to understand the effect of terrain on wind turbine power curves. *IEEE Transactions on Sustainable Energy*, 15(2):1127–1137, 2024.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1185–1196, 2025.

Yves-Marie Saint-Drenan, Romain Besseau, Malte Jansen, Iain Staffell, Alberto Troccoli, Laurent Dubus, Johannes Schmidt, Katharina Gruber, Sofia G Simões, and Siegfried Heier. A parametric model for wind turbine power curves incorporating environmental conditions. *Renewable Energy*, 157:754–768, 2020.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Sumaira Tasnim, Ashfaqur Rahman, Amanullah Maung Than Oo, and Md Enamul Haque. Wind power prediction in new stations based on knowledge of existing stations: A cluster based multi source domain adaptation approach. *Knowledge-Based Systems*, 145:15–24, 2018.

Guillermo Terrén-Serrano, Ranjit Deshmukh, and Manel Martínez-Ramón. Probabilistic day-ahead forecasting of system-level renewable energy and electricity demand. *Nature Communications*, 17(1):3307, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ding Wang, Min Xu, Zhu Guangming, Futao Luo, Jiaxin Gao, and Yuntian Chen. Enhancing wind power forecasting accuracy through lstm with adaptive wind speed calibration (c-lstm). *Scientific reports*, 15(1):5352, 2025.

Hexian Wang, Tongming Zhou, Chengzhen Jia, Yushan Liu, and Lingmei Wang. Deepwind: A foundation model for zero-shot wind power forecasting. *Energy*, 360:141793, 2026.

Yun Wang, Runmin Zou, Fang Liu, Lingjun Zhang, and Qianyi Liu. A review of wind speed and wind power forecasting with deep neural networks. *Applied energy*, 304:117766, 2021.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Shiwei Xu, Yongjun Wang, Xinglei Xu, Guang Shi, He Huang, and Yingya Zheng. A patchtst-gru based heterogeneous seq2seq model with numerical weather prediction refinement for multi-step wind power forecasting. *Scientific Reports*, 15(1):16547, 2025.

Yosuke Yamaguchi, Issei Suemitsu, Yuki Kajihara, and Wenpeng Wei. Citras-fm: Tiny time series foundation model for covariate-informed zero-shot forecasting. *arXiv preprint arXiv:2606.10798*, 2026.

Jie Yan, Corinna Möhrlen, Tuhfe Göçmen, Mark Kelly, Arne Wessel, and Gregor Giebel. Uncovering wind power forecasting uncertainty sources and their propagation through the whole modelling chain. *Renewable and Sustainable Energy Reviews*, 165:112519, 2022.

Yang Yang, Hao Lou, Jinran Wu, Shaotong Zhang, and Shangce Gao. A survey on wind power forecasting with machine learning approaches. *Neural Computing and Applications*, 36(21):12753–12773, 2024.

Jiexia Ye, Yongzi Yu, Weiqi Zhang, Le Wang, Jia Li, and Fugee Tsung. Empowering time series analysis with foundation models: A comprehensive survey. *Information Fusion*, 137:104601, 2027. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104601.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. *arXiv preprint arXiv:2505.04278*, 2025.

Wenyi Zhang, Wuyou Xiao, Xudong Li, and Zhao Xu. A hybrid convolutional-llm paradigm for robust wind power forecasting in data-scarce regimes. *Applied Energy*, 415:127914, 2026.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024.

A COMPUTATIONAL COMPLEXITY

We analyze the forward computation of the forecasting architecture per instance, retaining the coefficients of linear projections, convolutions, dot products, and weighted sums. One multiply–accumulate is counted as two FLOPs. Bias additions, normalization, nonlinearities (including softmax), masking, and data movement are excluded from this arithmetic count. Offline extraction of the cached satellite descriptors and AlphaEarth embeddings is also excluded. We use the previously defined S, C, D, L, L_{p}, T_{in}, T_{out}, $T = T_{\mathrm{in}} + T_{\mathrm{out}}$, and $V_{\mathrm{enc}} = C + 2$. For padded inputs, the sequence lengths in this analysis refer to the processed lengths, not the number of valid targets.

MeteoRouter. The shared $H_{\mathrm{out}} \to D$ projection is applied to all SC source-variable trajectories, costing $2SCH_{\mathrm{out}}D$ FLOPs. The SC query-trajectory dot products cost $2SCD$, and the source-weighted aggregation over $H_{\mathrm{out}}C$ meteorological values costs $2SCH_{\mathrm{out}}$. Thus,

$$F_{\mathrm{router}} = 2SCH_{\mathrm{out}}D + 2SCD + 2SCH_{\mathrm{out}} \approx 6CH_{\mathrm{out}}D, \tag{9}$$

where $S = 3$ denotes the three NWP source slots and all three are projected even when some sources are masked as unavailable.

Granularity-Aware Patch Embedding. Each observation is augmented with two markers, giving three input channels. The power projection uses a kernel of $n_i L_{\mathrm{p}}$ observations and produces T patches. The historical and future meteorological projections each use a kernel of L_{p} hourly observations and process C variables. Their respective costs are

$$\begin{aligned} F_{\mathrm{patch}} &= 6Tn_iL_{\mathrm{p}}D + 6T_{\mathrm{in}}CL_{\mathrm{p}}D + 6T_{\mathrm{out}}CL_{\mathrm{p}}D \\ &= 6(H_{\mathrm{in}} + H_{\mathrm{out}})D(1 + C). \end{aligned} \tag{10}$$

Here, we consider the hourly pretraining setting ($n_i = 1$). For the adopted context–horizon configurations, $TL_{\mathrm{p}} = H_{\mathrm{in}} + H_{\mathrm{out}}$.

Static Context Conditioning. Let d_{img} be the tile-class embedding width and d_{g} the hidden width of the instance-gating MLP. Projecting the tabular attributes, the 64-dimensional AlphaEarth embedding, and the flattened embeddings of K tiles costs

$$F_{\mathrm{feat}} = 2D(d_{\mathrm{tab}} + 64 + Kd_{\mathrm{img}}). \tag{11}$$

Cross-attention takes the $T(C+1)$ flattened patch tokens as queries and the three site feature tokens as keys and values. The query and output projections together cost $4T(C + 1)D^2$, while the key and value projections cost $12D^2$. The attention-score and weighted-value products together cost $12T(C + 1)D$. The instance gate has dimensions $2D \to d_{\mathrm{g}} \to 1$, adding $4Dd_{\mathrm{g}} + 2d_{\mathrm{g}}$. Static context conditioning is applied once before MPE, yielding

$$F_{\mathrm{cond}} = 4[T(C + 1) + 3]D^2 + 12T(C + 1)D + 4Dd_{\mathrm{g}} + 2d_{\mathrm{g}}, \tag{12}$$

$$F_{\mathrm{feat}} + F_{\mathrm{cond}} \approx 4[T(C + 1) + 3]D^2. \tag{13}$$

Meteo-Power Encoding. The patch-wise response extractor has dimensions $(C+1)D \to 2D \to D$. Applying its two linear layers to T_{in} historical patches costs $4T_{\mathrm{in}}(C + 1)D^2 + 4T_{\mathrm{in}}D^2$. Attention pooling uses one query and T_{in} keys/values, costing $4(T_{\mathrm{in}} + 1)D^2$ for projections and $4T_{\mathrm{in}}D$ for the two attention products. Therefore,

$$\begin{aligned} F_{\mathrm{MPE}} &= 4T_{\mathrm{in}}(C + 1)D^2 + 4T_{\mathrm{in}}D^2 + 4(T_{\mathrm{in}} + 1)D^2 + 4T_{\mathrm{in}}D \\ &\approx 4[T_{\mathrm{in}}(C + 3) + 1]D^2. \end{aligned} \tag{14}$$

The response token is computed once and broadcast to all patches. It is not recomputed at every patch or encoder block.

Variable-wise Multi-scale Transformer. Variable-wise attention operates independently on V_{enc} tokens within each of the T patches. In the implementation, the query projection also produces an output gate, with dimensions $D \to 2D$; the key, value, and output projections each have dimensions $D \to D$. These projections cost $(4 + 2 + 2 + 2)TV_{\mathrm{enc}}D^2$. The attention-score and weighted-value products together cost $4TV_{\mathrm{enc}}^2D$, giving the per-block attention count

$$F_{\mathrm{attn}} = 10TV_{\mathrm{enc}}D^2 + 4TV_{\mathrm{enc}}^2D. \tag{15}$$

For the Multi-scale FFN, let k_b denote the kernel size of active branch b. Each branch is a dense $D \to D$ temporal convolution applied to all V_{enc} variable sequences, costing $2TV_{\text{enc}}k_bD^2$. Concatenating the B branch outputs and projecting $BD \to D$ adds $2BTV_{\text{enc}}D^2$. Hence,

$$F_{\text{FFN}} = 2TV_{\text{enc}}D^2\left(\sum_{b=1}^{B} k_b + B\right). \tag{16}$$

The configuration uses four branches with kernel sizes $(1, 3, 3, 3)$ and dilations $(1, 1, 2, 3)$. Dilation enlarges the receptive field without increasing the number of kernel coefficients. Thus, $\sum_b k_b = 10$, $F_{\text{FFN}} = 28TV_{\text{enc}}D^2$, and the L-block backbone costs

$$F_{\text{enc}} = L\left(38TV_{\text{enc}}D^2 + 4TV_{\text{enc}}^2D\right). \tag{17}$$

Forecast Heads and Total Cost. For hourly deterministic forecasting ($n_i = 1$), the head maps each of the T_{out} future power tokens from D to L_{p} outputs, costing

$$F_{\text{head}} = 2T_{\text{out}}DL_{\text{p}} = 2H_{\text{out}}D. \tag{18}$$

Under the adopted configuration, the encoder remains the dominant source of computation. Retaining the leading terms from the backbone, static-context conditioning, and MPE, the total cost is approximated as

$$\begin{aligned} F_{\text{total}} \approx\ & 38LT(C+2)D^2 + 4[T(C+1)+3]D^2 \\ & + 4[T_{\text{in}}(C+3)+1]D^2. \end{aligned} \tag{19}$$

The total computational cost therefore scales linearly with the total temporal span, $H_{\text{in}} + H_{\text{out}}$. In particular, W2P-FM's backbone does not apply temporal attention independently to each variable, thereby avoiding the corresponding quadratic cost of $4LV_{\text{enc}}T^2D$.

B Dataset Details

All training and evaluation experiments are conducted on a corpus of 1,012 wind power series drawn from 33 source datasets across five continents. The corpus includes publicly available resources such as the Turkey Wind Turbine SCADA Dataset, Kelmarsh Wind Farm Data, and La Haute Borne Open Wind Farm Dataset (Effenberger & Ludwig, 2022; Clerc & Lingkan, 2025; Pombo et al., 2022; Byrne & MacArtain, 2023; Hansen et al., 2022). It comprises both turbine-level series and aggregates at the wind-farm or regional level; for aggregated sources, a site denotes the corresponding aggregation unit. Capacity refers to each site's nominal capacity and is reported in MW. Table 5 provides the complete source-level inventory.

C Detailed Experimental Settings and Additional Analyses

C.1 Corpus Construction and Data Partitions

Of the 1,012 sites, 938 are used for deterministic multi-scale pretraining, while the remaining 74 are reserved exclusively for cross-site zero-shot evaluation. The geographic distribution of these sites is summarized in Table 6. Samples from each site are chronologically split into training, validation, and test sets at a $7 : 1 : 2$ ratio. Full-shot performance is evaluated on future test intervals from sites included in pretraining, whereas zero-shot performance is evaluated on those from sites entirely unseen during pretraining. Power is represented and forecast in capacity-factor space, and all normalization statistics are estimated exclusively from the training partitions of the pretraining sites to prevent information leakage. Missing and identified anomalous power observations are first replaced with zeros. Runs of at most two consecutive zeros are then filled by interpolation using neighboring valid observations (Xu et al., 2025). During window construction, a window is retained only if the remaining zeros account for no more than one third of the readings in each of its historical and target intervals. The same preprocessing and window selection procedures are applied to all models. After preprocessing, the corpus contains approximately 29 million nonzero hourly power data points. The pretraining corpus is sampled hourly, with each instance comprising historical power, historical meteorological observations, available future NWP trajectories, and static site context.

Table 5: Geographic and capacity statistics of the 33 source datasets.

ID	Continent	Region/Country	#Sites	Mean Capacity (MW)
D01	Asia	Yalova, Turkey	1	3.600
D02	Asia	Gansu, China	1	201.000
D03	Asia	Heilongjiang, China	2	49.625
D04	Asia	Jiangsu, China	3	367.167
D05	Asia	Shandong, China	16	192.841
D06	Asia	Shanxi, China	268	72.459
D07	Asia	Shanxi, China	11	110.594
D08	Asia	Anhui, China	110	73.983
D09	Asia	Hubei, China	1	80.000
D10	Asia	Tianjin, China	1	20.000
D11	Asia	Guizhou, China	1	39.600
D12	Europe	Cornwall, UK	10	0.400
D13	Europe	Scotland, UK	21	2.300
D14	Europe	Northamptonshire, UK	6	2.050
D15	Europe	North Jutland, Denmark	40	0.500
D16	Europe	Meuse, France	4	2.050
D17	Europe	Scottish Borders, UK	14	2.050
D18	Europe	DTU Risø/SYSLAB, Denmark	1	0.011
D19	Europe	Tjæreborg Enge, Denmark	1	2.000
D20	Europe	Dundalk, Ireland	1	0.850
D21	Europe	Great Britain, UK	8	46.838
D22	Europe	Belgian North Sea, Belgium	7	215.457
D23	North America	Ontario, Canada	5	114.400
D24	North America	Alberta, Canada	5	146.800
D25	South America	Piauí, Brazil	20	0.900
D26	South America	Ceará, Brazil	32	0.800
D27	South America	Brazil	271	151.770
D28	South America	Uruguay	9	70.589
D29	South America	Peru	4	126.075
D30	Oceania	South Australia and Victoria	5	76.140
D31	Oceania	Eastern and Southern Australia	100	141.754
D32	Oceania	Western Australia	16	64.063
D33	Oceania	Southeastern Australia	17	87.668

Table 6: Geographic distribution of the pretraining and cross-site zero-shot corpus.

Continent	#Datasets	Pretrain	Zero-shot	Total
Asia	11	388	27	415
Europe	11	104	9	113
North America	2	10	0	10
South America	5	312	24	336
Oceania	4	124	14	138
Total	33	938	74	1,012

C.2 Meteorological Inputs and NWP Availability

We consider six meteorological variables: 100-m zonal and meridional winds, 2-m air and dew-point temperatures, surface pressure, and downward surface solar radiation. Historical meteorological inputs are obtained from ERA5, the fifth-generation atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020). Future forecasts are obtained from three sources. GFS refers to the Global Forecast System operated by the U.S. National Centers for Environmental Prediction (NCEP) (National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2015). EC denotes forecasts produced by ECMWF's Integrated Forecasting System (IFS) (ECMWF, 2026), while Tianji denotes forecasts supplied by Zhongke Tianji's meteorological data platform (`https://www.tjweather.com/`). A three-entry source-validity vector indicates NWP availability for each instance. MeteoRouter masks unavailable sources before the variable-wise softmax,

excluding them from both normalization and trajectory fusion. Since the EC product used in this study provides forecasts only up to 24 hours, it is excluded from the 168-hour forecasting task.

C.3 Static Site Context

The tabular modality comprises site capacity from source metadata, together with elevation and a terrain position index (TPI) computed over a 1-km neighborhood from the 30-m Copernicus DEM GLO-30. Capacity is log-transformed, while the remaining continuous attributes are robustly normalized using training-site medians and interquartile ranges. The satellite modality consists of a 3×3 neighborhood of tiles retrieved from Esri's ArcGIS World Imagery Wayback archive, with each tile mapped to a discrete geographic class descriptor by a frozen pretrained GeoRSCLIP model (Zhang et al., 2024). The AlphaEarth modality provides a 64-dimensional annual geospatial embedding extracted from Google Satellite Embedding V1 through Earth Engine and spatially averaged over a $100\,\mathrm{m} \times 100\,\mathrm{m}$ window centered on each site (Gorelick et al., 2017; Brown et al., 2025).

C.4 Training Objective

W2P-FM adopts an instance-balanced objective that first averages the pointwise losses over the forecast horizon of each instance and then averages the resulting instance-level losses over a batch. Let $\mathcal{B}$ denote a batch and $H_{\mathrm{out},i}$ the forecast horizon associated with instance i. For deterministic forecasting, the pointwise loss is the squared error, $\ell_{i,\tau} = (\hat{y}_{i,\tau} - y_{i,\tau})^2$. For probabilistic forecasting, $\ell_{i,\tau}$ is the multi-quantile loss (MQL), obtained by averaging the pinball loss over all quantile levels in $\mathcal{Q}$, where $\rho_q(u) = \max(qu, (q-1)u)$ and $u = y_{i,\tau} - \widehat{Q}^{(q)}_{i,\tau}$. The unified training objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{n_i H_{\mathrm{out},i}} \sum_{\tau=1}^{n_i H_{\mathrm{out},i}} \ell_{i,\tau}. \tag{20}$$

This two-level averaging assigns equal weight to each forecasting instance regardless of its forecast length.

C.5 Implementation Details.

W2P-FM is jointly pretrained at an hourly resolution across three context–horizon configurations, $(H_{\mathrm{in}}, H_{\mathrm{out}}) \in \{(24, 6), (72, 24), (72, 168)\}$, using a single shared backbone and parameterization. Across the three jointly pretrained scales, all samples are padded to common maxima of 72 historical hours and 168 forecast hours, with binary masks excluding padded target positions from both optimization and metric computation. W2P-FM is implemented in PyTorch and trained with distributed data parallelism on four NVIDIA A100 GPUs. The Variable-wise Multi-scale Transformer comprises three blocks with $D = 128$, 16 attention heads, and a dropout rate of 0.15, while Granularity-Aware Patch Embedding uses a 6-hour temporal span. Adam optimization is employed with an initial learning rate of 10^{-4}, cosine annealing with $T_{\max} = 15$ and $\eta_{\min} = 10^{-7}$, a batch size of 1,024, and early stopping with a patience of three epochs.

C.6 Baseline Configuration

We evaluate W2P-FM against representative baselines spanning industrial machine learning, recent time-series forecasting, wind-power-specific forecasting, time-series foundation models, and probabilistic forecasting. Their principal modeling characteristics are summarized as follows:

- **LightGBM** (Ke et al., 2017) is an efficient histogram-based gradient-boosted decision-tree framework that improves scalability through gradient-based one-side sampling and exclusive feature bundling.
- **Crossformer** (Zhang & Yan, 2023) combines dimension-segment-wise embedding with two-stage attention to model cross-temporal and cross-variable dependencies.
- **DUET** (Qiu et al., 2025) introduces temporal clustering and channel soft clustering to capture heterogeneous temporal patterns and sparse inter-variable dependencies.

- **iTransformer** (Liu et al., 2024c) represents complete variable trajectories as tokens and performs attention across variables to capture multivariate dependencies.
- **TimeBridge** (Liu et al., 2024a) employs Integrated and Cointegrated Attention to model short-term intra-variable dependencies and long-term cross-variable cointegration under non-stationarity.
- **XLinear** (Chen et al., 2026) employs MLP-based time-wise and variate-wise gating, with endogenous global tokens mediating interactions among exogenous variables.
- **TimeXer** (Wang et al., 2024) combines patch-wise self-attention with variate-wise cross-attention, using global endogenous tokens to incorporate exogenous information into endogenous temporal representations.
- **PatchTST-GRU** (Xu et al., 2025) combines a PatchTST encoder, a GRU decoder, NWP refinement, and fusion attention to integrate historical power patterns with future meteorological features for multi-step wind power forecasting.
- **C-LSTM** (Wang et al., 2025) augments LSTM with adaptive wind-speed calibration that dynamically fuses historical and forecast wind speeds through learnable weights.
- **Chronos-2** (Ansari et al., 2025) is a pretrained time-series foundation model that supports zero-shot univariate, multivariate, and covariate-informed forecasting through in-context learning.
- **TiRex-2** (Podest et al., 2026) is a recurrent xLSTM-based time-series foundation model for multivariate zero-shot forecasting with past and future-known covariates, combining time and variate mixing to preserve target causality and support efficient streaming inference.
- **CITRAS-FM** (Yamaguchi et al., 2026) is a compact 7M-parameter, patch-based decoder-only Transformer for univariate, multivariate, and covariate-informed zero-shot forecasting, introducing Shifted Attention to incorporate future-known covariates and CovSynth to synthesize observed and known pseudo-covariates for covariate-aware pretraining.
- **Moirai 2.0** (Liu et al., 2025) adopts a compact decoder-only architecture with single-patch encoding, multi-token prediction, and recursive multi-quantile decoding for efficient probabilistic forecasting.
- **WindFM** (Fan et al., 2025) discretizes multivariate wind observations into hierarchical tokens and autoregressively pretrains a decoder-only Transformer for zero-shot wind power forecasting.
- **DeepAR** (Salinas et al., 2020) trains a global autoregressive recurrent network across related time series to parameterize conditional predictive distributions.
- **TMDM** (Li et al., 2024) uses conditional representations derived from a Transformer as prior information in both the forward diffusion and reverse denoising processes for probabilistic multivariate time-series forecasting.
- **NsDiff** (Ye et al., 2025) introduces a location-scale noise model with input-dependent conditional variance and incorporates it into the diffusion process through an uncertainty-aware noise schedule to capture time-varying predictive uncertainty.

All trainable baselines follow the same chronological partitions, capacity-factor transformation, normalization statistics, target masks, and context–horizon configurations as W2P-FM. Their hyperparameters follow the corresponding official implementations and are selected using validation data. Future NWP is provided through native interfaces for exogenous inputs when supported and otherwise concatenated with historical inputs along the temporal dimension. LightGBM follows the conventional site-wise protocol, with a separate model trained and evaluated for each site, whereas other trainable baselines are trained at the dataset level. As most baselines lack native multi-source NWP support, the NWP source yielding the best validation performance is used for each baseline's test evaluation.

C.7 Cross-Region Zero-Shot Forecasting

W2P-FM is further evaluated in a more stringent cross-region zero-shot setting, where the training and test regions are fully disjoint. The 57 sites from 9 held-out regional datasets (D02–D05, D14,

D20, D22, D23, and D32 in Table 5) are designated as cross-region test sites, with the test interval of each site used for evaluation. Model pretraining uses only the training splits of sites from the remaining 24 regional datasets. Because no site from a target region is observed during training, this protocol provides a more stringent test of generalization to unseen regions than site-level holdout. As shown in Table 2, W2P-FM does not fully close the gap to supervised full-shot references, but consistently outperforms all representative zero-shot foundation models across all horizons and both metrics. Relative to the strongest zero-shot baseline at each horizon, it reduces MAE by 20.0%–31.7% and MSE by 40.4%–53.7%, demonstrating robust zero-shot generalization to unseen regions.

C.8 Frozen-Backbone Task Adaptation

For cross-granularity forecasting, the 15-minute evaluation uses datasets D03, D04, and D05, whereas the 10-minute evaluation uses D18, D20, and D26 in Table 5. For each granularity, the reported performance is averaged over the context–horizon configurations (24, 6) and (72, 24). For baselines that cannot natively accommodate mismatched power and meteorological resolutions, the hourly meteorological covariates are upsampled through temporal interpolation to match the corresponding power resolution before model input, without modifying the original baseline architecture.

For probabilistic forecasting, the pretrained backbone is frozen and only a new quantile head is optimized. At each forecast step, the head predicts 19 conditional quantiles at levels

$$\mathcal{Q} = \{0.05, 0.10, \ldots, 0.95\}. \tag{21}$$

Probabilistic adaptation is evaluated using site samples from seven geographically distributed datasets (D02, D08, D13, D14, D23, D26, and D33 in Table 5), with results reported separately for the context–horizon configurations (24, 6) and (72, 24).

C.9 Metrics

For both cross-site and cross-region zero-shot evaluation, MAE and MSE are computed separately for each test site at each forecast horizon and then averaged with equal weight across sites. For other evaluations, unless otherwise specified, metrics are computed separately for each dataset and then averaged with equal weight across datasets. Using the notation defined above, let $\mathcal{E}$ denote the set of evaluated instances. MAE is aggregated over all forecast targets,

$$\mathrm{MAE} = \frac{\sum_{i\in\mathcal{E}} \sum_{\tau=1}^{n_i H_{\mathrm{out},i}} |\hat{y}_{i,\tau} - y_{i,\tau}|}{\sum_{i\in\mathcal{E}} n_i H_{\mathrm{out},i}}, \tag{22}$$

whereas MSE follows the instance-balanced aggregation used for optimization,

$$\mathrm{MSE} = \frac{1}{|\mathcal{E}|} \sum_{i\in\mathcal{E}} \frac{1}{n_i H_{\mathrm{out},i}} \sum_{\tau=1}^{n_i H_{\mathrm{out},i}} (\hat{y}_{i,\tau} - y_{i,\tau})^2. \tag{23}$$

Both metrics are reported in capacity-factor space.

For the predictive distribution $\widehat{P}_{i,\tau}$ defined above, let $F_{\widehat{P}_{i,\tau}}$ denote its cumulative distribution function. The continuous ranked probability score is defined as

$$\mathrm{CRPS}\left(\widehat{P}_{i,\tau}, y_{i,\tau}\right) = \int_{-\infty}^{\infty} \left[F_{\widehat{P}_{i,\tau}}(z) - \mathbb{I}\{y_{i,\tau} \le z\}\right]^2 \mathrm{d}z. \tag{24}$$

Since each predictive distribution is represented by the 19 quantiles in $\mathcal{Q}$, CRPS is computed using the standard quantile approximation, namely twice the mean pinball loss across $\mathcal{Q}$, and then averaged over all forecast targets (Gneiting & Raftery, 2007).

We additionally report the interval score (IS). For a central $(1-\alpha)$ prediction interval $[L, U]$ and observation y, IS is defined as

$$\mathrm{IS}_\alpha(L, U; y) = \begin{cases} (U-L) + \dfrac{2}{\alpha}(L-y), & y < L, \\ U - L, & L \le y \le U, \\ (U-L) + \dfrac{2}{\alpha}(y-U), & y > U. \end{cases} \tag{25}$$

We use the central 80% interval with $\alpha = 0.2$, $L = \widehat{Q}_{i,\tau}^{(0.10)}$, and $U = \widehat{Q}_{i,\tau}^{(0.90)}$, and average IS over all forecast targets. Lower MAE, MSE, CRPS, and IS indicate better forecasting performance.

Table 7: Complete comparison between joint multi-horizon pretraining and independently trained single-horizon counterparts. Lower values are better. **Bold**: best.

Horizon	Continent	MSE		MAE	
		Single	Joint	Single	Joint
6	Asia	0.0196	**0.0190**	0.0954	**0.0928**
	Europe	0.0188	**0.0178**	0.0896	**0.0870**
	North America	**0.0215**	0.0218	**0.1006**	0.1008
	South America	**0.0203**	**0.0203**	**0.1024**	**0.1024**
	Oceania	0.0182	**0.0180**	0.0932	**0.0924**
24	Asia	0.0271	**0.0260**	0.1158	**0.1112**
	Europe	0.0253	**0.0240**	0.1028	**0.0993**
	North America	**0.0277**	**0.0277**	0.1162	**0.1137**
	South America	**0.0247**	0.0248	0.1154	**0.1151**
	Oceania	**0.0232**	0.0235	0.1054	**0.1045**
168	Asia	0.0446	**0.0422**	0.1548	**0.1467**
	Europe	0.0515	**0.0496**	0.1570	**0.1516**
	North America	0.0525	**0.0513**	0.1732	**0.1679**
	South America	**0.0392**	0.0393	0.1492	**0.1479**
	Oceania	0.0365	**0.0361**	0.1403	**0.1371**

C.10 Model Analysis

Joint Multi-Horizon Pretraining. To isolate the effect of joint multi-horizon supervision, we compare W2P-FM jointly pretrained on the context–horizon configurations (24, 6), (72, 24), and (72, 168) with three counterparts trained independently from scratch, each restricted to one configuration. For every horizon, the joint and independent settings use identical sites, chronological splits, horizon-specific training windows, input modalities, preprocessing, model architecture, loss function, and optimization settings. During joint pretraining, the three horizon-specific datasets are concatenated and randomly shuffled into a shared training stream without explicit horizon reweighting, and each sample contributes to the loss only at target positions associated with its own horizon. Thus, the training data and per-epoch exposure of each target horizon remain unchanged, while the shared parameters additionally receive supervision from the other horizons. Checkpoints are selected on validation data and evaluated separately on the corresponding horizon-specific test sets. The complete continent-level comparison across all three forecasting horizons is reported in Table 7.

Multi-Source NWP Analysis. We evaluate all seven nonempty subsets of $\{\mathrm{GFS}, \mathrm{EC}, \mathrm{TJ}\}$ using the same pretrained checkpoint without parameter updates. For a source subset $\mathcal{A}$, let $\widetilde{\mathbf{F}}_i^{\mathcal{A}}$ denote the normalized NWP trajectory produced by MeteoRouter when only the sources in $\mathcal{A}$ are available. Given a richer source set $\mathcal{A}$ and its nested subset $\mathcal{B} \subset \mathcal{A}$, the fusion-induced adjustment for instance i is measured by

$$d_i^{\mathcal{A},\mathcal{B}} = \left[\frac{1}{|\Omega_i| C} \sum_{\tau \in \Omega_i} \sum_{c=1}^{C} \left(\widetilde{F}_{i,\tau,c}^{\mathcal{A}} - \widetilde{F}_{i,\tau,c}^{\mathcal{B}} \right)^2 \right]^{1/2}, \tag{26}$$

where Ω_i contains the valid forecast steps and $C = 6$ is the number of meteorological variables. Since the trajectories are normalized using the same training-set statistics, this RMS measure is not dominated by differences in physical units or variable scales. We consider three comparison levels. The two-source versus single-source level contains six nested comparisons, with each of the three two-source combinations compared against each of its constituent single sources. The three-source versus two-source level compares the complete source set against each of the three two-source subsets, while the three-source versus single-source level compares it against each of the three individual sources. For the grouped analysis, 50 sites are randomly sampled from each region. Within each site and concrete source-set comparison, forecasting instances are ranked by $d_i^{\mathcal{A},\mathcal{B}}$ and partitioned into eight equal-frequency groups, G1–G8, from the smallest to the largest NWP adjustment. Group assignment is performed independently for each source-set comparison, such that the same forecasting instance may belong to different groups across comparisons. Each concrete richer–fewer source-set pair is treated as an independent paired comparison. For each comparison level and group, the MSEs

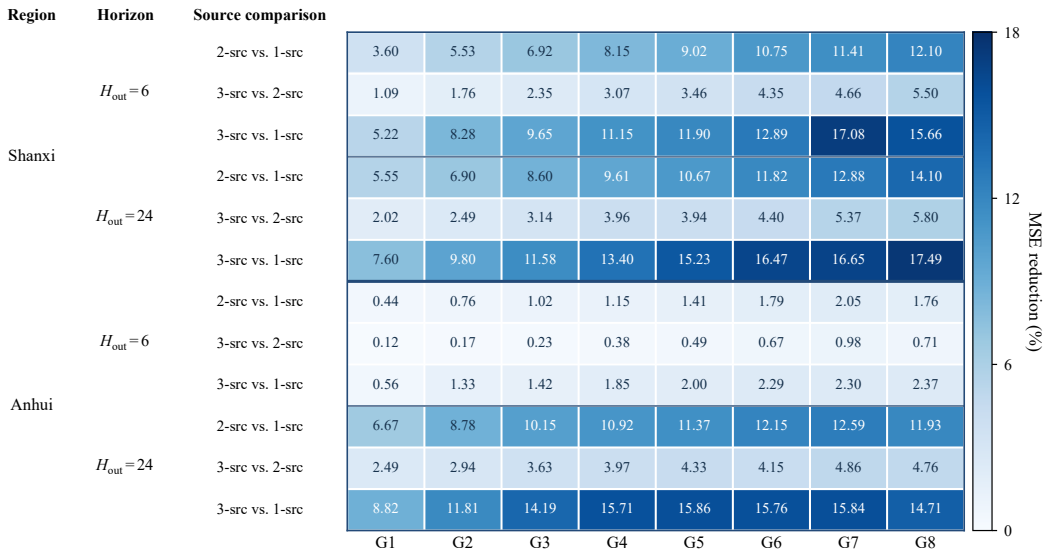

Figure 6: Complete grouped analysis of NWP adjustments induced by MeteoRouter. For each site, we compare pairs of configurations in which one uses more NWP sources and includes all the sources used by the other. Forecasting instances are independently divided into eight groups of equal size for each comparison, according to the root mean square difference between the corresponding normalized fused NWP trajectories. G1 and G8 represent the smallest and largest adjustments, respectively. Results for two sources versus one source are pooled across six pairs of configurations, whereas results for three sources versus two sources and three sources versus one source are each pooled across three pairs. Each cell reports the relative reduction in mean squared error achieved using more sources compared with fewer sources, calculated after pooling prediction errors over the instances assigned to the corresponding group for each site and pair.

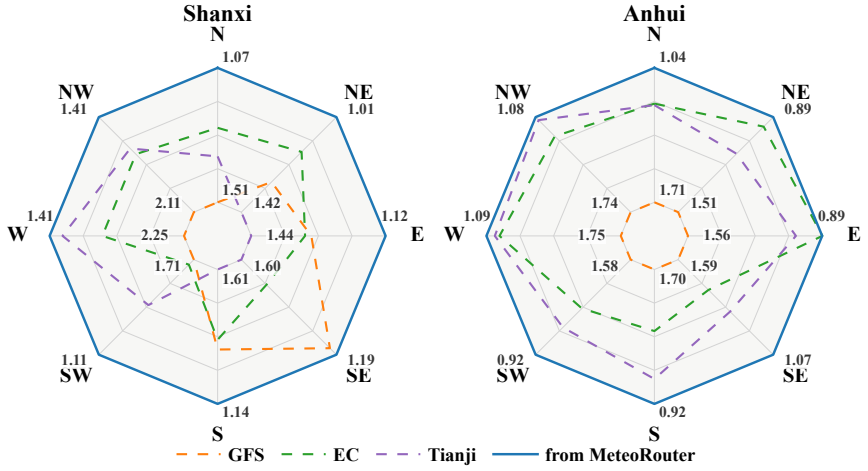

Figure 7: Wind-speed RMSE (m/s) relative to ERA5 in Shanxi (left) and Anhui (right), compared across eight wind directions. The lines correspond to GFS, EC, Tianji, and the NWP fused by MeteoRouter. Lower RMSE is plotted farther outward.

for configurations with more and fewer NWP sources are computed by pooling squared errors over all valid forecast points of the pair-specific group members across all constituent comparisons. The two MSEs use identical pair-specific group memberships and numbers of forecast points, and the same instance contributes once for every constituent comparison in which it is evaluated. The main text presents the three-source versus single-source subset, while Figure 6 reports the complete re-

sults for all three comparison levels. Across all three comparison levels, the complete results exhibit the same overall pattern: MSE reduction generally increases with the magnitude of fusion-induced NWP adjustments across regions and forecast horizons.

Moreover, we compare wind-speed forecast errors from the individual NWP sources with those from MeteoRouter fusion in both regions. As shown in Figure 7, no individual NWP source consistently provides the lowest wind-speed error across both regions and all wind directions. In contrast, the NWP fused by MeteoRouter achieves the lowest error in nearly all cases, providing further evidence for the effectiveness of adaptive multi-source NWP fusion. We further visualize three test samples under the $(72, 24)$ setting in Figure 8. In these samples, the fused wind trajectories more closely track ERA5, with power forecasts that better match the observations.

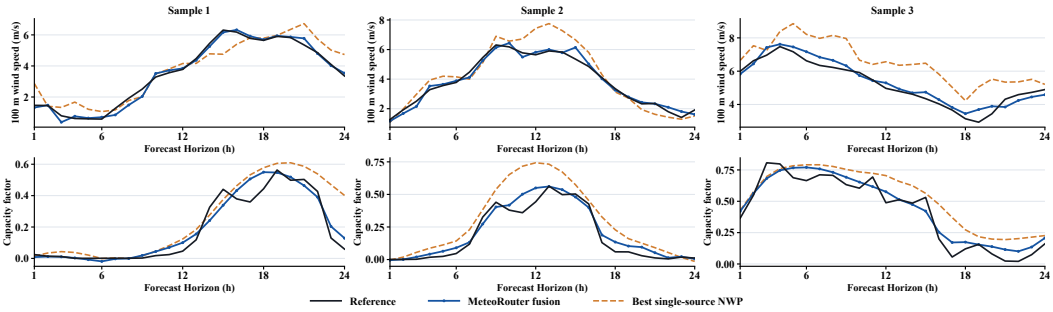

Figure 8: Visualization of MeteoRouter fusion and power forecasting results under the $(72, 24)$ context–horizon setting. The top row compares ERA5 wind speed with MeteoRouter fusion and the best single-source NWP, while the bottom row compares observed power with the corresponding power forecasts.

Table 8: Comparison of MeteoRouter and uniform averaging (Uniform) over available NWP sources. Lower values are better. **Bold**: best.

Setting	Horizon	MAE		MSE	
		Uniform	MeteoRouter	Uniform	MeteoRouter
Full-shot	6	0.0936	**0.0928**	0.0198	**0.0189**
	24	0.1081	**0.1072**	0.0261	**0.0250**
	168	0.1498	**0.1486**	0.0451	**0.0441**
Cross-site Zero-shot	6	0.1042	**0.1031**	0.0225	**0.0219**
	24	0.1212	**0.1197**	0.0307	**0.0294**
	168	0.1566	**0.1539**	0.0474	**0.0450**

To further assess the benefit of adaptive source weighting, we replace MeteoRouter with uniform averaging over the available NWP sources and pretrain the resulting variant on the same data with all other settings unchanged. Table 8 shows that MeteoRouter achieves lower MAE and MSE at all three horizons in both full-shot and cross-site zero-shot forecasting. Averaged across horizons, the relative MSE reductions are 3.66% and 3.99%, respectively. These results suggest that MeteoRouter's variable-wise routing improves forecasting by adaptively weighting reliable NWP sources.

Static-Context Similarity Analysis. The analysis uses the same 74 zero-shot sites as the cross-site zero-shot evaluation. For each target site, its proximity to the pretraining corpus is quantified independently from surface cover, terrain attributes, and AlphaEarth representations. Surface-cover proximity is measured using a six-dimensional vector comprising the proportions of five land-cover categories—cropland/grassland, forest/shrubland, mountain/hill, water/wetland, and complex/other—with a satellite-consistency score. The surface-cover distance between two sites is defined as the root-mean-square difference between their corresponding vectors. Terrain proximity is computed using all available tabular site attributes. These attributes are standardized using statistics estimated exclusively from pretraining sites with complete terrain information, after which the root-mean-square distance is calculated. For both distance-based measures, the target-site score is the mean distance to its ten nearest pretraining sites. AlphaEarth proximity is measured directly

Table 9: Effect of backbone scale on full-shot and cross-site zero-shot forecasting. L denotes the number of backbone blocks and D the hidden dimension. Lower values are better. **Bold**: best.

Configuration	Full-shot						Cross-site Zero-shot					
	6 h		24 h		168 h		6 h		24 h		168 h	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
$L = 2,\ D = 64$	0.0967	0.0198	0.1122	0.0263	0.1535	0.0450	0.1072	0.0228	0.1254	0.0307	0.1587	0.0454
$L = 3,\ D = 64$	0.0944	0.0192	0.1092	0.0255	0.1513	0.0449	0.1043	0.0221	0.1216	0.0299	0.1563	0.0455
$L = 2,\ D = 128$	0.0937	0.0190	0.1086	0.0252	0.1510	**0.0439**	0.1053	0.0226	0.1234	0.0306	0.1583	0.0460
$L = 3,\ D = 128$	**0.0928**	**0.0189**	**0.1072**	**0.0250**	**0.1486**	0.0441	**0.1031**	**0.0219**	**0.1197**	**0.0294**	**0.1539**	**0.0450**
$L = 3,\ D = 256$	0.0929	0.0192	**0.1072**	0.0255	0.1498	0.0462	0.1048	0.0230	0.1229	0.0316	0.1589	0.0493
$L = 4,\ D = 256$	0.0930	0.0193	0.1075	0.0257	0.1518	0.0477	0.1065	0.0237	0.1249	0.0325	0.1603	0.0504

from the original 64-dimensional embeddings. Cosine similarities are computed between each target site and all pretraining sites, and the mean similarity to the ten most similar sites is retained. The zero-shot sites are independently sorted by each static measure and partitioned into three contiguous groups. Lower surface-cover and terrain distances indicate greater proximity, whereas higher AlphaEarth similarity indicates greater proximity.

Temporal Interaction Ablation. We examine the contributions of MPE and MSFFN to temporal modeling and assess the effect of introducing temporal attention. To this end, we conduct controlled ablations using sites from nine datasets (D06, D08, D13, D17, D23, D27, D28, D31, and D33 in Table 5), with the test interval of each participating site used for evaluation. All variants adopt identical data partitions, input modalities, preprocessing, model dimensions, optimization settings, and training schedules, differing only in their temporal-interaction pathways. We evaluate five configurations: (i) the complete MPE+MSFFN design without temporal attention; (ii) MSFFN without MPE; (iii) MPE with MSFFN replaced by a standard position-wise FFN; (iv) temporal attention without MPE, with MSFFN replaced by a standard position-wise FFN; and (v) the complete MPE+MSFFN design augmented with temporal attention. Predictive accuracy is evaluated using the MSE averaged across the three forecast horizons. Computational efficiency is assessed by comparing the mean per-batch inference latency of MPE+MSFFN and the temporal-attention variant under identical input settings, batch size, and hardware, with forecast horizons ranging from 6 to 720 hours.

C.11 Model Scaling

We examine model capacity by varying the number of backbone blocks L and the hidden dimension D, while keeping all other architectural and optimization settings fixed. Table 9 reports full-shot and zero-shot performance across three forecast horizons. The configuration, $(L, D) = (3, 128)$, achieves the best result in eleven of the twelve comparisons and consistently performs best across all zero-shot settings. Wider or deeper variants yield no considerable improvement and generally weaken cross-site transfer, indicating that W2P-FM benefits from a compact, appropriately sized backbone.

C.12 Pretraining Data Scaling

We investigate whether increasing pretraining data yields sustained improvements in W2P-FM's forecasting accuracy and cross-site generalization. To this end, we retain 25%, 50%, 75%, or 100% of each pretraining site's original training data. We pretrain W2P-FM from scratch on each training subset using the same architecture and optimization settings. The validation and test sets are kept fixed, and all experiments use normalization statistics computed from the full training set.

The experimental results in Figure 9 show that MSE consistently decreases with increasing pretraining data across all three horizons in both full-shot and cross-site zero-shot forecasting. Compared with the 25% subset, training on the full dataset reduces MSE by an average of 20.8% in full-shot forecasting and 19.5% in cross-site zero-shot forecasting. Increasing the training fraction from 75% to 100% still yields average reductions of 6.6% and 6.3%, respectively. These results show that scaling pretraining data improves both forecasting accuracy and generalization to unseen sites. The continued gains suggest that performance has not yet saturated within the evaluated range and that a larger pretraining corpus could yield further improvements.

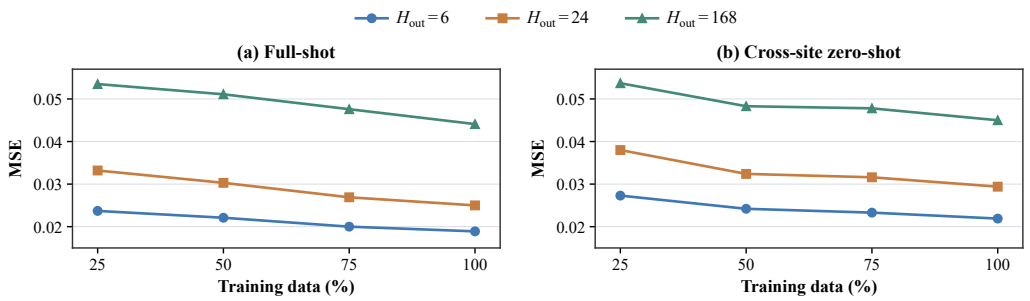

Figure 9: Effect of pretraining data volume on (a) full-shot and (b) cross-site zero-shot forecasting across three horizons. Percentages refer to the retained fraction of each site's original training period. Lower MSE is better.

D LIMITATIONS

Data coverage and quality. The coverage and quality of our corpus are constrained by the available source data. Although it spans five continents, geographic coverage remains uneven, with no sites from Africa and limited representation in North America. The availability and reliability of NWP inputs also vary across sources, regions, and forecast horizons. In addition, power series are provided at different aggregation levels, from individual turbines to wind farms and regional units, depending on the original source. Where the underlying turbine records are unavailable, aggregated generation cannot be converted into measured turbine-level series through preprocessing alone. These differences limit the detail available for characterizing generation behavior within aggregated units.

Static site context. Source datasets differ in coordinate accuracy and the availability and detail of site attributes, making it difficult to derive unified static site descriptions from raw information alone. The corresponding generation units also vary considerably in size and spatial extent. Consequently, context extracted around the available coordinates may not fully represent the terrain and surrounding environment of geographically dispersed turbines, particularly for entire wind farms or regional aggregation units. Moreover, the benefits of static context conditioning depend partly on how well similar target-site contexts are represented in the available pretraining corpus and may diminish under substantial static-context shifts.

Operational deployment. Our evaluation is based on offline training and test data. Assessing deployment benefits requires additional site-specific operational information and integration with the decision-making processes of wind-farm or system operators. The extent to which the observed forecasting improvements translate into practical benefits for generation scheduling, reserve allocation, and electricity-market operation remains to be validated.

E COMPLETE EXPERIMENTAL RESULTS

Complementing the aggregate comparisons in the main text, complete continent-level full-shot results are reported in Tables 10–12. Detailed results for probabilistic and cross-granularity adaptation are presented in Tables 13 and 14, respectively, while inference efficiency is reported in Table 15.

Table 10: Complete continent-level full-shot results for the 6-hour forecast horizon. Results are reported as mean ± standard deviation over three runs. Lower values are better. **Bold**: best mean; underlined: second-best mean.

Model	Metric	Asia	Europe	North America	South America	Oceania
W2P-FM	MAE	**0.0928** ± 0.0003	**0.0870** ± 0.0008	**0.1008** ± 0.0001	**0.1024** ± 0.0005	**0.0924** ± 0.0003
	MSE	**0.0190** ± 0.0001	**0.0178** ± 0.0001	**0.0218** ± 0.0002	**0.0203** ± 0.0001	**0.0180** ± 0.0000
W2P-FM-NS	MAE	0.1030 ± 0.0008	0.0922 ± 0.0004	0.1032 ± 0.0004	0.1044 ± 0.0001	0.0945 ± 0.0006
	MSE	0.0230 ± 0.0003	0.0194 ± 0.0000	0.0224 ± 0.0001	0.0214 ± 0.0000	0.0187 ± 0.0001
Crossformer	MAE	0.1219 ± 0.0020	0.0991 ± 0.0008	0.1076 ± 0.0004	0.1070 ± 0.0002	0.0954 ± 0.0002
	MSE	0.0281 ± 0.0001	0.0205 ± 0.0004	0.0233 ± 0.0001	0.0217 ± 0.0001	0.0188 ± 0.0000
DUET	MAE	0.1288 ± 0.0002	0.1311 ± 0.0004	0.1564 ± 0.0002	0.1260 ± 0.0000	0.1356 ± 0.0002
	MSE	0.0333 ± 0.0000	0.0349 ± 0.0002	0.0449 ± 0.0001	0.0298 ± 0.0000	0.0344 ± 0.0001
TimeBridge	MAE	0.1194 ± 0.0003	0.1221 ± 0.0016	0.1349 ± 0.0002	0.1165 ± 0.0000	0.1232 ± 0.0004
	MSE	0.0297 ± 0.0001	0.0314 ± 0.0006	0.0354 ± 0.0000	0.0263 ± 0.0000	0.0296 ± 0.0002
TimeXer	MAE	0.1205 ± 0.0007	0.1265 ± 0.0001	0.1444 ± 0.0002	0.1180 ± 0.0001	0.1243 ± 0.0002
	MSE	0.0298 ± 0.0003	0.0332 ± 0.0000	0.0394 ± 0.0000	0.0267 ± 0.0001	0.0297 ± 0.0001
XLinear	MAE	0.1225 ± 0.0003	0.1256 ± 0.0001	0.1491 ± 0.0000	0.1189 ± 0.0000	0.1346 ± 0.0007
	MSE	0.0311 ± 0.0002	0.0332 ± 0.0000	0.0423 ± 0.0000	0.0271 ± 0.0001	0.0350 ± 0.0004
P-GRU	MAE	0.1211 ± 0.0010	0.1127 ± 0.0030	0.1119 ± 0.0011	0.1146 ± 0.0000	0.0991 ± 0.0000
	MSE	0.0295 ± 0.0008	0.0254 ± 0.0012	0.0246 ± 0.0003	0.0241 ± 0.0000	0.0199 ± 0.0000
C-LSTM	MAE	0.1369 ± 0.0071	0.1279 ± 0.0035	0.1229 ± 0.0032	0.1182 ± 0.0039	0.1071 ± 0.0002
	MSE	0.0353 ± 0.0036	0.0326 ± 0.0018	0.0287 ± 0.0012	0.0256 ± 0.0017	0.0223 ± 0.0001
iTransformer	MAE	0.1197 ± 0.0002	0.1120 ± 0.0002	0.1133 ± 0.0018	0.1126 ± 0.0002	0.1038 ± 0.0000
	MSE	0.0280 ± 0.0000	0.0266 ± 0.0001	0.0261 ± 0.0004	0.0244 ± 0.0000	0.0221 ± 0.0000
LightGBM	MAE	0.1170 ± 0.0002	0.1029 ± 0.0001	0.1066 ± 0.0000	0.1076 ± 0.0003	0.0984 ± 0.0002
	MSE	0.0272 ± 0.0001	0.0230 ± 0.0001	0.0231 ± 0.0001	0.0233 ± 0.0001	0.0204 ± 0.0000

Table 11: Complete continent-level full-shot results for the 24-hour forecast horizon. Results are reported as mean ± standard deviation over three runs. Lower values are better. **Bold**: best mean; underlined: second-best mean.

Model	Metric	Asia	Europe	North America	South America	Oceania
W2P-FM	MAE	**0.1112** ± 0.0009	**0.0993** ± 0.0011	**0.1137** ± 0.0002	**0.1151** ± 0.0001	**0.1045** ± 0.0002
	MSE	**0.0260** ± 0.0001	**0.0240** ± 0.0001	**0.0277** ± 0.0004	**0.0248** ± 0.0002	0.0235 ± 0.0001
W2P-FM-NS	MAE	0.1235 ± 0.0012	0.1076 ± 0.0004	0.1158 ± 0.0002	0.1153 ± 0.0001	0.1057 ± 0.0010
	MSE	0.0318 ± 0.0001	0.0270 ± 0.0000	0.0281 ± 0.0001	0.0251 ± 0.0000	**0.0232** ± 0.0003
Crossformer	MAE	0.1444 ± 0.0055	0.1205 ± 0.0010	0.1233 ± 0.0000	0.1212 ± 0.0009	0.1096 ± 0.0002
	MSE	0.0384 ± 0.0025	0.0300 ± 0.0004	0.0296 ± 0.0004	0.0266 ± 0.0003	0.0248 ± 0.0000
DUET	MAE	0.1735 ± 0.0003	0.1919 ± 0.0003	0.2038 ± 0.0003	0.1494 ± 0.0018	0.1784 ± 0.0005
	MSE	0.0537 ± 0.0001	0.0659 ± 0.0000	0.0674 ± 0.0002	0.0391 ± 0.0008	0.0539 ± 0.0002
TimeBridge	MAE	0.1622 ± 0.0001	0.1723 ± 0.0013	0.1641 ± 0.0008	0.1353 ± 0.0013	0.1504 ± 0.0012
	MSE	0.0482 ± 0.0002	0.0565 ± 0.0009	0.0473 ± 0.0001	0.0328 ± 0.0005	0.0404 ± 0.0004
TimeXer	MAE	0.1661 ± 0.0006	0.1840 ± 0.0003	0.1824 ± 0.0001	0.1368 ± 0.0005	0.1590 ± 0.0002
	MSE	0.0503 ± 0.0004	0.0626 ± 0.0003	0.0569 ± 0.0008	0.0337 ± 0.0001	0.0451 ± 0.0001
XLinear	MAE	0.1701 ± 0.0011	0.1805 ± 0.0010	0.1949 ± 0.0001	0.1391 ± 0.0003	0.1816 ± 0.0005
	MSE	0.0521 ± 0.0007	0.0605 ± 0.0003	0.0628 ± 0.0000	0.0347 ± 0.0001	0.0552 ± 0.0001
P-GRU	MAE	0.1469 ± 0.0000	0.1241 ± 0.0006	0.1259 ± 0.0003	0.1287 ± 0.0010	0.1135 ± 0.0004
	MSE	0.0402 ± 0.0007	0.0311 ± 0.0001	0.0303 ± 0.0004	0.0297 ± 0.0006	0.0259 ± 0.0001
C-LSTM	MAE	0.1602 ± 0.0045	0.1641 ± 0.0026	0.1337 ± 0.0007	0.1374 ± 0.0003	0.1226 ± 0.0002
	MSE	0.0441 ± 0.0023	0.0549 ± 0.0016	0.0335 ± 0.0003	0.0332 ± 0.0001	0.0297 ± 0.0001
iTransformer	MAE	0.1546 ± 0.0003	0.1576 ± 0.0028	0.1325 ± 0.0003	0.1321 ± 0.0011	0.1215 ± 0.0003
	MSE	0.0439 ± 0.0005	0.0490 ± 0.0022	0.0338 ± 0.0001	0.0315 ± 0.0005	0.0293 ± 0.0001
LightGBM	MAE	0.1422 ± 0.0001	0.1168 ± 0.0000	0.1182 ± 0.0001	0.1252 ± 0.0000	0.1122 ± 0.0001
	MSE	0.0375 ± 0.0001	0.0299 ± 0.0002	0.0282 ± 0.0001	0.0298 ± 0.0000	0.0262 ± 0.0000

Table 12: Complete continent-level full-shot results for the 168-hour forecast horizon. Results are reported as mean ± standard deviation over three runs. Lower values are better. **Bold**: best mean; underlined: second-best mean.

Model	Metric	Asia	Europe	North America	South America	Oceania
W2P-FM	MAE	**0.1467** ± 0.0008	**0.1516** ± 0.0004	**0.1679** ± 0.0006	**0.1479** ± 0.0006	**0.1371** ± 0.0004
	MSE	**0.0422** ± 0.0000	**0.0496** ± 0.0003	**0.0513** ± 0.0000	**0.0393** ± 0.0003	**0.0361** ± 0.0002
W2P-FM-NS	MAE	0.1548 ± 0.0005	0.1588 ± 0.0004	0.1691 ± 0.0008	0.1487 ± 0.0004	0.1384 ± 0.0005
	MSE	0.0468 ± 0.0002	0.0532 ± 0.0007	0.0521 ± 0.0000	0.0398 ± 0.0004	**0.0361** ± 0.0000
Crossformer	MAE	0.1837 ± 0.0012	0.1981 ± 0.0052	0.1840 ± 0.0028	0.1548 ± 0.0011	0.1458 ± 0.0000
	MSE	0.0566 ± 0.0010	0.0699 ± 0.0029	0.0579 ± 0.0001	0.0414 ± 0.0006	0.0395 ± 0.0000
DUET	MAE	0.2052 ± 0.0002	0.2313 ± 0.0018	0.2353 ± 0.0016	0.1780 ± 0.0003	0.2153 ± 0.0002
	MSE	0.0728 ± 0.0002	0.0926 ± 0.0013	0.0866 ± 0.0006	0.0528 ± 0.0003	0.0731 ± 0.0001
TimeBridge	MAE	0.2005 ± 0.0002	0.2195 ± 0.0015	0.2081 ± 0.0010	0.1626 ± 0.0003	0.1878 ± 0.0005
	MSE	0.0698 ± 0.0000	0.0836 ± 0.0016	0.0697 ± 0.0005	0.0456 ± 0.0002	0.0580 ± 0.0003
TimeXer	MAE	0.2041 ± 0.0000	0.2286 ± 0.0014	0.2275 ± 0.0004	0.1707 ± 0.0007	0.2042 ± 0.0010
	MSE	0.0713 ± 0.0001	0.0911 ± 0.0002	0.0836 ± 0.0001	0.0496 ± 0.0000	0.0685 ± 0.0004
XLinear	MAE	0.2030 ± 0.0004	0.2279 ± 0.0002	0.2269 ± 0.0000	0.1732 ± 0.0003	0.2189 ± 0.0001
	MSE	0.0707 ± 0.0007	0.0871 ± 0.0005	0.0809 ± 0.0007	0.0505 ± 0.0001	0.0735 ± 0.0001
P-GRU	MAE	0.1754 ± 0.0013	0.1849 ± 0.0003	0.1827 ± 0.0010	0.1569 ± 0.0005	0.1487 ± 0.0010
	MSE	0.0561 ± 0.0017	0.0622 ± 0.0002	0.0572 ± 0.0001	0.0426 ± 0.0003	0.0401 ± 0.0003
C-LSTM	MAE	0.1902 ± 0.0010	0.2152 ± 0.0005	0.1917 ± 0.0001	0.1697 ± 0.0017	0.1594 ± 0.0014
	MSE	0.0618 ± 0.0003	0.0798 ± 0.0006	0.0615 ± 0.0000	0.0486 ± 0.0005	0.0449 ± 0.0005
iTransformer	MAE	0.1946 ± 0.0009	0.2196 ± 0.0002	0.1988 ± 0.0010	0.1649 ± 0.0011	0.1690 ± 0.0003
	MSE	0.0629 ± 0.0003	0.0821 ± 0.0011	0.0652 ± 0.0004	0.0466 ± 0.0006	0.0487 ± 0.0001
LightGBM	MAE	0.1839 ± 0.0001	0.1789 ± 0.0003	0.1782 ± 0.0001	0.1641 ± 0.0001	0.1504 ± 0.0001
	MSE	0.0558 ± 0.0000	0.0588 ± 0.0001	0.0553 ± 0.0000	0.0474 ± 0.0001	0.0418 ± 0.0001

Table 13: Complete probabilistic forecasting results averaged equally over seven datasets. Chronos-2 and Moirai 2.0 are initialized from their official pretrained checkpoints and fine-tuned for probabilistic forecasting. Results are reported as mean ± standard deviation over three runs. Lower values are better. **Bold**: best mean; underlined: second-best mean.

Model	6 h		24 h	
	CRPS	IS	CRPS	IS
DeepAR	0.0999±0.0011	0.6445±0.0009	0.1134±0.0005	0.7658±0.0011
Crossformer+MQL	0.0893±0.0008	0.5682±0.0103	0.1006±0.0003	0.6419±0.0014
TMDM	0.0963±0.0029	0.6431±0.0155	0.1376±0.0404	0.9908±0.3910
NsDiff	0.1273±0.0028	0.8348±0.0200	0.1388±0.0027	0.8995±0.0104
Chronos-2	0.0973±0.0007	0.6133±0.0027	0.1031±0.0010	0.6291±0.0039
Moirai 2.0	0.1097±0.0016	0.6619±0.0066	0.1449±0.0013	0.8220±0.0038
Head-only FT	0.0769±0.0008	0.4884±0.0062	0.0828±0.0007	0.5211±0.0059
Full FT	**0.0755**±0.0007	**0.4716**±0.0015	**0.0809**±0.0006	**0.5048**±0.0049

Table 14: Complete cross-granularity forecasting results at resolutions of 15 and 10 minutes. I/O Adapt. denotes lightweight adaptation of the granularity-specific power patch projection and forecast head with the shared backbone frozen. Chronos-2 and Moirai 2.0 are initialized from their official pretrained checkpoints and fine-tuned for forecasting at the corresponding temporal granularity. Results are reported as mean ± standard deviation over three runs. Lower values are better. **Bold**: best mean; underlined: second-best mean.

Granularity	Model	6 h		24 h	
		MAE	MSE	MAE	MSE
15 min	Crossformer	0.1243±0.0128	0.0280±0.0024	0.1449±0.0003	0.0360±0.0010
	iTransformer	0.1171±0.0005	0.0269±0.0009	0.1675±0.0014	0.0456±0.0047
	P-GRU	0.1152±0.0006	0.0252±0.0005	0.1403±0.0008	0.0346±0.0003
	Chronos-2	0.1096±0.0019	0.0284±0.0006	0.1256±0.0009	0.0346±0.0002
	Moirai 2.0	0.1215±0.0021	0.0340±0.0013	0.1746±0.0005	0.0589±0.0016
	I/O Adapt.	**0.0963**±0.0002	**0.0199**±0.0003	**0.1143**±0.0002	0.0271±0.0006
	Full FT	0.0990±0.0001	0.0203±0.0002	0.1149±0.0009	**0.0267**±0.0001
10 min	Crossformer	0.1175±0.0002	0.0254±0.0019	0.1368±0.0022	0.0330±0.0001
	iTransformer	0.1226±0.0045	0.0275±0.0001	0.1734±0.0026	0.0487±0.0005
	P-GRU	0.1322±0.0010	0.0297±0.0001	0.1216±0.0010	0.0270±0.0028
	Chronos-2	0.1245±0.0018	0.0311±0.0005	0.1303±0.0011	0.0336±0.0007
	Moirai 2.0	0.1366±0.0015	0.0363±0.0017	0.1790±0.0006	0.0585±0.0015
	I/O Adapt.	**0.1068**±0.0028	**0.0215**±0.0004	**0.1142**±0.0014	**0.0241**±0.0010
	Full FT	0.1157±0.0003	0.0241±0.0008	0.1216±0.0006	0.0267±0.0001

Table 15: Per-batch inference latency of task-specific and foundation models under the context–horizon configuration (72, 168). All models are evaluated on the same hardware with a batch size of 256. Results are reported in milliseconds as mean ± standard deviation.

Task-Specific Models		Foundation Models	
Model	Latency (ms/batch)	Model	Latency (ms/batch)
Crossformer	25.558 ± 0.011	Chronos-2	228.10 ± 1.18
DUET	3.356 ± 0.010	TiRex-2	1035.65 ± 0.94
TimeBridge	15.414 ± 0.005	CITRAS-FM	496.78 ± 0.41
TimeXer	3.361 ± 0.005	Moirai 2.0	569.48 ± 0.67
XLinear	0.495 ± 0.002	WindFM	3468.21 ± 2.11
P-GRU	2.992 ± 0.000	W2P-FM-NS	18.253 ± 0.005
C-LSTM	0.468 ± 0.002	W2P-FM	20.158 ± 0.003
iTransformer	1.548 ± 0.007		