ALM: LIGHTWEIGHT ADAPTATION OF FOUNDATION MODELS FOR JOINT WEATHER-POWER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Weather information is crucial for forecasting power variables such as wind power, photovoltaic power, and load. Existing power forecasting methods typically treat weather variables as exogenous inputs and generate site-level forecasts. However, such localized modeling makes it difficult to capture how continuous weather processes propagate across space, evolve over time, and persistently affect power dynamics. Weather foundation models are well suited to representing these continuous atmospheric processes, yet how to extend their forecasting capability to joint weather-power prediction remains insufficiently explored. To address these limitations, we propose **ALM**, a lightweight adaptation approach that supports weather-power joint forecasting and bidirectional information exchange between weather and power backbones. ALM consists of three steps: prefusion **A**lignment, **L**atent memory, and postfusion **M**odulation, which support joint improvements in weather and power forecasting while keeping most backbone parameters frozen. We evaluate the method on a gridded weather-power benchmark for the contiguous United States (CONUS) and further validate its forecasts in downstream power applications. Compared with full fine-tuning, ALM reduces mean 72-hour RMSE by 19.9% across 12 reported variables while using only 9.7% trainable parameters. ALM also reaches the shortest per-epoch training time among all compared methods. Our code is available in the Supplementary Material.

1 INTRODUCTION

Weather information is essential for forecasting renewable generation and electricity demand (Wang et al., 2021; Chen et al., 2023; Chu et al., 2024). Current statistical and machine learning methods typically use weather variables as exogenous inputs and predict these power variables at individual sites or stations (Verdone et al., 2025; Chu et al., 2024). Such local modeling primarily characterizes weather-power relationships for individual prediction targets. It struggles to capture persistent weather effects on power because weather evolves through continuous spatiotemporal processes (Verdone et al., 2025). Yet renewable energy forecasting at multi-site and regional scales remains less studied than site-level prediction (Verdone et al., 2025; Chu et al., 2024).

Numerical weather prediction (NWP) represents these atmospheric processes by solving physical equations, but requires substantial computational resources (Bauer et al., 2015). With recent advances in deep learning, models such as Pangu-Weather and GraphCast have achieved accurate and efficient global weather forecasts (Bi et al., 2023; Lam et al., 2023). Foundation models such as ClimaX and Aurora provide reusable representations across variables, resolutions, and forecasting tasks (Nguyen et al., 2023; Bodnar et al., 2025). FuXi-Energy further shows that forecasts from deep weather models can support renewable energy prediction (Zhong et al., 2026). These weather foundation models offer an effective means of representing continuous atmospheric processes. However, transferring these representations to power forecasting through full fine-tuning is costly and may weaken the model's original weather prediction performance. This creates a pressing need for lightweight adaptation that incorporates power variables while preserving weather forecasting performance.

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.

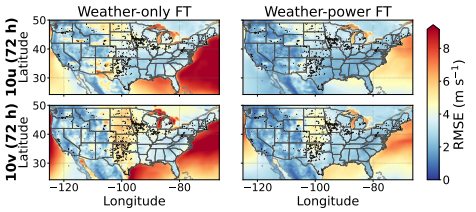


Figure 1: Comparison of Aurora forecast RMSE before (left) and after (right) incorporating power variables. The introduction of wind-power variables helps reduce forecast errors surrounding the power plant (marked by black triangles).

Moreover, existing methods largely assume a unidirectional information flow in which weather supports power prediction and power remains only a target (Wang et al., 2021; Zhong et al., 2026). Our preliminary experiments show that adding gridded power variables to an Aurora-based model reduces forecast errors for some weather variables (Figure 1). This result suggests that power variables may contain useful information about regional weather states. A small number of NWP and data assimilation studies have used turbine anemometer or supervisory control and data acquisition (SCADA) wind observations to improve local wind simulations (Cheng et al., 2017; Ivanova et al., 2025). However, it remains underexplored whether a weather-power forecasting model jointly trained on both prediction objectives can effectively incorporate power-side representations into the weather branch and improve performance on both tasks.

To address these questions, we introduce **ALM**, a lightweight adapter that enables bidirectional information exchange between pretrained weather and power models to improve their forecasts. ALM first applies *prefusion alignment* to reduce the representation mismatch between weather and power features. At multiple backbone levels, *latent memory* provides a shared space for aligned weather and power representations to interact. Finally, *postfusion modulation* regulates how each branch uses the fused information for prediction. ALM enables such bidirectional interaction while leaving most backbone parameters frozen. We evaluate ALM on joint forecasting of spatiotemporally aligned, gridded weather and power variables. At the 72-hour lead time, ALM achieves the lowest RMSE for all 12 weather and power variables and the best result on 23 of the 24 reported metrics. It also attains the largest mean RMSE improvement (33.7%) over the zero-shot models and the shortest per-epoch training time among the joint adaptation methods, with 90.3% fewer trainable parameters than full fine-tuning.

Our contributions are threefold:

- We propose an adaptation framework that extends weather foundation models to weather-power joint forecasting. The framework incorporates spatiotemporally continuous weather representations into power forecasting, while power representations in turn inform weather prediction.
- We propose **ALM**, a lightweight adaptation module that combines *prefusion alignment*, *multi-scale latent memory*, and *postfusion modulation* to reduce the number of trainable parameters and per-epoch training time when adapting pretrained models.
- We conduct extensive experiments on a unified gridded weather-power benchmark. At the 72-hour lead time, ALM improves mean RMSE over the corresponding zero-shot baselines by 33.7% and reduces downstream RMSE by 54.5%, 20.3%, and 66.7% for plant-level photovoltaic, plant-level wind, and county-level load forecasting, respectively. It also has the shortest per-epoch training time among joint adaptation methods, 4.2% below the next-fastest method.

2 Related Works

Weather-Power Forecasting. Weather information is widely used for wind power, photovoltaic power, and load forecasting. Recent studies exploit spatially resolved or learned weather forecasts for renewable power prediction (Keisler & Le Naour, 2025; Zhong et al., 2026), while joint forecasting models learn weather-power dependencies through multi-task objectives or shared representations (Wei et al., 2023; Chen et al., 2026). Knowledge from weather foundation models has also been transferred to photovoltaic forecasting through spatiotemporal knowledge distillation (He et al., 2025). Nevertheless, most existing approaches remain task-specific and primarily use weather information to support power prediction, leaving power-conditioned weather correction within a joint forecasting model largely unexplored.

Parameter-Efficient Adaptation of Foundation Models. Parameter-efficient fine-tuning (PEFT) adapts pretrained models by updating a small parameter subset, including adapter modules (Houlsby

et al., 2019), prompt tuning (Lester et al., 2021), and low-rank updates (Hu et al., 2022). For weather foundation models, recent work has explored personalized or task-adaptive lightweight adaptation (Chen et al., 2024; Cao et al., 2026), alongside general-purpose Earth-system foundation models (Bodnar et al., 2025). Pretrained time-series or language models have likewise been adapted for wind power and load forecasting (Yan et al., 2025; Fan et al., 2025; Liao et al., 2025; Meng et al., 2026). These lines of work generally adapt weather and power models separately, rather than enabling parameter-efficient interaction between their representations.

Existing studies therefore cover weather-power forecasting and foundation-model adaptation, but rarely unify them within a reciprocal weather-power forecasting model. FuXi-Energy uses energy-oriented weather forecasts for downstream wind and solar prediction (Zhong et al., 2026), but does not jointly adapt the weather and power branches. STKD-PV distills weather foundation model knowledge for photovoltaic forecasting (He et al., 2025), but supports only unidirectional weather-to-power transfer. Overall, power variables are seldom used through an explicit power-to-weather information pathway for atmospheric prediction, and atmospheric foundation-model representations are only beginning to be transferred to power forecasting. An efficient adaptation mechanism that allows weather and power representations to exchange information in both directions, while preserving the forecasting capability of the pretrained branches, remains underexplored.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let $\mathcal{D}_{\mathrm{T}}$ denote a target-domain dataset in which weather and power variables are spatiotemporally aligned on a common grid. At forecast initialization time t, the surface weather fields, pressure-level weather fields, and gridded power fields are represented separately as

$$X_t^s \in \mathbb{R}^{T_{\mathrm{in}} \times V_s \times H \times W}, \qquad X_t^a \in \mathbb{R}^{T_{\mathrm{in}} \times V_a \times L \times H \times W}, \qquad X_t^p \in \mathbb{R}^{T_{\mathrm{in}} \times V_p \times H \times W}, \qquad T_{\mathrm{in}} = 2. \quad (1)$$

where T_{in} is the number of historical states, V_s and V_a count surface and atmospheric variables, L is the number of pressure levels, and H and W are the latitude and longitude grid dimensions. The power field comprises $V_p = 3$ variables: wind power, photovoltaic power, and load. We assume access to a pretrained weather foundation model and a pretrained power model, represented by the following one-step forecast mappings with a 6-hour lead time:

$$f^w_{\theta^w_0} : (X_t^s, X_t^a) \mapsto (Y_{t+\delta}^s, Y_{t+\delta}^a), \qquad f^p_{\theta^p_0} : X_t^p \mapsto Y_{t+\delta}^p, \qquad \delta = 6\ \mathrm{h}. \quad (2)$$

where $f^w_{\theta^w_0}$ and $f^p_{\theta^p_0}$ are the pretrained predictors. The joint forecasting model is defined as

$$\hat{\mathcal{Y}}_{t+\delta} = f_{\Theta_0,\phi}(\mathcal{X}_t), \qquad \Theta_0 = (\theta_0^w, \theta_0^p). \quad (3)$$

where $\mathcal{X}_t = (X_t^s, X_t^a, X_t^p)$ and $\mathcal{Y}_{t+\delta} = (Y_{t+\delta}^s, Y_{t+\delta}^a, Y_{t+\delta}^p)$ denote the joint input and target, respectively. Let $\Theta_{\mathrm{t}} \subseteq \Theta_0$ denote the selected trainable subset of pretrained parameters, let ϕ denote newly introduced trainable parameters, and let Θ denote the full parameter set after these parameters are introduced. The target-domain adaptation problem is formulated:

$$(\Theta_{\mathrm{t}}^\star, \phi^\star) = \arg\min_{\Theta_{\mathrm{t}},\phi} \mathbb{E}_{(\mathcal{X},\mathcal{Y})\sim\mathcal{D}_{\mathrm{T}}}[\mathcal{L}(f_{\Theta_{\mathrm{t}},\phi}(\mathcal{X}; \Theta_0 \setminus \Theta_{\mathrm{t}}), \mathcal{Y})] \quad \text{s.t.} \quad \frac{|\Theta_{\mathrm{t}}| + |\phi|}{|\Theta|} \le \rho, \quad (4)$$

where $\mathcal{L}$ is the joint prediction loss and ρ is a trainable-parameter ratio bound. The notation $f_{\Theta_{\mathrm{t}},\phi}(\cdot; \Theta_0 \setminus \Theta_{\mathrm{t}})$ indicates that only the selected subset Θ_{t} and the newly introduced parameters ϕ are updated. Accordingly, the joint forecasting model is adapted by optimizing the selected pretrained subset Θ_{t} and the newly introduced parameters ϕ under a limited trainable-parameter budget.

3.2 WEATHER-POWER JOINT FINE-TUNING FRAMEWORK

Figure 2(a) shows how pretrained Aurora (Bodnar et al., 2025) and PowerSwin models exchange representations through three scale-specific ALM blocks. PowerSwin is a two-dimensional Swin-Transformer U-Net whose hierarchy is analogous to Aurora. Neither backbone architecture is mandatory, provided that the weather and power features can be dimensionally aligned with ALM. Appendix D.4 reports cross-backbone results using ClimaX and PowerConv. Both model branches comprise an Encoder, a Backbone, and a Decoder. The Encoder tokenizes physical fields, the Backbone contains three resolution-reducing encoder stages followed by three resolution-restoring decoder stages, and the Decoder reconstructs forecast fields.

Because Aurora represents atmospheric variables with level-aware tokens whereas the power Backbone operates on two-dimensional tokens, a learnable pseudo-level bridge first expands the power

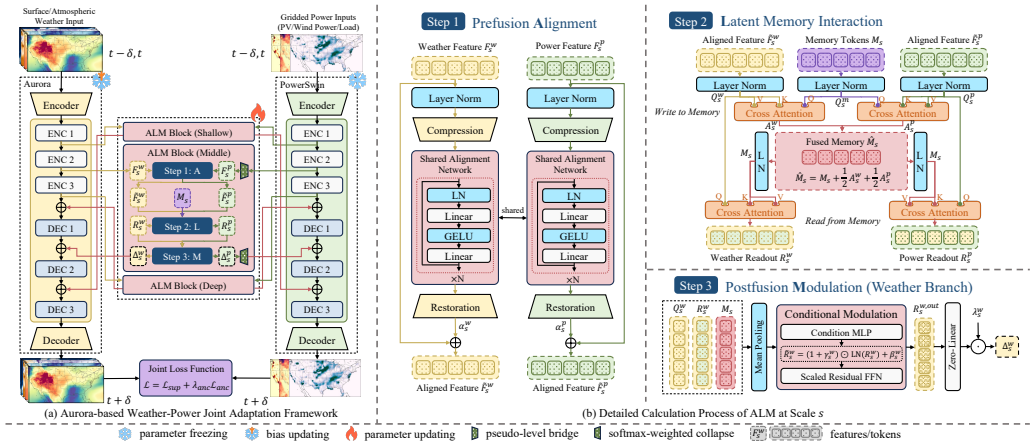


Figure 2: Overview of the proposed weather-power joint forecasting framework. (a) Shallow, middle, and deep ALM blocks mediate bidirectional information exchange between the pretrained Aurora and PowerSwin models. Most backbone parameters remain frozen during joint adaptation. (b) At each scale, ALM performs prefusion alignment, bidirectional write/read interaction through latent memory, and branch-specific postfusion modulation. Step 3 shows postfusion modulation for the weather branch. The power-side postfusion modulator has the same structure but separate parameters.

tokens using level embeddings before they enter ALM. Let $\mathcal{S} = \{\text{shallow}, \text{middle}, \text{deep}\}$ denote the three interaction scales. The weather and power branches respectively pass the following multi-scale features to the ALM blocks:

$$\{F_s^w\}_{s\in\mathcal{S}} = B_{\text{enc}}^w(E^w(X_t^s, X_t^a)), \qquad \{F_s^p\}_{s\in\mathcal{S}} = B_{\text{enc}}^p(E^p(X_t^p)). \tag{5}$$

where $b \in \{w, p\}$ denotes the weather and power branches, E^b is the top-level Encoder, B_{enc}^b denotes the Backbone encoder stages, and F_s^b is the feature passed to ALM at scale s. The channel widths are $d_s \in \{d, 2d, 4d\}$ from the shallow to the deep scale, where d is the shallow feature dimension. At each scale s, ALM maps the paired features (F_s^w, F_s^p) to residual corrections (Δ_s^w, Δ_s^p). The bridge then reduces the power residual Δ_s^p to two-dimensional tokens through softmax-weighted pooling. The weather residual Δ_s^w and the restored power residual Δ_s^p are subsequently injected into the scale-matched decoder stages of their respective Backbones. Finally, the top-level Decoders reconstruct the latent representations into weather and power forecasts.

3.3 ALM

ALM aims to address two crucial challenges: introducing power variables without destabilizing pretrained weather representations, and enabling useful information exchange in both directions with a small trainable footprint. As shown in Figure 2(b), an ALM block first aligns the paired weather and power features and writes them into a shared latent memory through branch-specific write operations. Each branch then retrieves information relevant to its forecast from the fused memory. Finally, branch-specific postfusion modulation converts the retrieved information into residual updates for the corresponding decoder. Alongside the ALM parameters, we apply BitFit (Ben Zaken et al., 2022) to Aurora, which updates only its bias parameters while keeping all other pretrained Aurora parameters frozen. We further apply progressive rollout fine-tuning at 24- and 72-hour horizons, as detailed in Appendix B.2.

3.3.1 Prefusion Alignment

Although the weather and power inputs are spatiotemporally aligned, their features can still differ substantially in magnitude and distribution. For every $s \in \mathcal{S}$, prefusion alignment maps each feature into a compact alignment space of width $d_{\text{ca},s} = \max(1, \lfloor r_{\text{pf}} d_s \rfloor)$:

$$Z_s^b = W_s^{\text{cmp},b}\text{LN}_b(F_s^b). \tag{6}$$

where $r_{\text{pf}} \in (0, 1]$ controls the relative width of the compact alignment space, and LN_b is branch-specific layer normalization. The two branches use branch-specific LayerNorm and compression/restoration mappings around a shared alignment network. For its layer $\ell \in \{1, \ldots, N\}$, the

transformation and residual recursion are

$$G_s^{(\ell)}(H) = W_{s,2}^{(\ell)}\mathrm{Drop}\Big(\mathrm{GELU}\Big(W_{s,1}^{(\ell)}\mathrm{LN}(H)\Big)\Big), \qquad H_s^{b,(\ell)} = H_s^{b,(\ell-1)} + G_s^{(\ell)}\Big(H_s^{b,(\ell-1)}\Big). \tag{7}$$

where $H_s^{b,(0)} = Z_s^b$, and the output is $\bar{Z}_s^b = H_s^{b,(N)}$. Restoration mappings then return the aligned representations to their original channel spaces:

$$\tilde{F}_s^b = F_s^b + \alpha_s^b W_s^{\mathrm{rst},b}\bar{Z}_s^b, \tag{8}$$

where α_s^b controls the magnitude of the alignment residual. The restoration mappings are zero-initialized, so prefusion alignment begins as an identity mapping and does not perturb either pre-trained feature stream at start.

3.3.2 Latent Memory Interaction

ALM associates each $s \in \mathcal{S}$ with an independent learnable memory bank $M_s \in \mathbb{R}^{K_s \times d_s}$ that serves as a bidirectional write-read interface, where K_s is the number of memory tokens. This shared compact interface is related in function to attention bottlenecks for multimodal fusion (Nagrani et al., 2021), but here the memory is updated by both branches and then read separately to preserve branch-specific predictions. The aligned branch features and memory are normalized separately before attention:

$$Q_s^b = \mathrm{LN}_b(\tilde{F}_s^b), \qquad Q_s^m = \mathrm{LN}_m(M_s). \tag{9}$$

We use the following generic multi-head attention operator:

$$\mathrm{Attn}(Q,K,V) = \mathrm{Concat}_{h=1}^{H}\left[\mathrm{softmax}\left(\frac{(QW_h^Q)(KW_h^K)^\top}{\sqrt{d_h}}\right)(VW_h^V)\right]W^O. \tag{10}$$

During memory writing, the memory tokens serve as queries and each aligned branch independently provides keys and values:

$$A_s^b = \mathrm{Attn}(Q = Q_s^m, K = Q_s^b, V = Q_s^b). \tag{11}$$

The two branch-specific writeouts update the shared memory with equal prior weight:

$$\hat{M}_s = M_s + \frac{1}{2}A_s^w + \frac{1}{2}A_s^p. \tag{12}$$

where $\hat{M}_s$ is the fused memory state. Each branch then reads the fused memory:

$$\bar{M}_s = \mathrm{LN}_{\mathrm{out}}(\hat{M}_s), \qquad R_s^b = \mathrm{Attn}(Q = Q_s^b, K = \bar{M}_s, V = \bar{M}_s). \tag{13}$$

Thus, the two readouts selectively transfer relevant information across branches, from power to weather through R_s^w and from weather to power through R_s^p. Independent memory banks at the shallow, middle, and deep scales preserve scale-specific interactions.

3.3.3 Postfusion Modulation

Readout relevance varies across samples, branches, and scales, so a fixed residual transformation is insufficient. ALM therefore conditions each branch's modulation on its local state and the fused memory. For either branch, token-wise mean pooling gives

$$c_s^b = \mathrm{Concat}\big[\mathrm{pool}(Q_s^b), \mathrm{pool}(R_s^b), \mathrm{pool}(\bar{M}_s)\big]. \tag{14}$$

where c_s^b is the condition vector and pool averages over the token dimension. The branch-specific condition network predicts channel-wise scale and shift vectors:

$$(\gamma_s^b, \beta_s^b) = \mathrm{MLP}_{\mathrm{cond}}^b\big(\mathrm{LN}(c_s^b)\big). \tag{15}$$

The raw readout is normalized and modulated by a branch-conditional feature-wise affine transformation (Perez et al., 2018):

$$\tilde{R}_s^b = (1+\gamma_s^b)\odot \mathrm{LN}(R_s^b) + \beta_s^b. \tag{16}$$

where $\tilde{R}_s^b$ is the modulated readout and $\odot$ denotes element-wise multiplication. The modulated readout is further refined through a residual FFN and projected to obtain the target residual:

$$R_s^{b,\mathrm{out}} = R_s^b + \lambda_s^{\mathrm{pf},b}\mathrm{FFN}_s^b(\tilde{R}_s^b), \qquad \Delta_s^b = \lambda_s^b W_{\mathrm{zero}}^b(R_s^{b,\mathrm{out}}). \tag{17}$$

where $\lambda_s^{\mathrm{pf},b}$ and λ_s^b scale the modulation and decoder residuals. Zero-initialized output projections make ALM a no-op at initialization and introduce cross-domain corrections gradually.

3.4 Optimization Objective

ALM is optimized using a supervised joint-prediction loss and a weather-output anchoring loss.

$$\mathcal{L}_{\text{sup}} = \frac{\gamma}{V_{\text{surf}} + V_a} \left[\alpha \sum_{v \in \mathcal{V}_{\text{surf}}} w_v^s \, \text{MAE}(\hat{y}_v, y_v) + \beta \sum_{v \in \mathcal{V}_a} w_v^a \, \text{MAE}(\hat{y}_v, y_v) \right]. \tag{18}$$

where $\hat{y}_v$ and y_v denote the prediction and target of variable v, respectively, and w_v is its variable-specific weight. The coefficients α and β balance the surface and atmospheric forecasting losses. The power variables are included in $\mathcal{V}_{\text{surf}}$ and optimized jointly with the surface weather variables. To stabilize the weather branch, the joint weather predictions are anchored to a stop-gradient forward pass of the same weather branch.

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{anc}} \mathcal{L}_{\text{anc}} = \mathcal{L}_{\text{sup}} + \frac{\lambda_{\text{anc}}}{N_{\text{anc}}} \sum_{i=1}^{N_{\text{anc}}} \text{MSE}(\hat{y}_i^w, y_i^{\text{teacher}}). \tag{19}$$

where y_i^{teacher} is the corresponding no-gradient output of the current weather branch after power variables are removed, N_{anc} is the number of anchored outputs, and λ_{anc} controls the anchoring strength. The teacher is recomputed from the current weather branch at each rollout step and is not a separately frozen checkpoint.

4 Experiments

Datasets. We construct a spatiotemporally aligned weather-power dataset from ERA5 (Hersbach et al., 2020), gridded wind and solar capacity-factor fields over the contiguous United States (CONUS), and hourly county-level electricity demand profiles (Obika et al., 2025). The meteorological fields comprise five pressure-level variables (z, u, v, t, q) at 13 levels from 1000 to 50 hPa and four surface variables $(2t, 10u, 10v, \text{msl})$, while the power fields comprise wind power, photovoltaic power, and load. All variables cover a $0.25^\circ \times 0.25^\circ$ CONUS grid of size 105×236 over 2016–2023. The source records are hourly, while the model advances at 6-hour intervals and each forecast is initialized from two historical states. The temporal split assigns 2016–2021 to training, 2022 to validation, and the independent year 2023 to testing. Weather fields are normalized using the statistics adopted during Aurora pretraining, whereas wind power, photovoltaic power, and load use statistics estimated from the 2016–2021 training split. Further dataset and preprocessing details are provided in Appendix A.

Baselines. We organize the baselines according to their experimental roles. The basic methods are Aurora Zero-Shot (A-ZS) (Bodnar et al., 2025), Aurora weather-only FT (A-WFT), Aurora weather-power FT (A-WPFT), and PowerSwin Zero-Shot (P-ZS). The parameter-efficient fine-tuning (PEFT) methods are BitFit (Ben Zaken et al., 2022), LoRA (Hu et al., 2022), DoRA (Liu et al., 2024a), AdaLoRA (Zhang et al., 2023), PiSSA (Meng et al., 2024), SSF (Lian et al., 2022), AdaptFormer (Chen et al., 2022), VPT (Jia et al., 2022), and WeatherPEFT (Cao et al., 2026). The time-series forecasting methods are PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024b), TimeXer (Wang et al., 2024b), TimeMixer (Wang et al., 2024a), and SimpleTM (Chen et al., 2025). Time-series forecasters are evaluated only in E2 as power-only baselines for the downstream applications. Within each experiment, all methods use the prescribed data splits, target lead times, and evaluation criteria. Except for the two zero-shot baselines, all baselines undergo the same 24-hour and 72-hour rollout fine-tuning stages. Complete baseline configurations are provided in Appendix C.

4.1 E1: Main Experimental Results

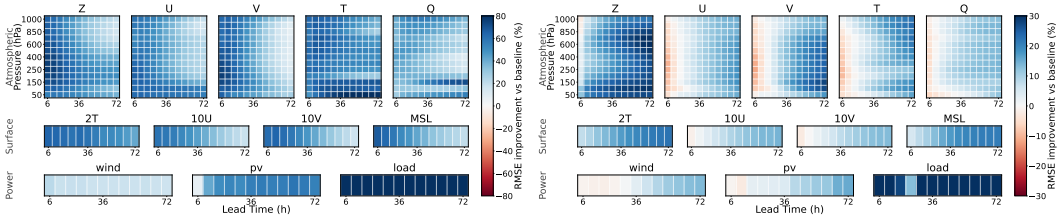
Tables 1 and 2 compare the different baselines for weather and power forecasting at the 72-hour lead time. We report latitude-weighted RMSE and anomaly correlation coefficient (ACC) for weather variables and RMSE and MAE for power variables, with complete definitions provided in Appendix B.2. Complete results at lead times of 6, 24, and 72 hours are provided in Tables 7 and 8 of Appendix D.1. At 72 hours, ALM achieves the lowest RMSE and highest ACC for every evaluated weather variable and leads on five of the six power metrics. Compared with A-WPFT, ALM reduces mean weather RMSE by 23.2% and wind and photovoltaic power RMSE by 14.6% and 13.6%, respectively, while maintaining a comparable load RMSE.

Table 1: Weather forecasting performance comparison. The best and second-best results at the 72-hour lead time are shown in bold and underlined, respectively.

Method	z		u		v		t		q		$2t$		$10u$		$10v$		msl	
	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
A-ZS	1018.49	0.227	10.47	0.212	9.37	0.194	6.57	0.190	0.00109	0.143	5.59	0.452	3.43	0.232	3.76	0.191	691.26	0.247
A-WFT	973.45	0.287	9.63	0.352	10.22	0.339	4.78	0.399	0.000957	0.262	5.38	0.659	3.74	0.302	3.85	0.337	933.32	0.171
A-WPFT	928.99	0.359	9.62	0.419	9.92	0.391	4.26	0.420	0.00107	0.291	5.34	0.656	3.38	0.415	3.76	0.409	636.80	0.434
BitFit	1137.80	0.259	12.35	0.260	13.94	0.231	4.72	0.340	0.00114	0.193	5.71	0.496	4.63	0.211	4.67	0.230	947.13	0.275
LoRA	812.39	0.458	8.96	0.454	10.75	0.386	3.80	0.498	0.000943	0.301	4.27	0.710	3.36	0.422	3.62	0.425	662.31	0.425
DoRA	878.76	0.395	9.00	0.451	10.57	0.388	3.80	0.492	0.000915	0.320	4.25	0.709	3.29	0.443	3.60	0.442	699.41	0.427
AdaLoRA	767.41	0.467	8.82	0.459	9.78	0.389	3.65	0.504	0.000955	0.314	4.42	0.682	3.29	0.431	3.59	0.426	639.45	0.446
PiSSA	850.98	0.428	8.93	0.457	10.15	0.365	3.74	0.495	0.000968	0.302	4.24	0.714	3.48	0.403	3.71	0.405	703.84	0.416
SSF	768.54	0.444	8.96	0.414	8.97	0.374	3.93	0.448	0.00102	0.261	5.78	0.436	3.42	0.351	3.68	0.348	671.02	0.387
AdaptFormer	793.31	0.447	8.61	0.458	9.96	0.385	3.63	0.510	0.000940	0.310	4.56	0.698	3.19	0.438	3.57	0.418	677.79	0.397
VPT	875.33	0.317	9.70	0.347	9.36	0.306	4.44	0.332	0.00113	0.201	5.85	0.498	3.74	0.272	3.98	0.249	733.98	0.318
WeatherPEFT	802.03	0.433	8.83	0.447	9.60	0.372	3.71	0.488	0.000975	0.287	4.65	0.644	3.25	0.418	3.54	0.411	658.10	0.412
ALM	**638.53**	**0.587**	**7.49**	**0.560**	**8.14**	**0.506**	**3.08**	**0.603**	**0.000796**	**0.405**	**3.31**	**0.794**	**2.82**	**0.529**	**3.13**	**0.524**	**554.41**	**0.551**

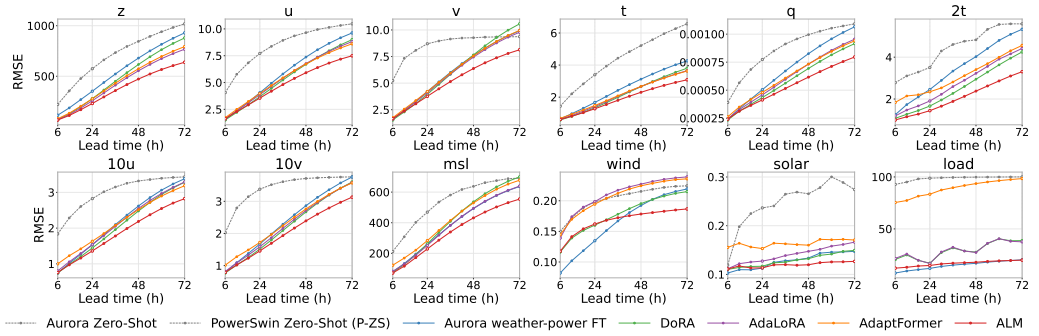
Table 2: Power forecasting performance comparison.

Method	wind		solar		load	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
P-ZS	0.224	0.156	0.274	0.179	100.06	21.96
A-WPFT	0.219	0.117	0.147	0.0837	20.25	**5.94**
BitFit	0.232	0.125	0.275	0.174	39.96	16.31
LoRA	0.236	0.124	0.144	0.0862	37.02	14.31
DoRA	0.215	0.114	0.149	0.0897	38.72	14.78
AdaLoRA	0.239	0.136	0.166	0.0973	37.56	14.29
PiSSA	0.226	0.117	0.137	0.0799	46.73	14.14
SSF	0.224	0.126	0.310	0.215	39.36	16.62
AdaptFormer	0.236	0.128	0.171	0.111	98.41	24.91
VPT	0.240	0.137	0.297	0.209	60.10	21.55
WeatherPEFT	0.237	0.127	0.195	0.118	38.20	14.99
ALM	**0.187**	**0.0994**	**0.127**	**0.0655**	**19.97**	6.02



(a) Relative to zero-shot baselines.

(b) Relative to DoRA.



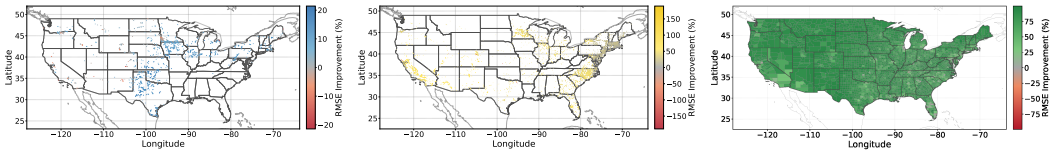
(c) RMSE by lead time.

Figure 3: Relative RMSE improvement scorecards and lead-time curves on the 2023 test set. Figure (a) uses A-ZS for weather and P-ZS for power, while Figure (b) uses DoRA for all variables.

Figure 3(a) and (b) compare ALM with the corresponding zero-shot baselines and second-ranked PEFT method DoRA, respectively. Relative to the zero-shot baselines, ALM delivers substantial RMSE improvements for every weather and power variable at every lead time. Against DoRA, the gains are mixed at short lead times but become broadly positive and generally strengthen as the forecast horizon increases. Figure 3(c) shows that ALM achieves the lowest RMSE for all weather

Table 3: Downstream power forecasting performance at a 72-hour lead time (MW).

Method	Plant solar forecasting		Plant wind forecasting		County load forecasting	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
P-ZS	3.69	2.14	30.92	22.49	455.07	139.82
A-WPFT	2.22	1.37	32.40	24.40	**105.47**	**32.70**
DoRA	2.35	1.36	32.00	23.28	253.07	76.91
PiSSA	2.00	1.18	36.03	26.06	311.61	83.32
iTransformer	1.93	1.05	31.26	23.24	255.44	70.55
PatchTST	1.84	0.97	25.92	22.41	291.97	101.32
SimpleTM	**1.65**	1.03	24.93	21.29	187.61	51.55
TimeMixer	1.98	0.99	27.75	22.74	254.45	86.49
TimeXer	1.80	1.00	26.77	21.67	197.18	50.38
ALM	1.68	**0.92**	**24.64**	**18.98**	151.34	46.58



(a) Wind (b) Solar (c) Load

Figure 4: Spatially resolved average RMSE improvement of ALM over iTransformer for downstream wind, solar, and load forecasting.

variables at almost every lead time, while its error curves rise more slowly as lead time increases. These results show that ALM limits error growth more effectively and delivers increasingly consistent gains at longer forecast horizons.

4.2 E2: Downstream Power Applications

Gridded power variable fields cannot adequately assess real-world forecasting performance, so we convert them into plant-level photovoltaic generation, plant-level wind generation, and county-level load forecasts for downstream evaluation. The gridded methods produce 12-step autoregressive forecasts at 6-hour intervals from E1 checkpoints and map them to plants and counties, whereas five independently trained time-series baselines directly forecast the corresponding object-level sequences. The complete data construction, training, mapping, and evaluation protocol is provided in Appendix B.3.

As summarized in Table 3, ALM achieves the lowest 72-hour plant-level wind RMSE/MAE and photovoltaic MAE, while ranking second for photovoltaic RMSE and both county-level load metrics. Relative to A-WPFT, ALM reduces photovoltaic RMSE/MAE by 24.3%/32.8% and wind RMSE/MAE by 24.0%/22.2%, respectively. Despite using only 9.7% of A-WPFT's trainable parameter count, ALM outperforms it on all four renewable-generation metrics, while A-WPFT remains strongest on both county-level load metrics.

Figure 4 compares ALM with iTransformer across photovoltaic plants, wind plants, and counties. ALM shows predominantly positive spatial gains for wind and solar, with only localized unfavorable sites. The load map is nearly uniformly positive across counties. These geographically widespread gains demonstrate that ALM's joint gridded modeling supports downstream forecasting more effectively than conventional site-level approaches.

4.3 E3: Ablation Studies

We ablate the three ALM components, the BitFit subset, the power-to-weather injection intensity, and the rollout fine-tuning stages using the same configuration as E1. To control the power-to-weather injection intensity, we scale the weather residual before decoder injection as $\Delta_s^{w,\mathrm{int}} = \eta \Delta_s^w$. Each structural *w/o* configuration and the $\eta = 0$ intervention receive only single-step fine-tuning. Complete results across all reported metrics and lead times are provided in Tables 10 and 11 of Appendix D.3.

Table 4: Ablation studies results at the 72-hour lead time. PA, LM, and PM denote prefusion alignment, latent memory, and postfusion modulation, respectively; η controls the power-to-weather injection intensity.

Method	z	u	v	t	q	$2t$	$10u$	$10v$	msl	wind	solar	load
A-ZS/P-ZS	1018.49	10.47	9.37	6.57	0.00109	5.59	3.43	3.76	691.26	0.224	0.274	100.06
w/o PA	785.94	9.21	9.89	3.79	0.000971	4.12	3.43	3.77	684.85	0.216	0.176	25.61
w/o LM	770.15	9.26	9.60	3.73	0.000977	4.04	3.41	3.72	692.23	0.217	0.167	30.90
w/o PM	762.66	9.09	9.79	3.76	0.000980	4.12	3.43	3.80	687.87	0.223	0.172	31.43
w/o BitFit	756.61	8.96	9.69	3.71	0.000966	4.08	3.34	3.68	690.89	0.218	0.169	24.69
$\eta = 0$	807.26	8.89	10.30	3.68	0.000947	4.17	3.23	3.64	669.91	0.216	0.172	24.74
ALM-6h	709.50	8.43	9.27	3.49	0.000879	3.87	3.07	3.47	631.04	0.211	0.166	21.45
ALM-24h	673.80	8.10	8.80	3.33	0.000859	3.66	3.03	3.42	598.88	0.196	0.131	**19.34**
ALM-72h	**638.53**	**7.49**	**8.14**	**3.08**	**0.000796**	**3.31**	**2.82**	**3.13**	**554.41**	**0.187**	**0.127**	19.97

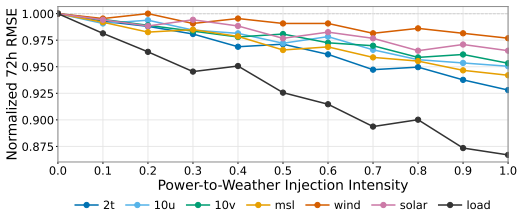


Figure 5: Weather and power forecasting response to the power-to-weather injection intensity η. RMSE is normalized by its value at $\eta = 0$.

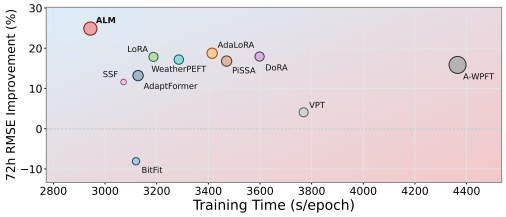


Figure 6: Mean 72-hour forecasting improvement versus per-epoch training cost. Marker area represents trainable parameter count on a logarithmic scale.

Table 4 shows that removing any structural component or BitFit updates, or setting $\eta = 0$, raises the 72-hour RMSE of every weather and power variable relative to ALM-6h. Rollout supervision further improves long-horizon performance: ALM-24h improves all 12 outputs over ALM-6h, and ALM-72h is best on 11 of 12. The results therefore support both the necessity of ALM's components and the value of progressively longer rollout fine-tuning.

Figure 5 varies η from zero to one and shows an overall decline in normalized 72-hour RMSE as the power-to-weather injection intensity increases. All seven curves finish below one at $\eta = 1$, indicating that full residual injection provides a useful complementary signal for weather prediction.

4.4 E4: Computational Efficiency

We compare ALM with A-WPFT and nine generic PEFT methods using mean 72-hour RMSE improvement, per-epoch training time, and trainable parameter count.

Figure 6 shows ALM in the favorable upper-left region, combining substantial 72-hour forecasting improvements with the shortest training time. Although ALM uses more trainable parameters than generic PEFT methods, it trains faster and uses far fewer parameters than full fine-tuning.

5 Conclusion

We proposed ALM, a lightweight framework for adapting weather foundation models to weather-power joint forecasting. Compared with full fine-tuning, ALM reduces mean 72-hour RMSE by 19.9% across 12 weather and power variables while using only 9.7% trainable parameters. The downstream applications, component ablations, and computational analysis further support ALM's downstream applicability, the positive marginal contributions of its components, and its training efficiency. Future work will extend weather-power joint forecasting to global energy networks and other domains such as industrial and commercial systems to develop a general-purpose forecasting framework for complex real-world systems.

AI USE STATEMENT

Large language models were used to assist with literature search tasks, code implementation, and manuscript language polishing. The authors independently reviewed and verified the cited sources, scientific claims, technical details, experimental protocols, and reported results.

REPRODUCIBILITY STATEMENT

The model architecture and optimization objectives are described in Section 3. Dataset construction, training protocols, hyperparameters, evaluation metrics, and baseline settings are provided in Appendices A–C. Code and configuration files are included in the Supplementary Material.

REFERENCES

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

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Shilei Cao, Hehai Lin, Jiashun Cheng, Yang Liu, Guowen Li, Xuehe Wang, Juepeng Zheng, Haoyuan Liang, Meng Jin, Chengwei Qin, Hong Cheng, and Haohuan Fu. Task-adaptive parameter-efficient fine-tuning for weather foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oANkBaVci5.

Shengchao Chen, Guodong Long, Jing Jiang, and Chengqi Zhang. Personalized adapter for large meteorology model on devices: Towards weather foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 84897–84943. Curran Associates, Inc., 2024.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678. Curran Associates, Inc., 2022.

Xinfang Chen, Weiran Chen, Venkata Dinavahi, Yiqing Liu, and Jilin Feng. Short-term load forecasting and associated weather variables prediction using ResNet-LSTM based deep learning. *IEEE Access*, 11:5393–5405, 2023.

Yan Chen, Rongbin Liu, Tingxiao Ding, Jin Su, Yuhan Yang, Rongxing Jiang, Yingying Jiang, Yiming Qin, and Kui Huang. Modeling temporal lag correlations for wind farm meteorology-power joint forecasting: A universal framework. *International Journal of Electrical Power & Energy Systems*, 178:111947, 2026.

William Y. Y. Cheng, Yubao Liu, Alfred J. Bourgeois, Yonghui Wu, and Sue Ellen Haupt. Short-term wind forecast of a data assimilation/weather forecasting system with wind turbine anemometer measurement assimilation. *Renewable Energy*, 107:340–351, 2017.

Yinghao Chu, Yiling Wang, Dazhi Yang, Shanlin Chen, and Mengying Li. A review of distributed solar forecasting with remote sensing and deep learning. *Renewable and Sustainable Energy Reviews*, 198:114391, 2024.

Aron P. Dobos. Pvwatts version 5 manual. Technical Report NREL/TP-6A20-62641, National Renewable Energy Laboratory, 2014.

J. E. Dobson, E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley. Landscan: A global population database for estimating populations at risk. *Photogrammetric Engineering & Remote Sensing*, 66(7):849–857, 2000.

Yuwei Fan, Tao Song, Chenlong Feng, Chao Liu, and Dongxiang Jiang. Wind power prediction using foundation large time series models enhanced by time series prompt in exogenous and tuning forms. *Applied Energy*, 400:126535, 2025.

Fang He, Jiaqi Fan, Yang Deng, Xiaoyang Zhang, Ka Tai Lau, and Dan Wang. Weather foundation model enhanced decentralized photovoltaic power forecasting through spatio-temporal knowledge distillation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9665–9673. International Joint Conferences on Artificial Intelligence Organization, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, Andras Horanyi, Joaquin Munoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elias Holm, Marta Janiskova, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noel Thepaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Tsvetelina Ivanova, Sara Porchetta, Sophia Buckingham, Gertjan Glabeke, Jeroen van Beeck, and Wim Munters. Improving wind and power predictions via four-dimensional data assimilation in the WRF model: case study of storms in february 2022 at belgian offshore wind farms. *Wind Energy Science*, 10:245–268, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, pp. 709–727. Springer Nature Switzerland, 2022.

Julie Keisler and Etienne Le Naour. WindDragon: Automated deep learning for regional wind power forecasting. *Environmental Data Science*, 4:e19, 2025.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 109–123. Curran Associates, Inc., 2022.

Wenlong Liao, Shouxiang Wang, Dechang Yang, Zhe Yang, Jiannong Fang, Christian Rehtanz, and Fernando Porté-Agel. TimeGPT in load forecasting: A large time series model perspective. *Applied Energy*, 379:124973, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024a. URL https://proceedings.mlr.press/v235/liu24bn.html.

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 *The Twelfth International Conference on Learning Representations*, 2024b.

N. Martin and J. M. Ruiz. Calculation of the pv modules angular losses under field conditions by means of an analytical model. *Solar Energy Materials and Solar Cells*, 70(1):25–38, 2001.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072. Curran Associates, Inc., 2024.

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.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14200–14213, 2021.

National Renewable Energy Laboratory. 2020 annual technology baseline, 2020. URL https://atb-archive.nrel.gov/electricity/2020/.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Kodi Obika, Wesley J. Cole, and Marie Rivers. Hourly electricity demand profiles for each county in the contiguous united states, 2025. URL https://data.openei.org/submissions/8562.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):3942–3951, 2018.

R. Perez, P. Ineichen, R. Seals, J. Michalsky, and R. Stewart. Modeling daylight availability and irradiance components from direct and global irradiance. *Solar Energy*, 44(5):271–289, 1990.

U.S. Energy Information Administration. Form eia-860 detailed data with previous form data, n.d.a. URL https://www.eia.gov/electricity/data/eia860/. Accessed August 13, 2026.

U.S. Energy Information Administration. Form EIA-923 detailed data with previous form data, n.d.b. URL https://www.eia.gov/electricity/data/eia923/. Accessed August 16, 2026.

U.S. Energy Information Administration. Hourly electric grid monitor: Form EIA-930, n.d.c. URL https://www.eia.gov/electricity/gridmonitor/. Accessed August 16, 2026.

U.S. Geological Survey. Annual nlcd collection 1 science products, 2024. URL https://www.sciencebase.gov/catalog/item/655ceb8ad34ee4b6e05cc51a.

Alessio Verdone, Massimo Panella, Enrico De Santis, and Antonello Rizzi. A review of solar and wind energy forecasting: From single-site to multi-site paradigm. *Applied Energy*, 392:126016, 2025.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498. Curran Associates, Inc., 2024b.

Junqiang Wei, Xuejie Wu, Tianming Yang, and Runhai Jiao. Ultra-short-term forecasting of wind power based on multi-task learning and LSTM. *International Journal of Electrical Power & Energy Systems*, 149:109073, 2023.

Jie Yan, Yujia Li, Han Wang, Shuang Han, Wenlong Shang, and Yongqian Liu. Wind power forecasting based on large time series model. *Engineering*, 2025. ISSN 2095-8099.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Xiaohui Zhong, Lei Chen, Hao Li, Xu Fan, Wenxu Qian, Jun Liu, Enwang Luo, and Libo Wu. Improving renewable energy forecasting with FuXi-Energy. *Applied Energy*, 412:127657, 2026.

APPENDIX

A DATASET AND PREPROCESSING

A.1 ERA5

ERA5 is the fifth-generation global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts and distributed through the Copernicus Climate Data Store (Hersbach et al., 2020). It combines observations with a numerical weather prediction system to provide hourly, multivariate, three-dimensional estimates of the atmospheric state. We use hourly records from 1 January 2016 through 31 December 2023 and crop them to the CONUS domain. The five pressure-level variables are geopotential, zonal wind, meridional wind, temperature, and specific humidity at $1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100,$ and 50 hPa. The four surface variables are 2-m temperature, 10-m zonal wind, 10-m meridional wind, and mean sea-level pressure. The land–sea mask, surface geopotential, and soil type are included as static fields. The resulting domain covers approximately $50.00°$N–$24.00°$N and $125.00°$W–$66.25°$W on a 105×236 grid at $0.25°$ resolution. ERA5 is available from `https://cds.climate.copernicus.eu/`.

A.2 GRIDDED WIND AND SOLAR CAPACITY-FACTOR FIELDS

The construction and validation of the wind and solar variables are supported by publicly available records provided by the U.S. Energy Information Administration (EIA). Monthly plant-level net generation is obtained from Form EIA-923 (U.S. Energy Information Administration, n.d.b), while generator locations, operating status, technology type, and nameplate capacities are obtained from Form EIA-860 (U.S. Energy Information Administration, n.d.a). Hourly wind and solar generation at the balancing-authority level is obtained from Form EIA-930 (U.S. Energy Information Administration, n.d.c). These records are harmonized over 2016–2023 and converted into hourly capacity-factor fields on the $0.25°$ CONUS grid. Training losses use capacity factors standardized with statistics computed from the 2016–2021 training split, whereas evaluation is performed after inverse transformation. The spatial and temporal downscaling procedure is detailed in Appendix A.4.

A.3 HOURLY ELECTRICITY DEMAND PROFILES FOR EACH COUNTY IN THE CONTIGUOUS UNITED STATES

County-level electricity demand is obtained from the Open Energy Data Initiative (OEDI) dataset *Hourly Electricity Demand Profiles for Each County in the Contiguous United States* (Obika et al., 2025). We align these records with the 2016–2023 ERA5 period and convert them into the gridded load field `load`. Load is standardized with the mean and variance computed exclusively from 2016–2021. The original county identifiers and polygons are retained for the downstream county-level reconstruction and evaluation. The dataset is available at `https://data.openei.org/submissions/8562`.

A.4 DATA ASSIMILATION

The three power variables are not spatially aligned with ERA5 before preprocessing. We harmonize them into hourly fields on the same $0.25°$ latitude–longitude grid used by the atmospheric variables. The resulting tensors follow the common layout $(\text{time}, \text{latitude}, \text{longitude})$ over the CONUS domain. All dynamic fields are aligned to an hourly UTC time axis. Here, *data assimilation* refers to the harmonization and physics-based construction of heterogeneous power data. It does not refer to variational or ensemble assimilation in numerical weather prediction.

Wind capacity factor. Hourly wind capacity factors are generated before model training using relevant historical environmental records on the common grid and a standard onshore wind-turbine model (Hersbach et al., 2020). For each target year, the wind conditions used to select the appropriate turbine power curve are calculated only from the preceding year. The selected curve follows the cut-in, rated, and cut-out settings reported in the National Renewable Energy Laboratory (NREL) Annual Technology Baseline (National Renewable Energy Laboratory, 2020). We apply mild spatial smoothing to reduce abrupt differences between neighboring grid cells. This gives the wind capacity

factor before fixed system losses $\mathrm{CF}_{g,t}^{\mathrm{curve}}$.

$$\mathrm{wind}_{g,t} = \mathrm{clip}\big(0.855\,\mathrm{CF}_{g,t}^{\mathrm{curve}}, 0, 1\big)\,, \tag{20}$$

Here, 0.855 represents the combined efficiency of the wind farm after the power-curve calculation, and the final capacity factor is limited to $[0, 1]$. For a forecast starting at t_0, the model only receives values at or before t_0. Values after t_0 are used only as prediction targets. The power curve for the target year is also selected using statistics from the preceding year. The resulting hourly wind field is aligned with the atmospheric grid without introducing future information.

Solar capacity factor. Hourly solar capacity factors are generated before model training from relevant historical environmental records and an EIA-860 equipment list from before the target year (U.S. Energy Information Administration, n.d.a). All source data are converted to the same grid and hourly time intervals. A timestamp t represents the interval $[t, t+1)\,\mathrm{h}$, and each step uses only data that were available by that time.

For each grid cell, the EIA-860 list from the preceding year is used to calculate the installed-capacity shares of fixed-tilt, single-axis, and dual-axis solar systems. If a plant's system type is missing, we fill it using the observed shares in the same state, then the same Census division, and finally the United States as a whole. Grid cells without recorded plants use the national shares. For each of the three system types, a standard photovoltaic conversion model accounts for panel orientation, the angle of incoming light, module output, DC power, and inverter conversion (Perez et al., 1990; Martin & Ruiz, 2001; Dobos, 2014). We use a DC-to-AC ratio of 1.2 and a total system loss of 14%. Single-axis systems use backtracking, a ground-coverage ratio of 0.35, and rotation limits of $\pm 60^\circ$. Dual-axis systems are assumed to face the sun directly. The three capacity factors are then combined according to their installed-capacity shares:

$$\mathrm{solar}_{g,t} = \sum_{k \in \{\mathrm{fixed,single,dual}\}} w_{g,k}\ \mathrm{clip}\left(\frac{P_{g,t,k}^{\mathrm{AC}}}{P_{\mathrm{AC},0}}, 0, 1\right). \tag{21}$$

For a forecast starting at t_0, the model only receives values at or before t_0. Values after t_0 are used only as prediction targets. The shares of the three solar system types are calculated from the preceding year's EIA-860 list. Snow losses and an additional empirical correction for cold conditions are not included. Thus, the resulting hourly AC capacity-factor field is aligned with the atmospheric grid without introducing future information.

Electricity load. The OEDI dataset *Hourly Electricity Demand Profiles for Each County in the Contiguous United States* provides hourly county-level demand $L(c, t)$ for 2016–2023 (Obika et al., 2025). We convert every ERA5 grid center into a $0.25^\circ \times 0.25^\circ$ cell polygon and intersect it with the harmonized county-equivalent boundaries. The released batch preprocessing pipeline applies this construction independently to each load year Y from 2016 to 2023, using only LandScan population (Dobson et al., 2000) and Annual National Land Cover Database (NLCD) products (U.S. Geological Survey, 2024) from year $Y - 1$. Population pixels are fractionally allocated according to their overlap with each county–grid intersection. The NLCD component retains non-road developed impervious surfaces with Impervious Descriptor 2 and developed land-cover classes $\{21, 22, 23, 24\}$. Population supplies the dominant allocation signal, while NLCD contributes only its positive excess share relative to population:

$$w(g, c) = 0.98\, w_{\mathrm{pop}}(g, c) + 0.02\, w_{\mathrm{NLCD\text{-}excess}}(g, c), \qquad \sum_g w(g, c) = 1. \tag{22}$$

Intersection area is used as a fallback when the county population is zero. Collecting the county-to-grid weights in a sparse matrix W, the gridded load at hour t is

$$G(t) = WL(t), \qquad G(g, t) = \sum_c w(g, c) L(c, t). \tag{23}$$

Column normalization makes the transformation exactly conservative for every county and hour. The resulting `load` field is stored as `float32`. Cells outside the allocation mask are set to missing. Quality control verifies hourly continuity, non-negativity, mask consistency, finite values within the allocation domain, and conservation of the original county totals.

Temporal alignment and normalization. After construction, wind, solar, and load are restricted to the common 2016–2023 period and matched to the ERA5 grid and hourly UTC timestamps. The power fields are stacked with the meteorological variables only after this alignment. Following the main experimental protocol, normalization statistics are computed exclusively from the training period (2016–2021) and then applied unchanged to validation and test data. Training losses are evaluated in normalized space, whereas reported power metrics are computed after inverse transformation in physical units. Predicted capacity factors are not truncated to $[0, 1]$ before RMSE or MAE is computed.

B Detailed Experimental Setup

B.1 Environment Configuration

All experiments were conducted on Ubuntu 22.04.4 LTS with Linux kernel 6.8.0-57-generic under the following configuration:

- **Hardware**: Two Intel Xeon Platinum 8468V processors providing 96 physical cores and 192 logical CPUs, 503 GiB RAM, and four NVIDIA A100 80GB PCIe GPUs with driver 595.71.05. Each distributed E1 run used two A100 GPUs on a single node.
- **Software Stack**: Python 3.12.3, PyTorch 2.6.0+cu124, torchvision 0.21.0+cu124, CUDA runtime 12.4, cuDNN 9.1.0, DeepSpeed 0.17.1 with ZeRO stage 0, and NCCL 2.21.5.
- **Key Dependencies**: NumPy 2.2.6, SciPy 1.14.1, pandas 2.2.3, xarray 2026.7.0, Zarr 3.2.1, netCDF4 1.7.2, timm 1.0.15, einops 0.8.1, and safetensors 0.5.3.

B.2 Detailed Main Experimental Setup

Data loading and reproducibility. The joint stages read preprocessed monthly shards stored as `float32`, use one data-loading worker per process, pin host memory, prefetch two batches per worker, and cache at most six days. Training and validation retain every second day in the 6-hour and 24-hour stages and every fourth day in the 72-hour stage, always with day-stride offset zero. Test windows are not subsampled by day. This produces 26,304/4,380 training/validation samples at 6 hours, 26,298/4,368 at 24 hours, and 13,146/2,184 at 72 hours. The joint-stage sampler does not shuffle samples. All stages use random seed 42, deterministic cuDNN kernels, and disabled cuDNN benchmarking.

Evaluation metrics. All metrics are computed after transforming the predictions and targets back to their physical scales. Let $\hat{y}^{q,\tau}_{n,v,i,j}$ and $y^{q,\tau}_{n,v,i,j}$ denote the prediction and target for variable v in output group $q \in \{s, a, p\}$, test sample n, lead time τ, and grid cell (i, j). We define the normalized latitude weight as $a_{i,j} = \cos(\varphi_i)/\sum_{i',j'} \cos(\varphi_{i'})$, where φ_i is the latitude at row i and $\sum_{i,j} a_{i,j} = 1$. The latitude-weighted root mean squared error is

$$\mathrm{RMSE}^{q}_{v,\tau} = \left[\frac{1}{N} \sum_{n=1}^{N} \sum_{i,j} a_{i,j} \left(\hat{y}^{q,\tau}_{n,v,i,j} - y^{q,\tau}_{n,v,i,j} \right)^2 \right]^{1/2}, \tag{24}$$

where N is the number of test samples evaluated at lead time τ. For ACC, we obtain the prediction and target anomalies $\tilde{\hat{y}}^{q,\tau}_{n,v,i,j}$ and $\tilde{y}^{q,\tau}_{n,v,i,j}$ by subtracting the day-of-year and hour-of-day climatology $c^{q}_{d,h,v,i,j}$ from $\hat{y}^{q,\tau}_{n,v,i,j}$ and $y^{q,\tau}_{n,v,i,j}$, respectively. The anomaly correlation coefficient averages the latitude-weighted spatial anomaly correlation over the N test samples:

$$\mathrm{ACC}^{q}_{v,\tau} = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{i,j} a_{i,j} \tilde{\hat{y}}^{q,\tau}_{n,v,i,j} \tilde{y}^{q,\tau}_{n,v,i,j}}{\left[\sum_{i,j} a_{i,j} \left(\tilde{\hat{y}}^{q,\tau}_{n,v,i,j} \right)^2 \sum_{i,j} a_{i,j} \left(\tilde{y}^{q,\tau}_{n,v,i,j} \right)^2 \right]^{1/2}}. \tag{25}$$

Lower RMSE and higher ACC indicate better weather forecasts. For wind power, photovoltaic power, and load, we report RMSE and MAE, with lower values indicating better performance. For pressure-level weather variables, the reported value first aggregates the corresponding pressure-level scores for each variable.

Training strategy. Table 5 summarizes the four training stages and their key settings. PowerSwin is first pretrained on gridded power variables and subsequently integrated with pretrained Aurora for single-step 6-hour joint fine-tuning. Each successive stage is initialized from the best validation checkpoint of the preceding stage. During 24-hour and 72-hour rollout fine-tuning, each training iteration performs K autoregressive forecast steps. Starting from the observed state at t, each 6-hour prediction becomes the next input:

$$\hat{\mathcal{X}}_t^{(0)} = \mathcal{X}_t, \qquad \hat{\mathcal{Y}}_{t+k\delta}^{(k)} = f_{\Theta_{\mathrm{t}},\phi}\Big(\hat{\mathcal{X}}_{t+(k-1)\delta}^{(k-1)}; \Theta_0 \setminus \Theta_{\mathrm{t}}\Big), \qquad k = 1, \ldots, K, \tag{26}$$

where $\hat{\mathcal{Y}}_{t+k\delta}^{(k)}$ is the prediction at step k, and $\hat{\mathcal{X}}^{(k)}$ is formed by advancing the input history with this prediction. Here $K \in \{4, 12\}$ for the 24-hour and 72-hour stages. The rollout fine-tuning objective accumulates these step-wise losses:

$$\mathcal{L}_{\text{roll-ft}}^{(K)} = \sum_{k=1}^{K} \omega_k \, \mathcal{L}\Big(\hat{\mathcal{Y}}_{t+k\delta}^{(k)}, \mathcal{Y}_{t+k\delta}\Big), \tag{27}$$

where $\mathcal{L}$ is defined in (19) and ω_k is the supervision weight assigned to rollout step k. During joint fine-tuning, ALM, the pseudo-level bridges, and the Aurora BitFit subset are trainable, while the remaining backbone parameters stay frozen.

Checkpoint selection. Table 5 summarizes the key stage-specific and shared settings used in E1. For checkpoint selection, normalized RMSE is averaged separately over the weather and power variables and combined as $(S_{\text{weather}} + 2S_{\text{power}})/3$. For rollout stages, this score is additionally averaged uniformly across the K forecast steps. The checkpoint with the lowest combined validation score is retained, and training stops after five validation checks without an improvement of at least 0.005. The 2023 test set is evaluated only after checkpoint selection.

Pseudo-level bridge. At each interaction scale s, the bridge uses four learned pseudo-levels to map power tokens $P_s \in \mathbb{R}^{B\times H_s W_s \times d_s}$ to Aurora's level-aware representation. With a learned level embedding $E_s \in \mathbb{R}^{4\times d_s}$, the expansion is

$$\mathcal{E}_s(P_s)_{b,\ell,i,j,:} = P_{s,b,i,j,:} + E_{s,\ell,:}, \qquad \mathcal{E}_s(P_s) \in \mathbb{R}^{B\times 4H_s W_s \times d_s}. \tag{28}$$

For a pseudo-level tensor $Z_s \in \mathbb{R}^{B\times 4H_s W_s \times d_s}$, four learned scalar logits $a_{s,\ell}$ define $\pi_{s,\ell} = \mathrm{softmax}(a_s)_\ell$, and the collapse operation is

$$\mathcal{C}_s(Z_s)_{b,i,j,:} = \sum_{\ell=1}^{4} \pi_{s,\ell} Z_{s,b,\ell,i,j,:}, \qquad \mathcal{C}_s(Z_s) \in \mathbb{R}^{B\times H_s W_s \times d_s}. \tag{29}$$

These four pseudo-levels are internal bridge dimensions and are distinct from the 13 ERA5 pressure levels.

Weather anchor. At every training rollout step, the anchor reference is obtained by an additional weather-only forward pass on the current rollout input after removing the power variables and bypassing the cross-branch ALM residuals. This pass is evaluated under `torch.no_grad()`, and it uses the current weather branch. Consequently, the Aurora parameters selected by BitFit can change between optimizer steps, while stop-gradient only prevents the reference forward pass from contributing gradients. The anchor averages the variable-wise MSE between the joint prediction and reference output over their common surface-weather and atmospheric variables. Its value is computed at each rollout step, added to that step's supervised loss, and then included in the uniformly averaged rollout objective.

Table 5: Key hyperparameter settings for PowerSwin pretraining and E1 joint adaptation.

Parameter	Value	Description
	Stage 0: PowerSwin pretraining	
Forecast setup	$T_{\text{in}} = 2$, lead time 6 h, $K = 1$	One-step prediction
Optimization	AdamW, LR 3×10^{-4}, weight decay 3×10^{-6}, constant schedule	All parameters trainable
Training	50 epochs, batch size 128, fp16 AMP, one GPU	-
Checkpoint selection	Lowest one-step validation MSE	Initializes E1
PowerSwin architecture	Patch size 4, width 512, depths $(2, 2, 2)$, heads $(8, 16, 32)$, window 6×12, MLP ratio 4	-
	Stage 1: 6-hour joint adaptation	
Initialization	Pretrained Aurora and selected PowerSwin, newly initialized ALM and bridges	PowerSwin frozen
Forecast setup	$T_{\text{in}} = 2$, lead time 6 h, $K = 1$	-
Training	30 epochs, scheduler horizon 20, train/validation/test batch sizes $4/8/8$ per GPU	Two GPUs, bf16
	Stage 2: 24-hour rollout fine-tuning	
Initialization and rollout	Best 6-hour checkpoint, $K = 4$, $\omega_k = 1/K$	-
Training	10 epochs, scheduler horizon 10, train/validation/test batch sizes $4/2/8$ per GPU	Two GPUs, bf16
	Stage 3: 72-hour rollout fine-tuning	
Initialization and rollout	Best 24-hour checkpoint, $K = 12$, $\omega_k = 1/K$	-
Training	6 epochs, scheduler horizon 6, train/validation/test batch sizes $4/2/8$ per GPU	Two GPUs, bf16
	Shared E1 settings	
Trainable parameters	ALM and bridge modules, Aurora BitFit parameters	PowerSwin frozen
Optimization	AdamW, ALM/bridge LR 5×10^{-5}, BitFit LR 3×10^{-4}, weight decay 0	-
LR schedule	Warmup ratio 0.10, cosine decay, minimum LR ratio 0.05, gradient clipping 1.0	-
Anchor and early stopping	$\lambda_{\text{anc}} = 0.1$, patience 5, minimum delta 0.005	-
Supervised-loss coefficients	$\gamma = 1$, $\alpha = 0.25$, $\beta = 1$	Equation (18)
Validation weights	Weather: 1.0, power: 2.0	Checkpoint selection
Latent memory	Tokens $(64, 32, 16)$, heads $(8, 16, 32)$	Shallow to deep
Prefusion alignment	Compact ratio 0.125, depth 1, dropout 0, weather scales $(0.5, 0.5, 0.25)$, power scales $(1, 1, 0.5)$	-
Postfusion modulation	Compact/hidden ratios 0.25/0.25, dropout 0, zero-initialized affine scale and shift	-

The rollout stages reduce the per-GPU validation batch size from 8 to 2 because retaining multi-step rollout states is memory intensive. The test batch size remains 8 because evaluation uses no-gradient inference and does not retain the training computation graph.

B.3 Detailed Downstream Power Application Setup

Downstream tasks. E2 tests whether improvements in gridded weather-power forecasting transfer to plant-level photovoltaic power, plant-level wind power, and county-level load forecasting. The photovoltaic and wind objects are defined by Plant Code in the 2023 EIA-860 inventory (U.S. Energy Information Administration, n.d.a). We retain generators with operating status `OP` or `OA` and positive nameplate capacity, aggregate them within each Plant Code, and restrict the inventory to the CONUS grid. This process yields 5,632 photovoltaic plants and 1,309 wind plants. The load task contains 3,109 CONUS county-equivalent units identified by five-digit Federal Information Processing Standards (FIPS) codes (Obika et al., 2025). We use separate inverse mappings to

recover plant power and county load from the common gridded fields. Photovoltaic and wind forecasts are evaluated by aggregating predictions by plant and calendar month, comparing them with EIA-923 monthly net generation, and dividing the monthly values by the corresponding number of hours (U.S. Energy Information Administration, n.d.b). County-load forecasts are evaluated against the corresponding county records at the target timestamps. All methods use the 2023 test set and report 6-, 24-, and 72-hour RMSE and MAE in MW. Let $\mathcal{V}_\tau$ denote the valid plant–month pairs for photovoltaic and wind power or county–time pairs for load at lead time τ. For prediction $\widehat{y}_j^{(\tau)}$ and reference y_j, the pooled metrics are

$$\mathrm{RMSE}_\tau = \sqrt{\frac{1}{|\mathcal{V}_\tau|}\sum_{j\in\mathcal{V}_\tau}\left(\widehat{y}_j^{(\tau)} - y_j\right)^2}, \qquad \mathrm{MAE}_\tau = \frac{1}{|\mathcal{V}_\tau|}\sum_{j\in\mathcal{V}_\tau}\left|\widehat{y}_j^{(\tau)} - y_j\right|. \tag{30}$$

Each reported value therefore pools all valid evaluation pairs and does not average object-wise RMSE values.

Evaluation of gridded models. P-ZS uses the pretrained PowerSwin checkpoint, while the other gridded methods use their fixed E1 checkpoints. Each method performs a 12-step autoregressive rollout, with indices 0, 3, and 11 evaluated at lead times of 6, 24, and 72 hours, respectively. For photovoltaic and wind power, the inverse mapping extracts the predicted capacity factor at the forecast-grid cell assigned to each plant. The capacity-factor output is used directly without clipping and is converted to plant power using the total operating nameplate capacity from EIA-860:

$$\widehat{P}_{i,t}^{(\tau)} = \widehat{\mathrm{CF}}_{g(i),t}^{(\tau)} C_i, \qquad \tau \in \{6, 24, 72\}, \tag{31}$$

where $g(i)$ denotes the grid cell assigned to plant i, and C_i is the total nameplate capacity of the corresponding photovoltaic or wind generators within its Plant Code. Because EIA-923 reports monthly energy in MWh, the observed monthly mean power is

$$\overline{P}_{i,m}^{\mathrm{obs}} = \frac{E_{i,m}^{923}}{H_m}, \tag{32}$$

where $E_{i,m}^{923}$ is the monthly net generation for the corresponding plant and energy source, and H_m is the number of hours in month m. The predicted plant powers at all valid target timestamps in the same month are averaged as

$$\overline{\widehat{P}}_{i,m}^{(\tau)} = \frac{1}{|\mathcal{T}_{i,m}^{(\tau)}|}\sum_{t\in\mathcal{T}_{i,m}^{(\tau)}} \widehat{P}_{i,t}^{(\tau)}, \tag{33}$$

where $\mathcal{T}_{i,m}^{(\tau)}$ is the set of valid target timestamps for plant i, month m, and lead time τ. Predictions and references are matched by Plant Code, energy source, and calendar month. Pairs with missing generation, non-positive nameplate capacity, no valid predictions, or non-finite values are excluded. For load, let W be the fixed county-to-grid allocation matrix described in Appendix A.4. Its columns follow the canonical county FIPS list, and its rows follow the south-to-north and west-to-east forecast-grid order. The inverse mapping reconstructs county load from the flattened predicted field $\widehat{G}_{\mathrm{pred}}$ through

$$\widehat{L} = \left(W^\top W + \lambda I\right)^{-1} W^\top \widehat{G}_{\mathrm{pred}}, \qquad \lambda = 10^{-6}. \tag{34}$$

County predictions are matched with references by FIPS code and target timestamp. Missing and non-finite pairs are excluded. Before inversion, W is restricted to the physical rows represented by the forecast tensor, and longitudes are normalized to $[0, 360)$. Longitudes in $[-180^\circ, 180^\circ)$ are converted to $[0, 360^\circ)$ by adding 360° to negative values. The same conversion is applied before matching the forecast-grid rows with the rows of W. No method receives future observed weather states during forecasting.

Training and evaluation of time-series models. The plant extraction and county inverse mappings are first applied to the gridded power dataset to construct object-level sequences for the five time-series baselines. At 00, 06, 12, and 18 UTC, each plant receives the capacity factor at its assigned 0.25° grid cell, and county load is reconstructed with the same W-based inverse. The ERA5 variables $2t$, $10u$, $10v$, msl, total cloud cover, and surface solar radiation downwards are bilinearly interpolated to plant coordinates and county centroids (Hersbach et al., 2020). Each baseline trains

one global model for each of photovoltaic power, wind power, and load, with the task model shared across all corresponding objects. The models use 2016–2021 for training, 2022 for validation, and 2023 for testing. They take 56 six-hour historical steps and directly predict the next 12 object-level power values. They do not receive future observed weather states. Common optimization settings and method-specific architectures are listed in Table 6. During evaluation, plant capacity-factor outputs are converted to MW with the same EIA-860 capacities, aggregated by plant and calendar month, and compared with EIA-923 monthly net generation after both monthly values are divided by the corresponding number of hours. County outputs pass through the common county-to-grid-to-county operator before comparison with the county records. The same valid-pair rules and pooled RMSE and MAE definitions are used for both model families. The time-series baselines receive a longer history than the two states used by the gridded methods and directly predict in the object space. These two conditions give the time-series baselines an experimental advantage over the autoregressive gridded methods.

C Baseline Settings

Table 6 summarizes the method-specific configurations of the E1 and E2 baselines. Unless explicitly listed, E1 optimization, rollout, data, and checkpoint-selection settings follow the ALM protocol in Table 5, while the E2 time-series baselines follow the common downstream protocol in Appendix B.3.

Table 6: Baseline architectures and method-specific hyperparameters. Unlisted settings follow the corresponding ALM E1 or E2 protocol.

Category	Baseline	Year	Key configuration
Basic	A-ZS	2025	Frozen pretrained Aurora with no parameter updates.
	A-WFT	2025	Weather-only full fine-tuning, LR 3×10^{-5}.
	A-WPFT	2025	All joint-model parameters trainable, LR 3×10^{-5}.
	P-ZS	-	Frozen pretrained PowerSwin with no parameter updates.
PEFT	BitFit	2022	Bias parameters and new power-variable input/output parameters, LR 3×10^{-4}.
	LoRA	2022	Rank 8, $\alpha = 8$, dropout 0, one update shared across rollout steps.
	DoRA	2024	Rank 8, $\alpha = 16$, dropout 0.05, one update shared across rollout steps, LR 10^{-4}.
	AdaLoRA	2023	Initial/target rank 8, $\alpha = 16$, dropout 0.05, warmup ratios 0.10/0.20, mask interval 10, orthogonal penalty 0.1.
	PiSSA	2024	Rank 8 with principal-component initialization.
	SSF	2022	Feature-wise scale-and-shift modules at attention and MLP outputs.
	AdaptFormer	2022	Bottleneck width 64, residual scale 0.1.
	VPT	2022	Deep prompts, prompt length 20, hidden width 16, decoder heads trainable.
	WeatherPEFT	2026	Prompt length 20, hidden widths $(8, 6, 16, 16)$, selection ratio 0.001, $\gamma = 0.2$, BitFit subset trainable.
E1	Full/PEFT	-	DeepSpeed `WarmupLR`, warmup ratio 0.10, no cosine decay.
TS (E2 only)	Common protocol	-	Split 2016–2021/2022/2023, 56 six-hour inputs, 12 direct outputs, AdamW LR 3×10^{-4}, patience 5.
	PatchTST	2023	Patch length/stride $16/8$, width 256, 3 layers, 8 heads, channel-independent encoding.
	iTransformer	2024	Width 512, 3 layers, 8 heads, feed-forward width 2,048, object embedding 64.
	TimeXer	2024	Patch length 8, width 256, 2 layers, 8 heads, feed-forward width 512.
	TimeMixer	2024	Three scales, width 128, 2 layers, feed-forward width 256, moving-average kernel 25.
	SimpleTM	2025	Three scales, pseudo-length 56, kernel 3, width 128, 1 layer, 4 heads.

D Additional Experimental Results

D.1 Additional Main Experimental Results

Table 7 presents the complete weather RMSE and ACC results at lead times of 6, 24, and 72 hours. At 6 hours, the strongest PEFT baselines retain an advantage for several pressure-level variables,

whereas ALM performs best for $2t$ and msl and ties the highest ACC for z. At 24 and 72 hours, ALM achieves the best RMSE and ACC for every weather variable, with the clearest separation at 72 hours. The complete results show that ALM preserves the weather branch after adding power forecasting and delivers increasingly consistent weather gains at longer horizons, directly supporting the joint-forecasting objective.

Table 7: Complete weather forecasting results.

LT	Method	z		u		v		t		q		$2t$		$10u$		$10v$		msl	
		RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
6 h	A-ZS	220.98	0.915	4.06	0.827	5.23	0.735	1.43	0.880	0.000391	0.691	2.78	0.717	1.84	0.735	2.02	0.716	211.93	0.901
	A-WFT	95.56	0.882	1.63	0.967	1.58	0.959	0.682	0.954	**0.000231**	0.783	1.36	0.923	**0.736**	**0.956**	0.798	0.954	82.69	0.985
	A-WPFT	114.75	0.964	1.57	0.963	1.59	0.965	0.642	0.972	0.000240	0.783	1.25	0.932	0.743	**0.956**	0.788	0.959	78.40	0.985
	BitFit	86.89	0.981	1.81	0.958	1.79	0.956	0.671	0.968	0.000260	0.775	1.60	0.874	0.900	0.935	0.927	0.943	95.29	0.977
	LoRA	65.29	**0.990**	1.58	0.966	1.59	0.964	0.601	0.974	0.000240	0.782	1.14	0.935	0.788	0.950	0.801	0.957	79.08	0.985
	DoRA	70.69	0.987	**1.51**	**0.969**	**1.53**	**0.966**	**0.576**	**0.976**	0.000234	**0.784**	1.08	0.938	0.746	0.955	**0.768**	**0.960**	75.23	0.986
	AdaLoRA	69.93	0.988	1.61	0.965	1.63	0.963	0.616	0.973	0.000243	0.781	1.19	0.928	0.801	0.949	0.818	0.955	82.72	0.983
	PiSSA	**63.96**	**0.990**	1.52	**0.969**	1.54	**0.966**	0.580	**0.976**	0.000235	**0.784**	1.12	0.935	0.766	0.953	0.783	0.959	77.78	0.984
	SSF	98.13	0.981	2.01	0.947	2.07	0.943	0.767	0.960	0.000278	0.764	1.58	0.888	0.981	0.924	1.00	0.934	114.77	0.968
	AdaptFormer	79.85	0.985	1.72	0.960	1.74	0.957	0.680	0.968	0.000262	0.774	1.84	0.843	1.00	0.919	1.01	0.930	122.55	0.960
	VPT	153.19	0.945	2.64	0.911	2.75	0.904	1.05	0.930	0.000351	0.730	2.73	0.716	1.46	0.829	1.52	0.848	176.23	0.928
	WeatherPEFT	83.22	0.985	1.80	0.957	1.83	0.954	0.686	0.967	0.000262	0.772	1.35	0.910	0.893	0.936	0.913	0.945	98.52	0.976
	ALM	67.36	**0.990**	1.62	0.966	1.65	0.962	0.606	0.973	0.000242	0.780	**1.00**	**0.949**	0.757	0.954	0.789	0.958	**71.45**	**0.988**
24 h	A-ZS	574.42	0.604	7.71	0.488	8.70	0.272	3.39	0.543	0.000774	0.406	3.51	0.744	2.82	0.456	3.38	0.322	469.04	0.626
	A-WFT	352.38	0.742	4.35	0.839	4.25	0.752	1.87	0.813	0.000476	0.641	2.90	0.881	1.69	0.825	1.67	0.857	355.38	0.797
	A-WPFT	351.99	0.836	4.05	0.858	4.20	0.860	1.67	0.872	0.000504	0.657	2.45	0.909	1.54	0.860	1.67	0.863	257.79	0.891
	BitFit	314.20	0.873	4.98	0.799	5.17	0.800	1.65	0.870	0.000546	0.620	2.21	0.907	2.12	0.752	2.05	0.797	310.85	0.850
	LoRA	243.36	0.926	3.79	0.877	4.10	0.865	1.38	0.906	0.000453	0.667	1.76	0.940	1.52	0.861	1.56	0.874	246.71	0.903
	DoRA	277.90	0.896	3.68	0.885	3.95	0.874	1.34	0.911	0.000434	0.679	1.66	0.945	1.44	0.875	1.51	0.883	262.12	0.892
	AdaLoRA	257.47	0.918	3.87	0.872	4.10	0.864	1.41	0.903	0.000458	0.669	1.91	0.932	1.54	0.858	1.59	0.870	251.67	0.896
	PiSSA	247.89	0.922	3.73	0.880	3.94	0.873	1.34	0.911	0.000449	0.673	1.72	0.942	1.55	0.856	1.56	0.874	258.13	0.891
	SSF	331.69	0.872	4.71	0.809	4.95	0.801	1.81	0.850	0.000540	0.616	2.94	0.848	1.87	0.787	1.96	0.802	321.68	0.834
	AdaptFormer	276.76	0.905	3.98	0.861	4.21	0.853	1.48	0.894	0.000480	0.656	2.35	0.898	1.63	0.835	1.72	0.844	283.12	0.868
	VPT	437.03	0.758	5.75	0.717	6.03	0.701	2.33	0.749	0.000666	0.530	3.64	0.761	2.39	0.642	2.61	0.633	407.17	0.741
	WeatherPEFT	296.56	0.893	4.17	0.849	4.39	0.841	1.55	0.884	0.000491	0.646	2.16	0.914	1.66	0.832	1.73	0.845	287.00	0.866
	ALM	**230.61**	**0.933**	**3.55**	**0.891**	**3.76**	**0.882**	**1.28**	**0.917**	**0.000413**	**0.692**	**1.44**	**0.957**	**1.36**	**0.885**	**1.46**	**0.890**	**228.78**	**0.913**
72 h	A-ZS	1018.49	0.227	10.47	0.212	9.37	0.194	6.57	0.190	0.00109	0.143	5.59	0.452	3.43	0.232	3.76	0.191	691.26	0.247
	A-WFT	973.45	0.287	9.63	0.352	10.22	0.339	4.78	0.399	0.000957	0.262	5.38	0.659	3.74	0.302	3.85	0.337	933.32	0.171
	A-WPFT	928.99	0.359	9.62	0.419	9.92	0.391	4.26	0.420	0.00107	0.291	5.34	0.656	3.38	0.415	3.76	0.409	636.80	0.434
	BitFit	1137.80	0.259	12.35	0.260	13.94	0.231	4.72	0.340	0.00114	0.193	5.71	0.496	4.63	0.211	4.67	0.230	947.13	0.275
	LoRA	812.39	0.458	8.96	0.454	10.75	0.386	3.80	0.498	0.000943	0.301	4.27	0.710	3.36	0.422	3.62	0.425	662.31	0.425
	DoRA	878.76	0.395	9.00	0.451	10.57	0.388	3.80	0.492	0.000915	0.320	4.25	0.709	3.29	0.443	3.60	0.442	699.41	0.427
	AdaLoRA	767.41	0.467	8.82	0.459	9.78	0.389	3.65	0.504	0.000955	0.314	4.42	0.682	3.29	0.431	3.59	0.426	639.45	0.446
	PiSSA	850.98	0.428	8.93	0.457	10.15	0.365	3.74	0.495	0.000968	0.302	4.24	0.714	3.48	0.403	3.71	0.405	703.84	0.416
	SSF	768.54	0.444	8.96	0.414	8.97	0.374	3.93	0.448	0.00102	0.261	5.78	0.436	3.42	0.351	3.68	0.348	671.02	0.387
	AdaptFormer	793.31	0.447	8.61	0.458	9.96	0.385	3.63	0.510	0.000940	0.310	4.56	0.698	3.19	0.438	3.57	0.418	677.79	0.397
	VPT	875.33	0.317	9.70	0.347	9.36	0.306	4.44	0.332	0.00113	0.201	5.85	0.498	3.74	0.272	3.98	0.249	733.98	0.318
	WeatherPEFT	802.03	0.433	8.83	0.447	9.60	0.372	3.71	0.488	0.000975	0.287	4.65	0.644	3.25	0.418	3.54	0.411	658.10	0.412
	ALM	**638.53**	**0.587**	**7.49**	**0.560**	**8.14**	**0.506**	**3.08**	**0.603**	**0.000796**	**0.405**	**3.31**	**0.794**	**2.82**	**0.529**	**3.13**	**0.524**	**554.41**	**0.551**

Table 8 presents the complete RMSE and MAE results for gridded wind power, photovoltaic power, and load forecasting. A-WPFT is strongest on most 6- and 24-hour wind and load metrics, while ALM already achieves the best photovoltaic MAE at both horizons. At 72 hours, ALM ranks first on all three RMSE metrics and on wind and photovoltaic MAE, with only load MAE remaining marginally lower for A-WPFT. These results show that ALM converts the weather foundation model's continuous-process representations into stronger long-horizon forecasts for all three power variables.

Table 8: Complete power forecasting results.

LT	Method	wind		solar		load	
		RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
6 h	P-ZS	0.150	0.0888	0.119	0.0671	92.82	24.38
	A-WPFT	**0.0820**	**0.0412**	**0.103**	0.0601	**7.54**	**2.79**
	BitFit	0.138	0.0762	0.217	0.127	21.45	5.50
	LoRA	0.132	0.0703	0.118	0.0710	21.28	5.31
	DoRA	0.116	0.0609	0.110	0.0672	20.59	5.20
	AdaLoRA	0.139	0.0763	0.111	0.0642	21.59	5.44
	PiSSA	0.125	0.0657	0.112	0.0649	21.50	5.44
	SSF	0.142	0.0784	0.142	0.0919	22.00	5.99
	AdaptFormer	0.144	0.0841	0.156	0.0986	75.40	18.05
	VPT	0.146	0.0816	0.216	0.145	23.93	7.77
	WeatherPEFT	0.142	0.0775	0.117	0.0694	21.88	5.56
	ALM	0.118	0.0627	0.111	**0.0600**	12.22	4.63
24 h	P-ZS	0.198	0.124	0.237	0.159	98.81	20.33
	A-WPFT	**0.135**	**0.0672**	**0.113**	0.0651	**12.00**	**4.03**
	BitFit	0.195	0.107	0.193	0.115	16.40	8.47
	LoRA	0.199	0.104	0.124	0.0741	16.48	7.83
	DoRA	0.160	0.0821	0.117	0.0707	16.98	8.12
	AdaLoRA	0.199	0.112	0.127	0.0748	16.70	7.82
	PiSSA	0.184	0.0945	0.116	0.0674	17.59	7.90
	SSF	0.193	0.109	0.208	0.136	18.38	9.11
	AdaptFormer	0.195	0.108	0.153	0.0994	83.39	21.29
	VPT	0.203	0.115	0.234	0.158	29.14	12.48
	WeatherPEFT	0.197	0.107	0.140	0.0852	17.02	8.11
	ALM	0.162	0.0852	**0.113**	**0.0588**	14.80	4.80
72 h	P-ZS	0.224	0.156	0.274	0.179	100.06	21.96
	A-WPFT	0.219	0.117	0.147	0.0837	20.25	**5.94**
	BitFit	0.232	0.125	0.275	0.174	39.96	16.31
	LoRA	0.236	0.124	0.144	0.0862	37.02	14.31
	DoRA	0.215	0.114	0.149	0.0897	38.72	14.78
	AdaLoRA	0.239	0.136	0.166	0.0973	37.56	14.29
	PiSSA	0.226	0.117	0.137	0.0799	46.73	14.14
	SSF	0.224	0.126	0.310	0.215	39.36	16.62
	AdaptFormer	0.236	0.128	0.171	0.111	98.41	24.91
	VPT	0.240	0.137	0.297	0.209	60.10	21.55
	WeatherPEFT	0.237	0.127	0.195	0.118	38.20	14.99
	ALM	**0.187**	**0.0994**	**0.127**	**0.0655**	**19.97**	6.02

Figure 7 shows ALM's ACC improvement over each E1 weather baseline across variables and lead times. The gains are already broad against A-ZS and weaker adaptation methods at 6 hours, but remain more variable against the strongest PEFT baselines. By 72 hours, positive improvements cover all weather variables and all displayed baselines, with particularly large margins over A-ZS, VPT, and SSF. The heatmap confirms that ALM improves the weather branch across variables and that its weather-side gains are not limited to RMSE.

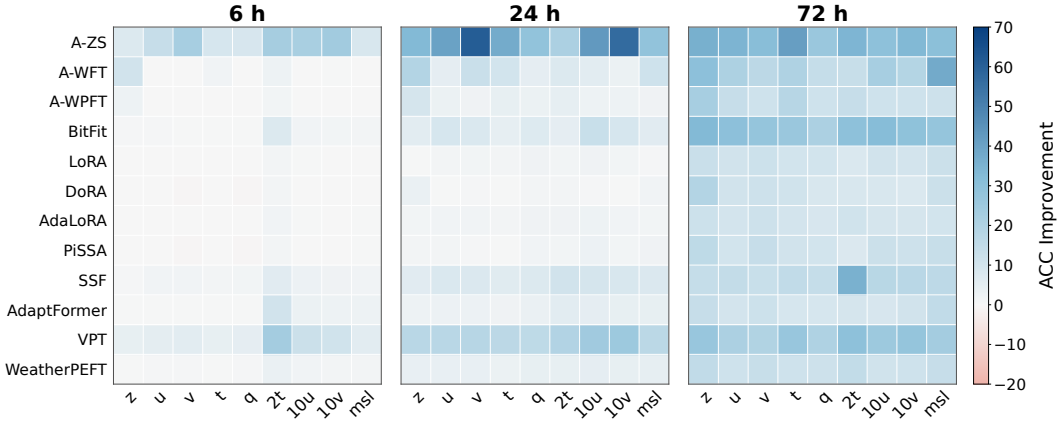


Figure 7: ACC improvement of ALM over E1 weather baselines at 6-, 24-, and 72-hour lead times.

Figure 8 shows weather ACC and power MAE throughout the 6–72-hour rollout. ALM remains on the upper envelope of the weather ACC curves as lead time increases and maintains lower power MAE than the generic PEFT baselines. For the power variables, its advantage is clearest for photovoltaic power and at longer wind horizons, whereas A-WPFT remains slightly better for load MAE. These complementary metrics support ALM's joint improvement of weather and power forecasting beyond the RMSE comparisons in the main text.

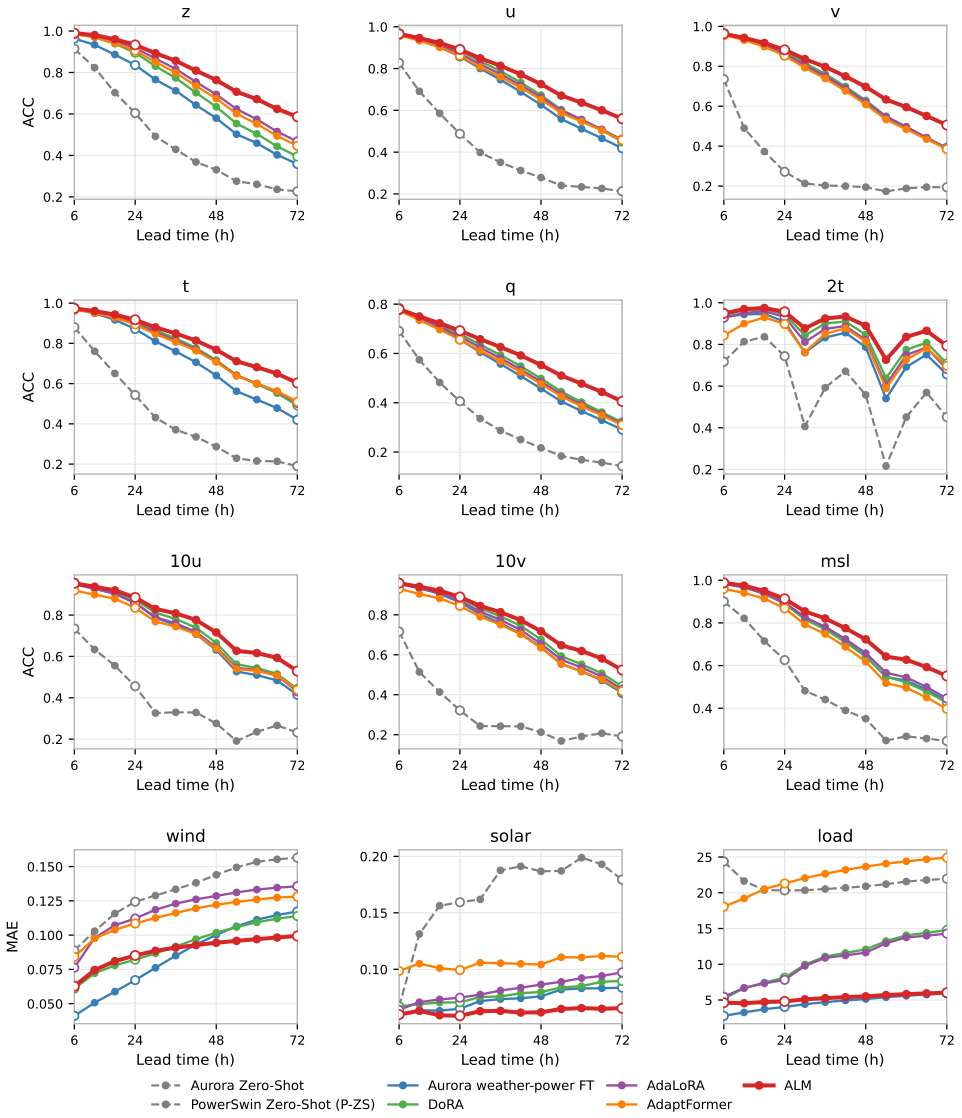
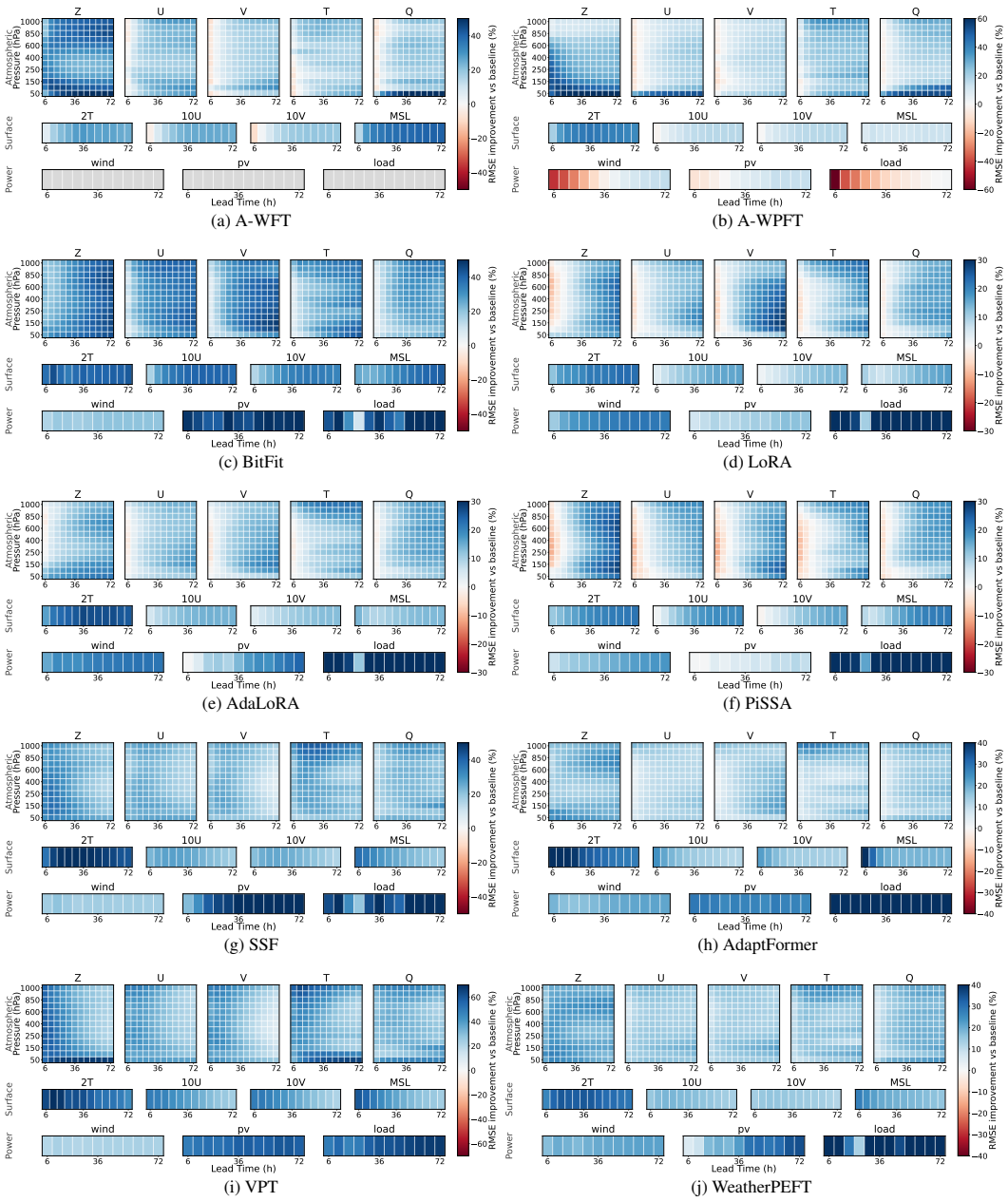


Figure 8: Weather ACC and power MAE lead-time curves omitted from the main-text RMSE figure.

Figure 9 shows ALM's variable-wise RMSE improvement over ten additional E1 baselines from 6 to 72 hours. Improvements over A-WFT and the generic PEFT methods generally expand with lead time, while the strongest baselines retain isolated advantages at shorter horizons. At 72 hours, the scorecards become predominantly positive across atmospheric, surface, and power variables, including against A-WPFT. This broader comparison shows that ALM's long-horizon gains persist across adaptation mechanisms and are not limited to a single reference method.



(a) A-WFT (b) A-WPFT

(c) BitFit (d) LoRA

(e) AdaLoRA (f) PiSSA

(g) SSF (h) AdaptFormer

(i) VPT (j) WeatherPEFT

Figure 9: RMSE improvement scorecards of ALM relative to ten E1 baselines across weather and power variables at lead times from 6 to 72 hours.

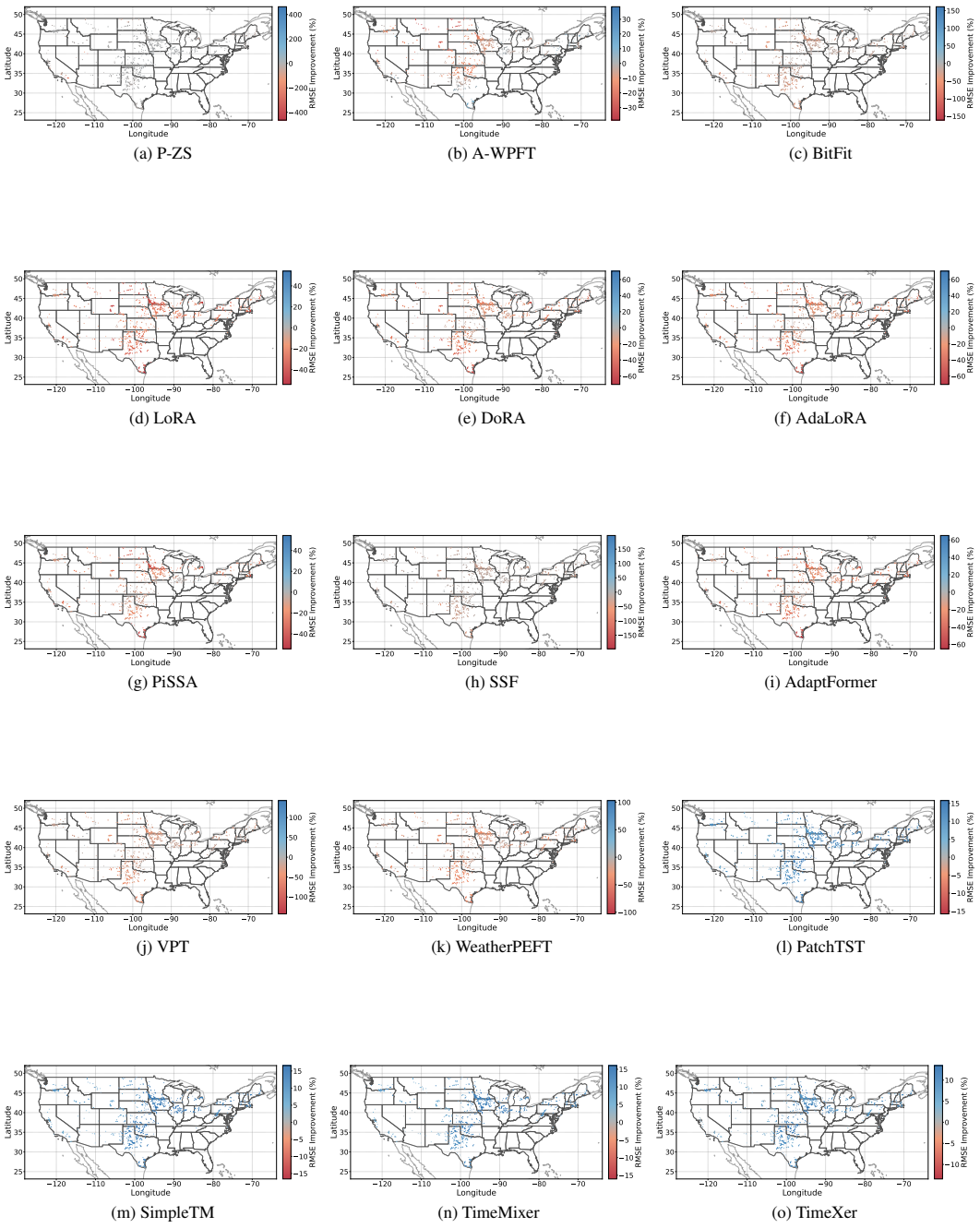
D.2 Additional Downstream Power Application Results

Table 9 presents the complete downstream RMSE and MAE results for all evaluated methods at lead times of 6, 24, and 72 hours. Across the 18 lead-time and metric combinations, ALM ranks first or second in every case. At 72 hours, ALM achieves the best plant-level wind RMSE/MAE and photovoltaic MAE, while ranking second for photovoltaic RMSE and both county-level load metrics. A-WPFT remains strongest for county load, which shows that ALM's gridded load advantage does not fully transfer through county reconstruction. This remaining gap may reflect the sensitivity of the regularized inverse to local spatial errors, although the results do not establish a specific mechanism. Overall, the table shows that ALM transfers joint gridded weather-power forecasting gains to consistently strong plant- and county-level predictions.

Table 9: Full downstream power forecasting results on the 2023 independent test set (MW).

LT	Method	Plant solar forecasting		Plant wind forecasting		County load forecasting	
		RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
6 h	P-ZS	1.84	0.95	21.01	14.25	507.95	163.04
	A-WPFT	1.43	0.83	**9.95**	**7.50**	**47.31**	**13.21**
	BitFit	2.97	1.64	19.97	13.43	123.12	29.95
	LoRA	1.73	1.05	18.84	13.76	124.26	30.51
	DoRA	1.59	0.99	17.70	11.82	120.54	29.04
	AdaLoRA	1.64	1.00	20.37	14.72	121.74	30.44
	WeatherPEFT	1.65	1.12	21.07	13.72	131.17	31.86
	VPT	2.86	2.01	21.82	14.59	157.07	39.09
	AdaptFormer	3.01	1.88	19.93	13.70	195.28	130.15
	SSF	2.30	1.34	21.44	14.56	126.67	32.33
	PiSSA	1.71	0.95	16.67	12.49	124.39	30.60
	iTransformer	1.75	0.92	23.10	16.81	131.21	37.78
	PatchTST	1.46	0.81	21.07	13.56	126.53	36.57
	SimpleTM	1.47	0.79	18.62	14.30	127.32	36.35
	TimeMixer	1.55	0.75	20.08	15.64	125.28	38.26
	TimeXer	1.56	0.85	19.96	15.74	140.92	36.31
	ALM	**1.35**	**0.72**	15.56	11.26	107.04	27.66
24 h	P-ZS	3.29	2.12	30.19	20.87	459.90	136.89
	A-WPFT	1.62	0.99	**18.31**	**12.78**	**67.66**	**19.71**
	BitFit	2.62	1.51	26.29	20.15	152.83	36.83
	LoRA	1.88	1.13	30.63	21.46	136.88	39.27
	DoRA	1.78	1.06	25.52	15.21	116.80	36.94
	AdaLoRA	2.03	1.21	27.61	19.57	139.53	38.77
	WeatherPEFT	2.25	1.26	32.18	21.75	165.09	41.11
	VPT	3.48	2.20	31.21	21.43	256.80	63.15
	AdaptFormer	2.28	1.65	30.94	20.98	268.15	149.90
	SSF	3.56	2.14	29.92	21.82	139.94	42.20
	PiSSA	1.67	1.01	26.12	18.62	127.81	38.47
	iTransformer	1.93	0.90	29.36	22.15	235.75	38.73
	PatchTST	1.69	0.99	25.47	20.04	234.13	38.22
	SimpleTM	1.71	0.90	25.73	23.33	164.39	34.29
	TimeMixer	1.92	0.89	26.57	20.89	204.34	37.31
	TimeXer	1.70	**0.87**	26.07	21.57	183.88	37.09
	ALM	**1.53**	0.88	24.16	14.99	112.71	33.32
72 h	P-ZS	3.69	2.14	30.92	22.49	455.07	139.82
	A-WPFT	2.22	1.37	32.40	24.40	**105.47**	**32.70**
	BitFit	3.05	1.97	35.31	25.40	341.94	81.07
	LoRA	1.91	1.40	34.71	27.87	304.66	83.23
	DoRA	2.35	1.36	32.00	23.28	253.07	76.91
	AdaLoRA	2.65	1.71	34.88	25.84	326.88	84.90
	WeatherPEFT	2.97	2.06	33.41	28.08	378.26	89.56
	VPT	4.25	3.04	39.40	26.38	450.67	117.98
	AdaptFormer	2.45	1.51	38.86	27.58	488.94	178.64
	SSF	4.82	3.06	35.94	24.73	268.11	86.44
	PiSSA	2.00	1.18	36.03	26.06	311.61	83.32
	iTransformer	1.93	1.05	31.26	23.24	255.44	70.55
	PatchTST	1.84	0.97	25.92	22.41	291.97	101.32
	SimpleTM	**1.65**	1.03	24.93	21.29	187.61	51.55
	TimeMixer	1.98	0.99	27.75	22.74	254.45	86.49
	TimeXer	1.80	1.00	26.77	21.67	197.18	50.38
	ALM	1.68	**0.92**	**24.64**	**18.98**	151.34	46.58

Figures 10–12 show object-level RMSE improvements over iTransformer for plant-level wind, plant-level photovoltaic, and county-level load forecasting, respectively. The baseline maps contain mixed favorable and unfavorable locations for wind and photovoltaic power, and several methods retain pronounced regional variation for county load. In comparison, Figure 4 shows predominantly positive ALM gains for both plant tasks and nearly uniform gains across counties. The spatial comparison therefore shows that ALM's joint gridded representations produce more consistent plant- and county-level forecasts than the PEFT and time-series baselines.



(a) P-ZS (b) A-WPFT (c) BitFit

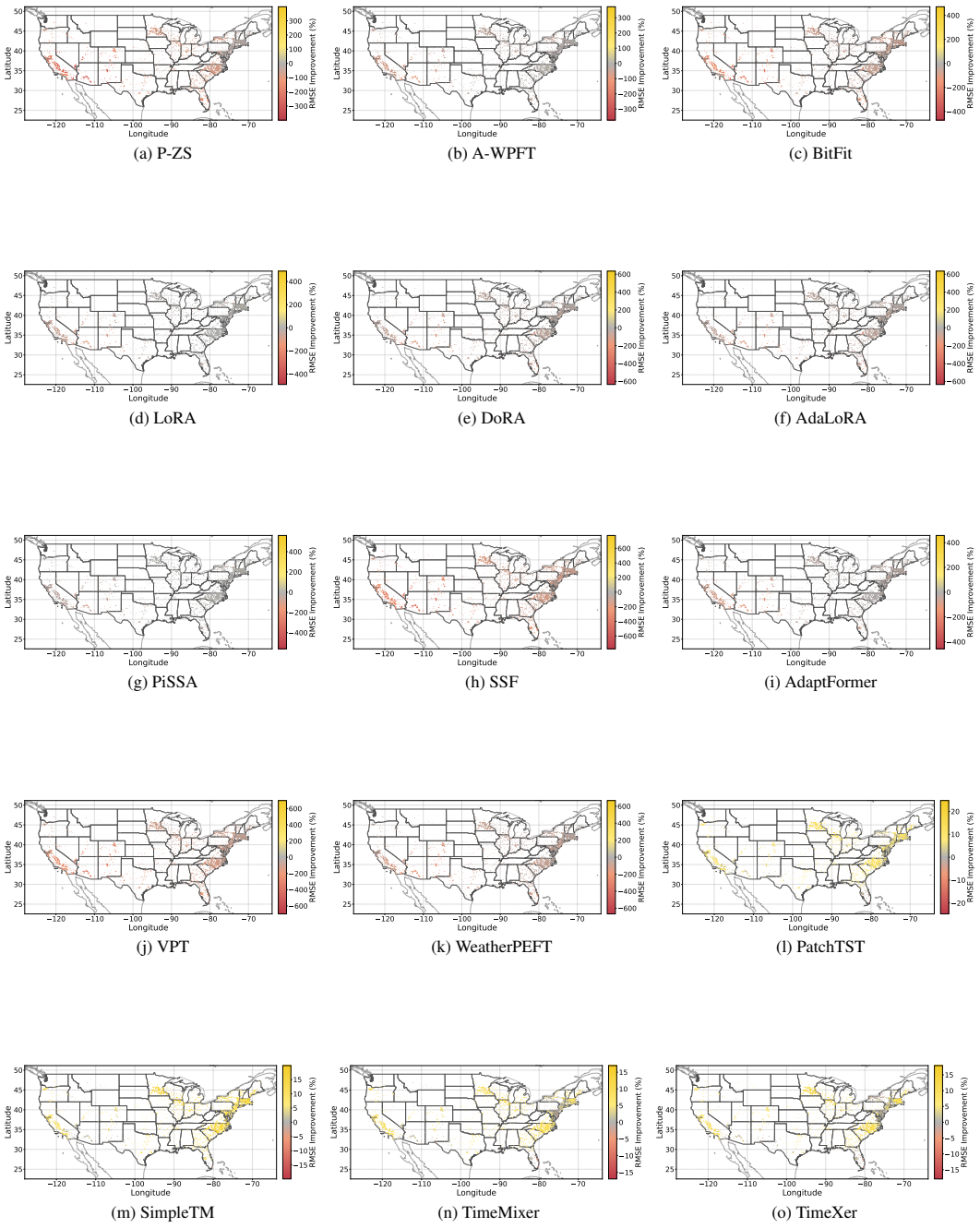
(d) LoRA (e) DoRA (f) AdaLoRA

(g) PiSSA (h) SSF (i) AdaptFormer

(j) VPT (k) WeatherPEFT (l) PatchTST

(m) SimpleTM (n) TimeMixer (o) TimeXer

Figure 10: Spatially resolved average RMSE improvement of comparison methods over iTransformer for plant-level wind forecasting at a 72-hour lead time.



(a) P-ZS (b) A-WPFT (c) BitFit

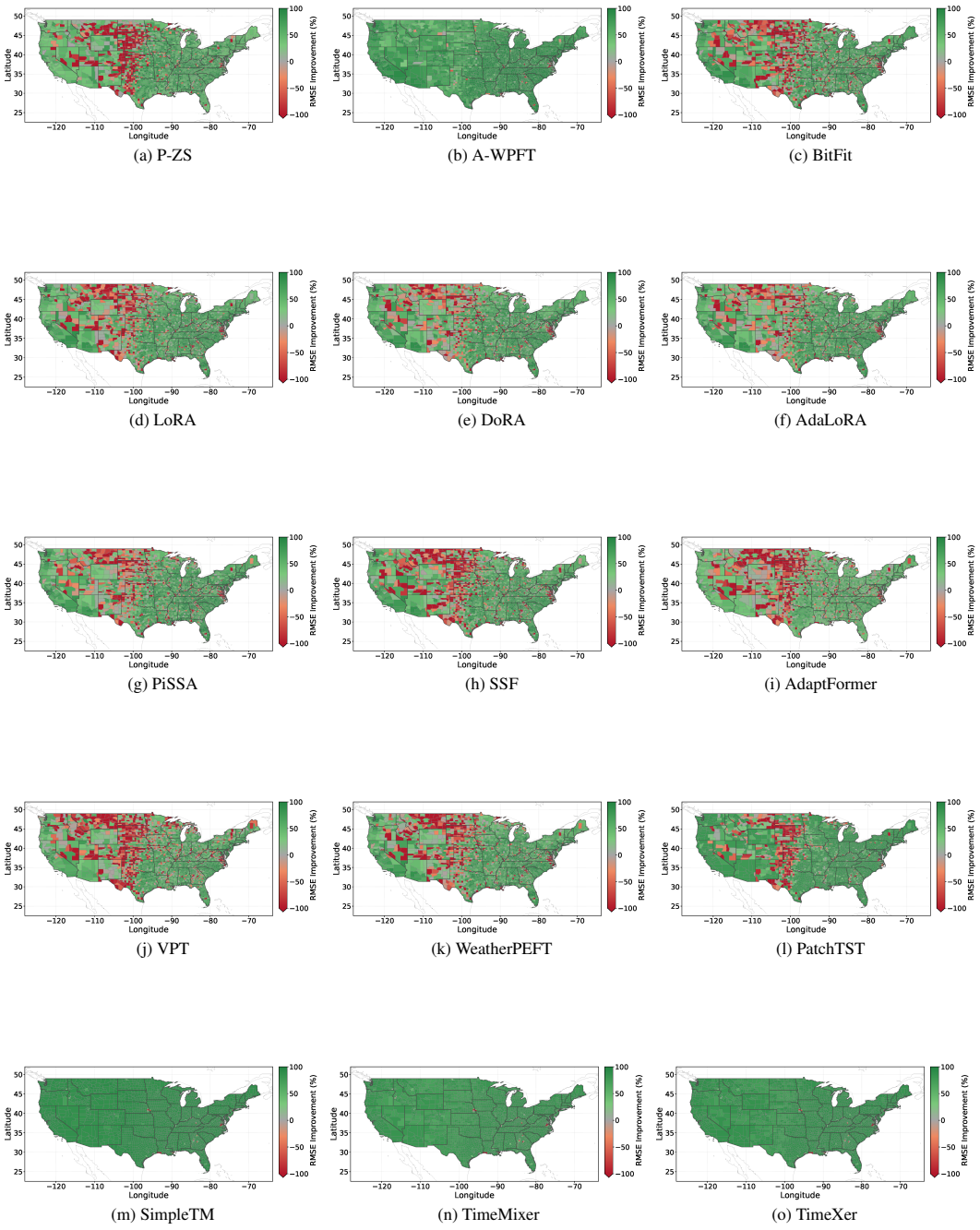
(d) LoRA (e) DoRA (f) AdaLoRA

(g) PiSSA (h) SSF (i) AdaptFormer

(j) VPT (k) WeatherPEFT (l) PatchTST

(m) SimpleTM (n) TimeMixer (o) TimeXer

Figure 11: Spatially resolved average RMSE improvement of comparison methods over iTransformer for plant-level photovoltaic forecasting at a 72-hour lead time.



(a) P-ZS (b) A-WPFT (c) BitFit

(d) LoRA (e) DoRA (f) AdaLoRA

(g) PiSSA (h) SSF (i) AdaptFormer

(j) VPT (k) WeatherPEFT (l) PatchTST

(m) SimpleTM (n) TimeMixer (o) TimeXer

Figure 12: Spatially resolved average RMSE improvement of comparison methods over iTransformer for county-level load forecasting at a 72-hour lead time.

D.3 Additional Ablation Experiments

Tables 10 and 11 present the complete component, information-path, and rollout-training ablations at all three lead times, using RMSE/ACC for weather and RMSE/MAE for power. At 72 hours, removing prefusion alignment, latent memory, postfusion modulation, or Aurora BitFit increases RMSE for every weather and power variable relative to ALM-6h under the same single-step training budget. Setting the power-to-weather injection intensity to $\eta = 0$ likewise degrades all weather RMSE values relative to ALM-6h. Rollout fine-tuning generally improves longer-horizon behavior, with occasional transfer across lead times. Notably, ALM-24h retains the lowest 72-hour load error. The full ablation results therefore show positive marginal contributions from the three ALM components under matched single-step training and a complementary benefit from progressive rollout fine-tuning.

Table 10: Full weather ablation results. η controls the power-to-weather injection intensity.

LT	Method	z		u		v		t		q		$2t$		$10u$		$10v$		msl	
		RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
6 h	A-ZS/P-ZS	220.98	0.915	4.06	0.827	5.23	0.735	1.43	0.880	0.000391	0.691	2.78	0.717	1.84	0.735	2.02	0.716	211.93	0.901
	w/o PA	68.24	0.944	1.69	0.922	1.73	0.919	0.637	0.929	0.000257	0.746	1.03	0.910	0.792	0.912	0.831	0.915	72.88	0.942
	w/o LM	70.48	0.943	1.70	0.922	1.74	0.919	0.638	0.929	0.000257	0.746	1.04	0.909	0.794	0.911	0.830	0.915	73.80	0.941
	w/o PM	69.61	0.943	1.70	0.922	1.74	0.919	0.640	0.929	0.000258	0.746	1.04	0.909	0.796	0.911	0.831	0.915	73.92	0.941
	w/o BitFit	72.23	0.943	1.86	0.917	1.91	0.913	0.691	0.925	0.000270	0.738	1.09	0.905	0.857	0.906	0.890	0.910	77.72	0.940
	$\eta = 0$	71.04	0.988	1.63	0.965	1.66	0.962	0.618	0.972	0.000243	0.779	1.04	0.946	0.771	0.952	0.792	0.958	76.47	0.986
	ALM-6h	**58.79**	**0.992**	**1.48**	**0.971**	**1.51**	**0.968**	**0.558**	**0.977**	**0.000227**	**0.785**	**0.911**	**0.957**	**0.698**	**0.960**	**0.720**	**0.965**	**62.29**	**0.990**
	ALM-24h	61.10	0.991	1.53	0.969	1.56	0.966	0.575	0.976	0.000232	0.784	0.929	0.955	0.712	0.959	0.749	0.962	65.16	0.989
	ALM-72h	67.36	0.990	1.62	0.966	1.65	0.962	0.606	0.973	0.00024	0.780	1.00	0.949	0.757	0.954	0.789	0.958	71.45	0.988
24 h	A-ZS/P-ZS	574.42	0.604	7.71	0.488	8.70	0.272	3.39	0.543	0.000774	0.406	3.51	0.744	2.82	0.456	3.38	0.322	469.04	0.626
	w/o PA	256.62	0.889	4.02	0.845	4.27	0.835	1.46	0.869	0.000477	0.651	1.68	0.908	1.56	0.838	1.67	0.839	260.32	0.867
	w/o LM	261.71	0.885	4.06	0.843	4.30	0.833	1.47	0.868	0.000478	0.649	1.70	0.907	1.56	0.837	1.68	0.838	264.24	0.863
	w/o PM	257.83	0.888	4.02	0.844	4.28	0.834	1.47	0.868	0.000479	0.647	1.69	0.907	1.56	0.837	1.67	0.840	260.83	0.866
	w/o BitFit	266.54	0.883	4.23	0.832	4.46	0.824	1.52	0.860	0.000490	0.637	1.71	0.905	1.62	0.827	1.72	0.832	269.10	0.861
	$\eta = 0$	266.24	0.911	3.86	0.873	4.14	0.861	1.41	0.902	0.000454	0.668	1.66	0.945	1.49	0.865	1.57	0.873	257.86	0.891
	ALM-6h	**224.42**	0.934	3.55	0.892	3.76	0.882	1.28	0.916	0.000420	0.686	1.49	0.955	1.36	0.887	1.45	0.890	228.17	0.911
	ALM-24h	224.66	**0.935**	**3.49**	**0.896**	**3.69**	**0.887**	**1.25**	**0.920**	**0.000405**	**0.697**	**1.40**	**0.959**	**1.34**	**0.890**	**1.43**	**0.893**	**226.77**	**0.913**
	ALM-72h	230.61	0.933	3.55	0.891	3.76	0.882	1.28	0.917	0.00041	0.692	1.44	0.957	1.36	0.885	1.46	0.890	228.78	**0.913**
72 h	A-ZS/P-ZS	1018.49	0.227	10.47	0.212	9.37	0.194	6.57	0.190	0.00109	0.143	5.59	0.452	3.43	0.232	3.76	0.191	691.26	0.247
	w/o PA	785.94	0.509	9.21	0.479	9.89	0.437	3.79	0.529	0.000971	0.336	4.12	0.727	3.43	0.453	3.77	0.446	684.85	0.455
	w/o LM	770.15	0.495	9.26	0.457	9.60	0.439	3.73	0.528	0.000977	0.334	4.04	0.728	3.41	0.442	3.72	0.440	692.23	0.429
	w/o PM	762.66	0.516	9.09	0.483	9.79	0.442	3.76	0.533	0.000980	0.331	4.12	0.723	3.43	0.455	3.80	0.443	687.87	0.458
	w/o BitFit	756.61	0.513	8.96	0.486	9.69	0.434	3.71	0.531	0.000966	0.332	4.08	0.726	3.34	0.457	3.68	0.442	690.89	0.437
	$\eta = 0$	807.26	0.441	8.89	0.456	10.30	0.371	3.68	0.502	0.000947	0.312	4.17	0.725	3.23	0.446	3.64	0.428	669.91	0.407
	ALM-6h	709.50	0.512	8.43	0.481	9.27	0.431	3.49	0.537	0.000879	0.344	3.87	0.747	3.07	0.472	3.47	0.451	631.04	0.456
	ALM-24h	673.80	0.548	8.10	0.514	8.80	0.455	3.33	0.564	0.000859	0.372	3.66	0.766	3.03	0.491	3.42	0.473	598.88	0.503
	ALM-72h	**638.53**	**0.587**	**7.49**	**0.560**	**8.14**	**0.506**	**3.08**	**0.603**	**0.00079**	**0.405**	**3.31**	**0.794**	**2.82**	**0.529**	**3.13**	**0.524**	**554.41**	**0.551**

Table 11: Full power ablation results. η controls the power-to-weather injection intensity.

LT	Method	wind		solar		load	
		RMSE↓	**MAE↓**	**RMSE↓**	**MAE↓**	**RMSE↓**	**MAE↓**
6 h	A-ZS/P-ZS	0.150	0.0888	0.119	0.0671	92.82	24.38
	w/o PA	0.118	0.0615	0.113	0.0600	9.33	3.55
	w/o LM	0.119	0.0622	0.112	0.0593	10.08	3.81
	w/o PM	0.121	0.0632	0.113	0.0601	11.75	4.27
	w/o BitFit	0.118	0.0616	0.112	0.0597	9.20	3.53
	$\eta = 0$	0.119	0.0618	0.113	0.0601	9.67	3.56
	ALM-6h	0.116	**0.0602**	0.111	0.0584	**8.91**	**3.31**
	ALM-24h	**0.115**	0.0605	**0.108**	**0.0575**	9.31	3.62
	ALM-72h	0.118	0.0627	0.111	0.0600	12.22	4.63
24 h	A-ZS/P-ZS	0.198	0.124	0.237	0.159	98.81	20.33
	w/o PA	0.181	0.0933	0.143	0.0793	13.43	4.45
	w/o LM	0.181	0.0936	0.139	0.0774	14.90	5.00
	w/o PM	0.185	0.0955	0.145	0.0810	17.14	5.70
	w/o BitFit	0.181	0.0933	0.142	0.0789	12.95	4.41
	$\eta = 0$	0.180	0.0931	0.142	0.0790	13.53	4.43
	ALM-6h	0.177	0.0913	0.134	0.0737	12.51	4.07
	ALM-24h	0.164	0.0857	0.114	**0.0584**	**12.35**	**3.95**
	ALM-72h	**0.162**	**0.0852**	**0.113**	0.0588	14.80	4.80
72 h	A-ZS/P-ZS	0.224	0.156	0.274	0.179	100.06	21.96
	w/o PA	0.216	0.112	0.176	0.101	25.61	7.14
	w/o LM	0.217	0.112	0.167	0.0962	30.90	8.59
	w/o PM	0.223	0.115	0.172	0.0989	31.43	9.46
	w/o BitFit	0.218	0.112	0.169	0.0968	24.69	7.04
	$\eta = 0$	0.216	0.111	0.172	0.0978	24.74	6.93
	ALM-6h	0.211	0.110	0.166	0.0936	21.45	6.19
	ALM-24h	0.196	0.103	0.131	0.0681	**19.34**	**5.54**
	ALM-72h	**0.187**	**0.0994**	**0.127**	**0.0655**	19.97	6.02

D.4 Cross-Backbone Generalization with ClimaX

This experiment evaluates whether ALM transfers from the hierarchical Aurora–PowerSwin configuration to a different weather foundation model and a convolutional electric branch. We replace Aurora with the open-source ClimaX foundation model (Nguyen et al., 2023) and replace PowerSwin with a residual convolutional U-Net, denoted *PowerConv*. The corresponding architecture is shown in Figure 13. The comparison is conducted entirely within the ClimaX protocol to evaluate architectural transfer. It does not compare absolute errors across the different spatial resolutions used by Aurora and ClimaX.

ClimaX adaptation protocol. We initialize the weather branch from the public ClimaX checkpoint pretrained on five CMIP6 datasets at $1.40625°$ resolution.[1] The checkpoint uses variable-specific patch embeddings, a learned variable-aggregation query, an eight-block Vision Transformer with width 1,024 and 16 attention heads, and a two-layer prediction head. All E1 fields are remapped to the checkpoint grid before the CONUS region is selected: ERA5 fields are bilinearly interpolated, wind and photovoltaic capacity factors are capacity-weighted within each target cell, and load is conservatively aggregated. The valid-cell masks are remapped together with the power fields and remain active in the training loss and evaluation metrics.

The weather inputs use the intersection between the E1 variables and the public ClimaX checkpoint: z, u, v, t, q at $50, 250, 500, 600, 700, 850$, and $925\,\mathrm{hPa}$, together with $2t$, $10u$, and $10v$. Mean sea-level pressure and the six unsupported E1 pressure levels are excluded instead of introducing randomly initialized weather embeddings that would confound backbone transfer with new-variable learning. ClimaX receives the latest weather state at t, consistent with its native forecasting interface, whereas PowerConv retains the two power states at $t - 6\,\mathrm{h}$ and t. Static land–sea mask, orography,

[1] Official ClimaX code and checkpoints: `https://github.com/microsoft/ClimaX`.

and latitude fields are supplied to ClimaX but are not prediction targets. Table 12 summarizes the differences from E1.

Table 12: ClimaX cross-backbone protocol relative to E1.

Aspect	Public ClimaX checkpoint	E1 setting	Cross-backbone setting
Grid	Global 1.40625° grid	CONUS 0.25° grid	Remap all fields, then select CONUS
Upper air	z, u, v, t, q on seven shared levels	Five variables on 13 levels	Use the seven-level intersection
Surface	$2t, 10u, 10v$ among E1 targets	$2t, 10u, 10v, \mathrm{msl}$	Exclude unsupported msl
Weather history	One state	Two states	Use the latest weather state
Weather backbone	Eight equal-width Transformer blocks	Hierarchical Aurora	Interact after blocks 2, 5, and 8
Electric backbone	Not applicable	PowerSwin	Three-scale PowerConv

PowerConv architecture. PowerConv receives wind power, photovoltaic power, and load at two consecutive input times, yielding $X_t^p \in \mathbb{R}^{6\times H_c\times W_c}$ on the remapped CONUS grid. A same-padded 3×3 convolution maps the six channels to a 64-channel shallow representation. Its encoder contains three stages with widths $(64, 128, 256)$ and two residual blocks per stage, with a stride-2 3×3 convolution between adjacent stages. Each residual block applies GroupNorm with eight groups, SiLU, a 3×3 convolution, GroupNorm, SiLU, and a second 3×3 convolution before residual addition. A 1×1 projection is used on the skip path whenever the channel width changes.

The decoder starts from the 256-channel deep feature and contains two restoration stages. Let E_1^+, E_2^+, and E_3^+ denote the outputs of the first, second, and third encoder stages after their respective ALM electric residuals have been added. The first restoration stage bilinearly upsamples E_3^+ to the spatial resolution of E_2^+, concatenates the two tensors along the channel dimension, applies a 1×1 mapping from $256 + 128$ to 128 channels, and refines the result with two residual blocks. The second restoration stage repeats this operation by upsampling to the resolution of E_1^+, concatenating the result with E_1^+, applying a 1×1 mapping from $128 + 64$ to 64 channels, and applying two residual blocks. Both decoder skip tensors therefore use the ALM-corrected encoder features instead of the pre-interaction features. The figure omits these two skip arrows for visual simplicity. A final 1×1 linear head predicts wind power, photovoltaic power, and load without clipping, while variable-specific masks restrict the loss and metrics to valid cells. PowerConv is pretrained for one-step 6-hour power forecasting using the E1 training and validation years, after which its backbone parameters are frozen during ALM adaptation.

ClimaX–ALM–PowerConv connections. ClimaX maintains a constant spatial token grid without a resolution hierarchy. Its outputs after Transformer blocks 2, 5, and 8 define the early, middle, and late weather interaction levels. The three PowerConv encoder outputs define the corresponding shallow, middle, and deep electric levels. At each level, a weather linear mapping and a power 1×1 convolution project the paired features to common widths $(64, 128, 256)$ before prefusion alignment, latent-memory interaction, and postfusion modulation. The ALM memory banks retain $(64, 32, 16)$ tokens from the early to late level. The weather residual is restored to width 1,024 and added to the ClimaX token sequence before the next Transformer group or, at the late level, before the final normalization and prediction head. The electric residual is added to the corresponding PowerConv encoder output, which is then used by the next encoder stage and retained as the decoder skip feature as specified above. The ClimaX and PowerConv heads produce weather and power forecasts, respectively, and both outputs enter the same joint loss used in E1.

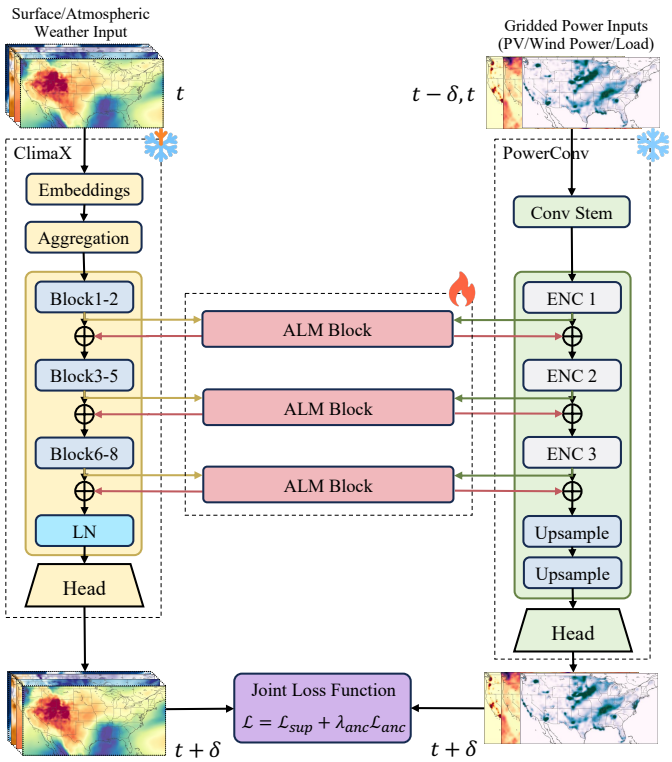


Figure 13: ClimaX–PowerConv weather-power joint adaptation through three ALM interaction levels. The two decoder skip connections and the internal ALM operations are specified in the accompanying text.

Training and baselines. The experiment retains the E1 split of 2016–2021 for training, 2022 for validation, and 2023 for testing, together with training-set normalization, a 6-hour forecasting interval, and evaluation at 6, 24, and 72 hours. After one-step PowerConv pretraining, all trainable configurations follow the E1 progressive schedule of 30 epochs with one rollout step, 10 epochs with four steps, and 6 epochs with 12 steps. At each rollout step, predicted weather and power fields replace their respective dynamic inputs, while the static ClimaX fields remain fixed. ALM training updates its three interaction blocks and ClimaX bias parameters while keeping the remaining ClimaX and PowerConv backbone parameters frozen. The weather anchor is the stop-gradient output of the current ClimaX branch with the ALM cross-branch residuals bypassed. The optimizer, loss weights, rollout-step weighting, checkpoint selection, and early-stopping rules otherwise follow the E1 settings.

The comparison includes the independent ClimaX and PowerConv zero-shot forecasts and three representative PEFT methods: BitFit, LoRA, and DoRA. For the PEFT baselines, wind power, photovoltaic power, and load are introduced as additional ClimaX variables through trainable variable-specific patch embeddings and output rows, and the selected PEFT update is applied to the resulting weather-power model. BitFit updates biases and normalization affine parameters. LoRA uses rank 8 on the attention and MLP linear mappings. DoRA uses the same rank with magnitude–direction decomposition. Every trainable method uses the same remapped data, progressive rollout schedule, joint loss, and checkpoint-selection rule. For z, u, v, t, and q, the tabulated weather metrics use the same aggregation over the seven retained pressure levels as the E1 tables.

Table 13: ClimaX-based 72-hour weather forecasting results.

Method	z		u		v		t		q		$2t$		$10u$		$10v$	
	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
ClimaX-ZS	1148.63	0.206	14.70	0.338	9.82	0.156	6.31	0.114	0.00167	0.164	8.60	0.402	4.20	0.198	4.27	0.195
ClimaX-Full	683.91	0.446	8.34	0.433	8.64	0.468	4.92	0.469	0.000961	**0.482**	4.61	0.687	**2.18**	0.484	3.98	0.478
BitFit	1097.26	0.267	12.79	0.374	11.83	0.212	6.02	0.291	0.00151	0.200	7.39	0.448	5.13	0.213	3.45	0.220
LoRA	857.73	0.381	9.82	0.414	9.72	0.381	6.22	0.371	0.00114	0.367	4.50	0.546	4.10	0.236	3.32	0.230
DoRA	836.00	0.348	9.05	0.430	9.26	0.354	5.56	0.486	0.000987	0.388	4.61	0.599	3.53	0.388	3.91	0.398
ALM	**656.01**	**0.473**	**8.19**	**0.455**	**7.35**	**0.483**	**4.32**	**0.504**	**0.000959**	0.433	**3.11**	**0.737**	3.07	**0.511**	**3.06**	**0.492**

Table 14: ClimaX-based 72-hour power forecasting results.

Method	wind		solar		load	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
PowerConv-ZS	0.267	0.186	0.268	0.194	85.36	39.95
ClimaX-Full	0.221	0.138	0.207	0.126	**20.69**	7.94
BitFit	0.292	0.182	0.250	0.195	48.96	23.51
LoRA	0.246	0.182	0.225	0.146	34.91	13.67
DoRA	0.232	0.117	0.224	0.121	36.05	11.72
ALM	**0.206**	**0.101**	**0.185**	**0.0940**	22.60	**5.92**

Tables 13 and 14 report the 72-hour results after replacing both pretrained branches and changing their feature geometries. ALM ranks first on 19 of the 22 metrics and second on the remaining three, reducing RMSE over the corresponding zero-shot branches by 25.2–63.8% across the eight weather targets and by 22.8%, 31.0%, and 73.5% for wind power, photovoltaic power, and load, respectively. Compared with full fine-tuning, ALM obtains lower RMSE on seven of eight weather targets and both renewable-power targets, while ClimaX-Full retains lower RMSE for $10u$ and load and higher ACC for q. These results show that ALM's joint forecasting gains transfer to ClimaX and PowerConv and are not specific to the Aurora–PowerSwin backbone pair.