Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Extending weather forecasts from hours to weeks with teleconnection-aware graph neural networks

Shan Zhao
Technical University of Munich
Zhitiong Xiong
Technical University of Munich
Xiao Xiang Zhu
xiaoxiang.zhu@tum.de

Technical University of Munich

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11430274/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Extending weather forecasts from hours to weeks with teleconnection-aware graph neural networks

Shan Zhao[1], Zhitong Xiong[1], Xiao Xiang Zhu[1,2*]

[1]Chair of Data Science in Earth Observation, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany.

[2]Munich Center for Machine Learning, 80333 Munich, Germany.

*Corresponding author(s). E-mail(s): xiaoxiang.zhu@tum.de;

Abstract

Subseasonal forecasting is essential for agricultural planning, water-resource management, and early warning of natural hazards, yet remains challenging because forecast errors grow rapidly with lead time. Recent machine-learning models have achieved remarkable foresting skills, but training such models requires substantial computational resources and associated energy consumption. Here, we propose a parameter-efficient framework that reuses a pretrained medium-range weather model and extends its forecast horizon from 24 hours to four weeks. The architecture retains the pretrained spatial-attention modules while introducing lightweight temporal attention to model dependencies across multiple atmospheric states. To account for the influence of large-scale boundary conditions at subseasonal timescales, we further incorporate climate teleconnection indices through a difference-based graph module. This physically motivated component learns how large-scale teleconnection patterns modulate the temporal evolution of geographically remote areas. Experiments show that the proposed framework reduces root-mean-square error for most evaluated atmospheric variables from the two-week forecast horizon onward. The framework requires updating only 10.9% of the total model parameters while enabling the processing of substantially longer temporal sequences. These results further indicate the transferability of spatial representations learned from medium-range forecasting to extended-range prediction.

1 Introduction

Subseasonal forecasting aims to predict the evolution of atmospheric conditions several weeks in advance. Skillful forecasts at these lead times are important for applications such as agriculture, water-resource planning[1], and disaster management[2,3].

The weather system is a highly complex dynamical system governed by multiscale processes and nonlinear interactions among the atmosphere, ocean and land. Traditional numerical weather modeling predicts this evolution by solving discretized forms of the governing physical equations from an estimated initial state. Although the performance has improved greatly in recent decade[4,5], their accuracy is still limited by numerical approximations and parameterizations of unresolved subgrid-scale processes[6]. Moreover, high-resolution simulations are computationally intensive and require substantial supercomputing resources.

Machine learning (ML) provides a solution by learning complex nonlinear relationships directly from large volumes data. In recent years, ML models in weather forecasting has advanced rapidly. They have achieved faster inference and better error scores on many meteorological variables, spanning tasks from medium-range weather forecast to decadal climate emulation[7]. FourCastNet introduced Fourier neural operators for short to medium range global weather forecasting[8]. PanguWeather subsequently developed a three-dimensional (3D) Earth-specific Transformer[9]. GraphCast represented the atmosphere using a multiscale graph and applied graph neural networks (GNNs) to propagate information over both local and

long-range spatial scales[10]. In parallel, ClimaX explored transformer-based foundation modelling across heterogeneous climate variables and datasets[11]. Similarly, Aurora[12] employs a 3D Swin Transformer[13] pretrained on diverse atmospheric and Earth-system datasets to support a range of forecasting tasks, including global weather prediction, air pollution forecasting, and ocean wave forecasting.

Subseasonal forecasting bridges the gap between medium-range weather prediction and seasonal forecasting, but it operates in a particularly challenging predictability lead time[14,15]. First, because of the chaotic nature of the atmosphere, information contained in the initial atmospheric state gradually loses predictive value as lead time increases[16]. Second, forecast errors can accumulate rapidly over extended horizons. Third, at subseasonal timescales, atmospheric evolution becomes increasingly influenced by slowly varying components of the Earth system and large-scale modes of climate variability. For example, the El Niño–Southern Oscillation (ENSO), which is the sea-surface temperature anomalies in the central and eastern tropical Pacific, can influence weather far beyond its region of origin through large-scale teleconnections[17,18].

Recent studies have begun to extend ML weather models toward subseasonal prediction[19]. However, many models uses the autoregressive rollout for extending the horizons. Such strategies are particularly susceptible to rapid error accumulation at extended lead times[20]. Moreover, these models often do not explicitly represent slowly varying large-scale climate drivers, potentially neglecting important sources of subseasonal predictability. A further challenge is computational cost. Despite their fast inference, most ML weather models rely on large architectures and computationally intensive training, making task-specific adaptation and downstream development difficult under limited computational resources. This high computational demand also increases energy consumption and associates with environmental concerns.

In this work, we propose a parameter-efficient strategy for transferring a pretrained medium-range weather model to extended-range forecasting. We ask if the pretrained spatial representations can be reused while adapting only lightweight modules to capture the additional temporal changes. Specifically, we reuse the 3D Earth-specific spatial-attention backbone of PanguWeather and introduce trainable temporal modules for capturing the temporal changes within multiple frames. To account for the influence of large-scale ocean-atmosphere patterns on local weather anomalies, we further introduce a teleconnection-aware graph module (TeleDiffGraph). We construct a heterogeneous graph with two types of nodes representing teleconnection indices and spatial weather anomalies. Edges between two types of nodes are initialized using prior knowledge, while their strengths are subsequently learned through differentiable message passing. To improve the spatial granularity of the prediction, we also added the spatial spectral loss to encourage the learned fields preserves the high frequency information in the weather dynamics. The proposed architecture provides a computationally efficient and physically motivated framework for extending the weather forecasting horizons.

2 Results

2.1 Global Teleconnection Graph

The teleconnection-aware graph module is built on a heterogeneous graph that connects 15 ocean-climate indices (OCIs) nodes with weather nodes. Each OCI node is represented by its 24-month history time series, while each weather node encodes the change in the local atmospheric state between two adjacent time steps. Cross-type edges are initialized using prior knowledge of the geographical regions impacted each teleconnection pattern (Method 5.3.2). Table 3 in Appendix A lists all input variables used in the model. The resulting graph structure is illustrated in Fig. 1.

During message passing, the OCI node feature first propagates to the connected weather nodes, which represents how large-scale climate indices modulate local weather changes. The 3D Earth-specific spatial-attention blocks capture spatial interactions primarily through local windowed self-attention, while shifted-window operations enable information exchange across neighboring windows in successive layers. Consequently, long-range dependencies are represented through progressive propagation across multiple spatial-attention blocks. The teleconnection-aware graph complements this process by introducing explicit nonlocal pathways through OCI nodes. Through iterative message aggregation and feature updates, geographically remote weather nodes associated with the same climate mode can exchange information indirectly via a shared OCI node. This allows the model to represent large-scale teleconnection-driven dependencies.

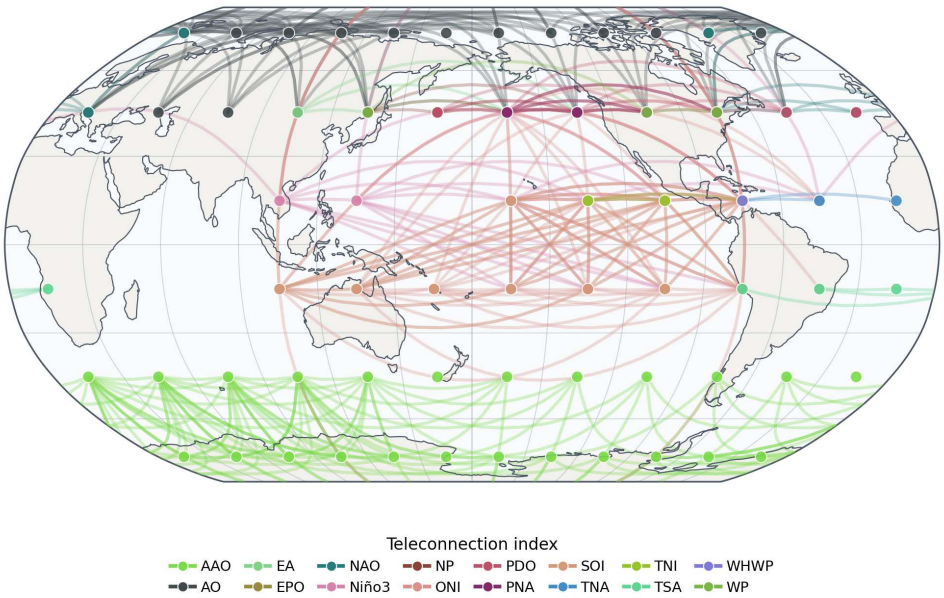


Fig. 1: Grid-to-Grid network converted from the teleconnection indices. We visualize 60% of the total connection for better visibility. In addition to the spatial interactions captured within individual 3D attention windows, the teleconnection graph introduces long-range dependencies between geographically distant regions associated with common large-scale climate patterns at the same layer.

2.2 Subseasonal Performance

We evaluate the forecasting skill of the proposed framework at subseasonal lead times ranging from 7 to 28 days. Forecast performance is assessed for five upper-air variables geopotential (Z), specific humidity (Q), temperature (T), and the zonal and meridional wind components (U, V) at 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa), together with four surface variables: 2-m temperature (T2M), 10-m zonal (U10) and meridional wind components (V10), and mean sea-level pressure (MSL). We compare our framework against four types of baselines. First, we include two ML weather forecasting models, PanguWeather[9] and GraphCast[10]. Second, the European Centre for Medium-Range Weather Forecasts (ECMWF)'s operational Integrated Forecasting System is included as a numerical weather prediction (NWP)[21] baseline. Third, ClimaX is used as a weather and climate foundation-model baseline. We used the finetuned ClimaX at each targeting horizon for comparison. In addition, climatology is included as a reference baseline. Results for these baseline models, together with the corresponding references, are taken from ChaosBench[20], which is a benchmark designed for evaluating subseasonal-to-seasonal forecasting models.

Figure 2 shows the latitude-weighted Root Mean Squared Error (RMSE) of our model compared to one ML (PanguWeather) and NWP baseline from 7- to 28-day lead times for four selected variables. Our model began to outperform the ML and NWP baseline with lower RMSE approximately the 14-day lead time onward. Despite these improvements relative to the ML forecasting baselines, climatology remains highly competitive at subseasonal lead times. Similar to the other models considered here, our framework does not consistently outperform climatology across variables and forecast horizons. An exception is the 50-hPa zonal wind (U50), for which the proposed framework achieves a lower RMSE than the climatological baseline. Fig. 3 and Fig. 4 further plot the RMSE and Anomaly Correlation Coefficient (ACC) of the atmospheric variables across 13 pressure levels at 3-4 weeks ahead compared with Pangu-Weather, GraphCast, ECMWF, ClimaX, and climatology. Overall, the proposed model achieves lower RMSE than the ML baselines. At both 21- and 28-day lead times, ACC generally decreases toward zero for most variables and models. The proposed model achieves competitive or improved ACC for selected upper-atmospheric variables, particularly for upper-level humidity and zonal wind, but performs below the strongest baselines for several variables in the middle and lower troposphere.

Figure 5 compares the normalized RMSE growth aggregated over all evaluated atmospheric variables. For most forecasting systems, the normalized error increases progressively with lead time, reflecting the accumulation of forecast errors as predictions are extended further from the initial state. The growth is particularly pronounced for models that rely on repeated autoregressive updates such as GraphCast and

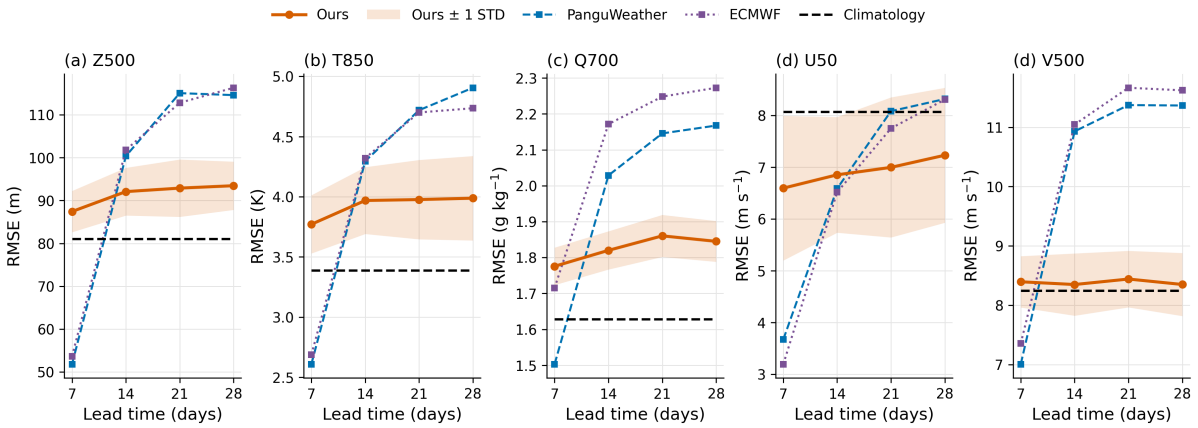


Fig. 2: Comparison of RMSE across our model and the baseline methods at forecast horizons from 1 to 4 weeks. The shaded area indicates the standard deviation across test samples. At short- to medium-range horizons, our model remains behind the baselines. From 14 days onward, however, our model achieves the lowest RMSE overall. Climatology serves as a particularly strong baseline in the subseasonal scale, where many ML models tend to lose skill. Notably, our model approaches climatological performance across several variables and even surpasses it for certain fields, such as U50.

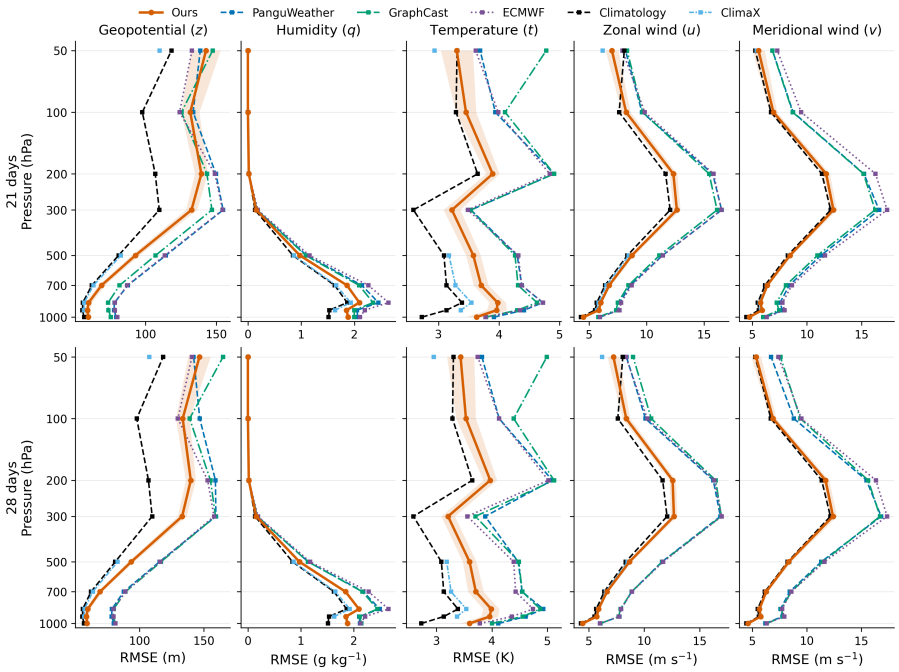


Fig. 3: Comparison of RMSE for pressure-level variables across different forecasting models. The shaded regions indicate the standard deviation across test samples. Our model yields lower RMSE than the other ML weather forecasting models for most variables and approaches the long-term climatological baseline at extended forecast horizons. Finetuned ClimaX remains competitive, particularly in the upper atmosphere, where it achieves lower RMSE for several variables.

PanguWeather. The proposed multi-frame formulation exhibits a slower normalized error-growth rate at extended lead times. Instead of advancing the atmospheric state one frame at a time, the model jointly predicts four future frames from four preceding states. Consequently, errors introduced at an individual prediction step are propagated through fewer recurrent updates.

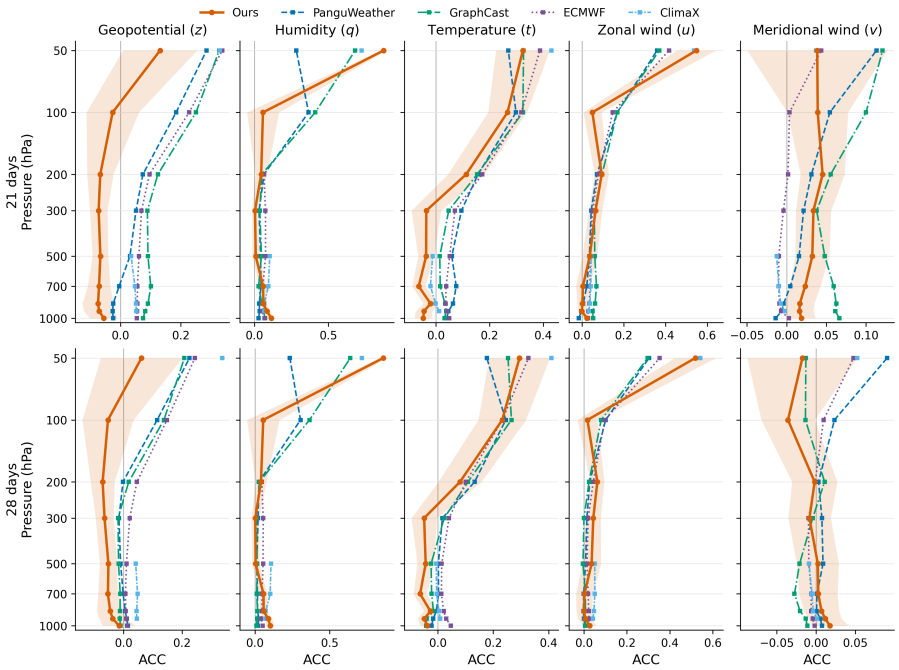


Fig. 4: Comparison of ACC for pressure-level variables across different forecasting models. Our model does not achieve the highest ACC for most variables; however, it remains competitive for wind components such as U and V, which exhibit finer-scale spatial variability than many of the other evaluated variables.

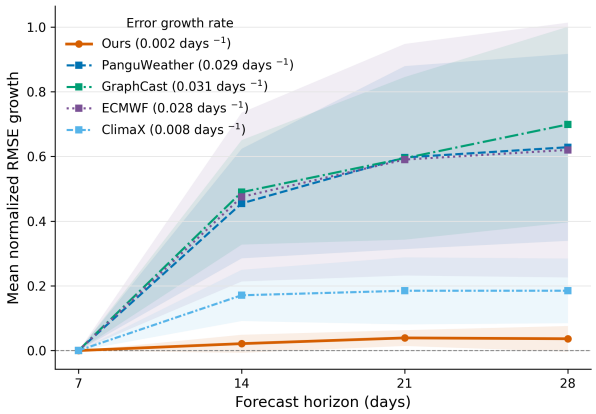


Fig. 5: Error accumulation with increasing forecast horizon. The mean normalized increase in RMSE is evaluated from 7 to 28 days, with shaded regions indicating variability across variables. Autoregressive weather models show pronounced error growth with increasing lead time, whereas ClimaX exhibits a flatter trajectory because predictions are made directly at the corresponding target horizons. Our model shows the lowest error growth rate across all evaluated horizons, suggesting that multi-frame prediction mitigates the accumulation of forecast errors.

2.3 Computational Efficiency

Our model contains 293.2 million parameters, of which only 32.0 million (10.9%) are trainable. The remaining parameters are inherited from the 3D Earth-specific spatial-attention backbone of PanguWeather at 24 hour horizon and kept frozen during adaptation. Fine-tuning was performed for only five epochs using 300 training samples and required approximately 20.6 GPU-hours on a single NVIDIA A100 80 GB GPU. Despite the four times longer temporal input-output sequence, only a small fraction of the model parameters needs to be updated.

Table 1 summarizes the model sizes and reported computational requirements of representative ML weather forecasting systems. The comparison illustrates the practical advantage of reusing pretrained

Table 1: Comparison of model sizes and reported computational requirements. The computational cost of our model corresponds to fine-tuning, whereas the costs of the baseline models correspond to training cost reported in Zhu et al.[22]. Frames T_{in}-T_{out} indicates the number of input (T_{in}) and output (T_{out}) time steps processed per training iteration.

Model	Frames	Spatial resolution	**Trainable parameters**	**Hardware**	**Training time**
GraphCast	2-1	0.25	36.7M	32 TPU v4	4 weeks
Pangu-Weather	1-1	0.25	256M	192 V100	15 days
ClimaX	1-1	5.65	110M	80 V100	3 days
Ours	4-4	0.25	**32.04M**	**1 A100**	**20 hours**

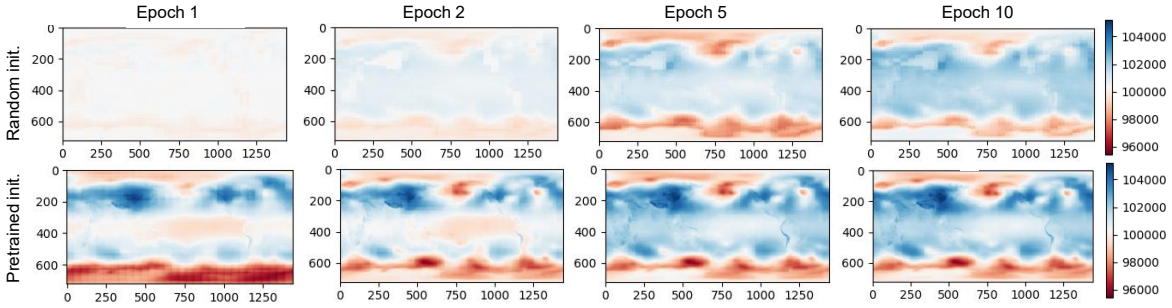


Fig. 6: Evaluation of pretrained spatial representations. Sample MSL predictions from models trained with random initialization (top) and initialized from pretrained 3D Earth-specific Transformer weights (bottom) are shown after epochs 1, 2, 5, and 10. The pretrained initialization leads to substantially faster convergence: large-scale pressure patterns are already recovered after the first epoch and remain spatially well structured throughout training. In contrast, the randomly initialized model develops comparable spatial organization only after several epochs.

spatial representations: adaptation to the extended-range forecasting task can be achieved with a single GPU and substantially fewer trainable parameters, rather than repeating the computationally expensive pre-training process. Besides, such lightweight adaptation strategy leaves room to incorporate richer temporal modelling without substantially increasing the number of trainable parameters. By factorizing spatial and temporal attention, the architecture reduces the additional attention memory complexity associated with modeling temporal dependencies from $\mathcal{O}(T^2N^2)$ for joint spatiotemporal attention to $\mathcal{O}(T^2N)$ for factorized temporal attention, where T denotes the number of time steps and N the number of spatial tokens, with N typically much larger than T (Method 5.2).

To further evaluate the efficiency of the pretrained spatial attention weights, we compare training with the pretrained weights against training from random initialization. Qualitative predictions at epochs 1, 2, 5, and 10 at Fig. 6 show that the pretrained spatial features improves faster capturing of the spatial structure earlier than the randomly initialized counterpart. Even at the first epoch, the pretrained model already reproduces the dominant spatial pattern of MSL, whereas the randomly initialized model requires more training to develop comparable structure. This indicates that spatial representations learned for medium-range forecasting remain transferable when the downstream task is extended to longer-horizon, multi-frame prediction.

3 Discussion

3.1 Temporal attentions

In our framework, temporal attention is introduced after the pretrained spatial blocks to model dependencies across input frames at corresponding spatial locations (Method 5.2). It quantifies how the representation of spatial features at a given spatial-vertical window depends on the corresponding representations across different input frames. With two input frames, mean temporal-attention matrices remain close to uniform across the first four layers, with weights near 0.5 for both frames. With three frames, deeper layers exhibit more differentiated temporal weighting. For example, layers 2 and 3 assign non-uniform weights across the three frames, whereas layers 1 and 4 remain closer to one-third. These

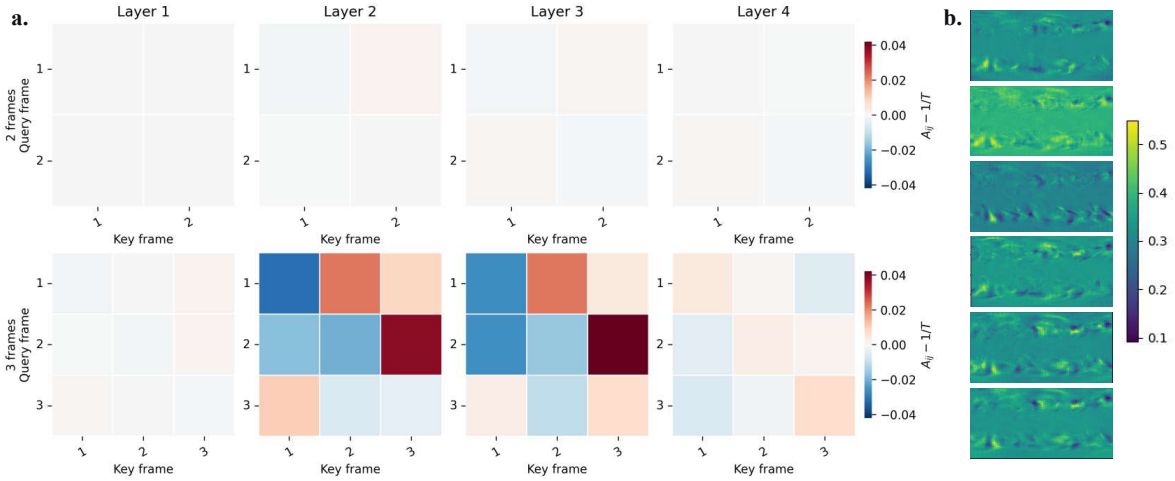


Fig. 7: Temporal attention patterns across layers and their spatial variability. (a) Temporal attention deviations from uniform weighting ($A_{ij} - 1/T$) averaged over space and heads, across four attention layers for models using two input frames (top) and three input frames (bottom). With two frames, the learned temporal attention remains close to uniform, whereas the three-frame configuration exhibits stronger temporal differentiation, particularly in the intermediate layers. (b) Sample visualization of weights frame 0 attending to frame 1, across six attention heads for the model using three input frames. The model learns spatially heterogeneous temporal attention, with distinct patterns at mid- to high latitudes, where atmospheric dynamics exhibit stronger regional variability.

observations suggest that a longer temporal context permits the model to learn non-uniform temporal preferences.

3.2 Ablation on model components

This section evaluates the contribution of the proposed model components by reporting RMSE for four surface variables at a 14-day forecast horizon. To reduce the computational cost of the ablation study, we use a relatively lightweight configuration with two input frames and two output frames.

As shown in Table 2, the autoregressive rollout of the 24-hour PanguWeather model (AR-Pangu) produces the largest RMSE across all four variables, with errors of 890.03 for MSL, 5.08 for U10, 5.26 for V10, and 3.32 for T2M. This indicates substantial error accumulation when the one-day model is repeatedly rolled out to the 14-day horizon.

Fully fine-tuning PanguWeather considerably reduces the forecast error, achieving RMSE values of 728.25, 4.05, 4.06, and 2.72 for MSL, U10, V10, and T2M, respectively. Compared with the autoregressive PanguWeather baseline, introducing temporal attention (Ours-TA) reduces RMSE across all four variables, achieving performance comparable to the fully fine-tuned PanguWeather model and slightly outperforming it for U10 (4.03 vs. 4.05). However, incorporating a conventional GNN to model interactions directly on the atmospheric states (Ours-Graph) increases RMSE across all four variables, with corresponding RMSE values of 744.95, 4.17, 4.10, and 2.87. In contrast, replacing the conventional GNN with DiffGraph (Ours) consistently improves performance across all four variables compared with both ablations. The full proposed model achieves the lowest RMSE for MSL (723.91) and U10 (4.01), outperforming the fully fine-tuned PanguWeather baseline, while matching its performance for V10 (4.06) and T2M (2.72). These results highlight that modeling graph interactions through temporal differences, rather than directly on the full atmospheric states provides a more effective representation of atmospheric dynamics, while temporal attention enables the integration of multiple input time steps for modeling such temporal dependencies.

4 Conclusion

In this work, we investigated how spatial representations learned by a short-range weather forecasting model can be efficiently reused and extended to multi-frame subseasonal prediction. We reuse the spatial attention which is a common module in many ML weather models, and augment it with two lightweight components: temporal attention and a teleconnection-aware difference graph module. The temporal

Table 2: Ablation study at a 14-day forecast horizon. Checkmarks indicate the components included in each model. Lower RMSE is better, and the best result for each variable is shown in bold.

Model	Components			RMSE at $H = 14$ days			
	Fully fine-tuned	Temporal attention	GNN	MSL	U10	V10	T2M
AR-Pangu				890.03	5.08	5.26	3.32
PanguWeather	✓			728.25	4.05	**4.06**	**2.72**
Ours-TA		✓		731.60	4.03	4.10	2.73
Ours-Graph		✓	✓(Graph)	744.95	4.17	4.09	2.87
Ours		✓	✓(DiffGraph)	**723.91**	**4.01**	**4.06**	**2.72**

attention follows a factorized space-time design, in which temporal interactions are modeled separately from the pretrained spatial processing. As a result, the additional computational cost only grows with the number of input frames, rather than requiring full joint space-time attention. The teleconnection-aware graph introduces a heterogeneous graph containing teleconnection-index nodes and spatial-grid nodes. Message passing first propagates information from large-scale circulation indices to the spatial grid. Through repeated teleconnection–grid–teleconnection interactions, geographically distant regions can exchange information indirectly, thus forming the long-range dependencies beyond the local spatial neighborhoods. The graph connectivity is initialized from prior knowledge of known teleconnection structures, while the corresponding edge weights are learned from data. Together, the model learns longer temporal context while updating only a small fraction of the original model parameters.

Although the proposed model does not consistently outperform the strongest baselines at short lead times, its relative advantage becomes larger at longer horizons, particularly from two weeks onward. Our experiments also show the benefit of extended temporal context.The sequence-to-sequence formulation reduces the number of autoregressive rollout steps compared with repeatedly applying a short-range model, which in turn slows the accumulation of forecast error at subseasonal lead times. The model also shows greater variation in temporal attention with the increasing number of input frames, indicating a more structured temporal dependencies learned with longer historical context.

Importantly, these improvements are obtained with limited additional training cost. Despite processing four frames in both the input and output sequences, the proposed adaptation updates only approximately 10.9% as many trainable parameters as full-model fine-tuning and is trained using only about 2.8% of the available reanalysis data. The comparison between pretrained and randomly initialized models further demonstrates faster convergence when the pretrained spatial features are used. These results indicate that spatial representations learned for short-range weather forecasting remain useful at longer prediction horizons, and that adapting temporal dynamics may be more important than relearning the full spatial representation from scratch.

The ablation experiments show that the DiffGraph module further reduces forecast errors relative to applying graph interactions directly to the full atmospheric state. By operating on temporal differences, the graph emphasizes evolving anomalies and state transitions rather than the absolute states.

However, despite the use of a spatial spectral loss to encourage the preservation of high-frequency information, the forecasts still exhibit oversmoothed spatial pattern relative to the ground truth. At longer lead times, most models, including ours, do not consistently outperform the climatological baseline across all variables, although our model approaches or even surpasses it for certain variables, such as U50. This raises the question of whether improvements in RMSE necessarily reflect better predictability or are partly driven by a tendency toward climatological mean states. Future work should therefore focus on better preserving atmospheric anomalies and small-scale structures, as well as assessing predictive skill beyond conventional error metrics. Besides, probabilistic forecast is important in seasonal to subseasonal scale. Future model development should consider the ensemble forecast[23] or probabilistic model with uncertainty estimation. The temporal and graph modules could also be extended to adaptive graph structures. We also plan to evaluate model performance on case studies and extreme events.

Overall, our results show that pretrained large-scale spatial representations can be retained and complemented with lightweight modules that introduce the temporal and long-range dependencies required for longer-range forecasting. This provides a practical way to develop weather models when computational resources are limited. The same idea may also be useful for other Earth-system applications, such as regional forecasting, downscaling, and extreme-event prediction, where the spatial representation remains relevant but the required temporal scale or interaction dynamics changes.

5 Method

Let $\mathbf{X}_{\text{upper}} \in \mathbb{R}^{B\times T_{\text{in}}\times C\times Z\times H\times W}$ denote the upper-air atmospheric variables, $\mathbf{X}_{\text{surface}} \in \mathbb{R}^{B\times T_{\text{in}}\times C\times H\times W}$ denote the surface variables. In addition, we have a collection of large-scale OCIs, $\mathbf{X}_{\mathbf{oci}} \in \mathbb{R}^{C\times T}$, where B is the batch size, T_{in} is the number of input time steps, C is the number of variables, and Z denotes the number of pressure levels.

The objective is to predict multiple future atmospheric states,

$$\widehat{\mathbf{Y}}_{\text{upper}}, \widehat{\mathbf{Y}}_{\text{surface}} = f_\theta\left(\mathbf{X}_{\text{upper}}, \mathbf{X}_{\text{surface}}, \mathbf{X}_{\text{oci}}, \mathbf{M}\right), \tag{1}$$

where $\widehat{\mathbf{Y}}_{\text{upper}}$ and $\widehat{\mathbf{Y}}^{\text{surface}}$ denote the predicted upper-air and surface variables, respectively and $\mathbf{M}$ is the static masks including the topography mask, land-sea mask and soil type mask.

5.1 Model Overview

Our model extends an Earth-specific weather Transformer with two complementary components: (i) temporal self-attention for modeling the evolution of each atmospheric token, and (ii) a teleconnection-informed difference graph that injects large-scale climate variability into local atmospheric changes.

The model follows a multi-resolution encoder-decoder architecture (Fig. 8). The input fields are first converted into three-dimensional atmospheric patches. Earth-specific window attention extracts spatial and vertical interactions, while temporal attention models dependencies along the temporal dimension by treating each spatial grid location as an independent batch element. A physics-motivated difference graph is initialized using prior knowledge of the potential regions influenced by each teleconnection pattern. The resulting graph propagates information from teleconnection nodes to the corresponding geographical grid cells and is iteratively refined across multiple spatial scales. The resulting features are decoded into multi-frame forecasts.

5.2 Temporal Modeling

Extending the input sequence can provide more supporting information for long-term prediction. However, applying joint space-time attention to all atmospheric tokens would require every token to attend simultaneously to all spatial, pressure-level, and temporal positions. Joint attention over T frames would then operate on TN tokens, resulting in a computational and memory complexity of $\mathcal{O}\left(T^2N^2\right)$, where $N = ZHW$ is the number of token for each 3D attention block. This is computationally expensive for global atmospheric fields, especially when N is large.

Inspired by TimeSformer[24], we divides space-time attention into two successive operations. Rather than learning joint attention over all pressure-space-time tokens, the model first applies Earth-specific spatial-pressure attention to each time step and subsequently applies temporal attention independently at every spatial-pressure token. This factorization also allows us to reuse the pretrained Earth-specific Transformer blocks without the necessary to learn spatial features from the scratch.

Specifically, token dimensions are reorganized such that the temporal sequence at each fixed spatial-pressure location is processed independently. We use $i = (z, h, w)$ to denote a particular pressure-latitude-longitude token. The feature sequence at location i is $\mathbf{H}_i \in \mathbb{R}^{T\times D}$. The temporal self-attention follows: For each location i, we have the temporal queries $\mathbf{Q}_i$, keys $\mathbf{K}_i$, and values $\mathbf{V}_i$. The temporal attention is subsequently evaluated along the T time steps:

$$\text{TA}\left(\mathbf{H}_i\right) = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i, \tag{2}$$

where d_h is the dimensionality of the values[25].

The resulting tensor is rearranged back to its original shape and will be used as the input of the TeleDiffGNN module.

The temporal-attention complexity is $\mathcal{O}\left(NT^2\right)$ because temporal attention is evaluated separately at each of the N spatial-vertical locations. Considering a much smaller T compared to N, this design avoids the $\mathcal{O}(T^2N^2)$ cost of joint space-time attention while preserving both atmospheric spatial structure and temporal dependencies.

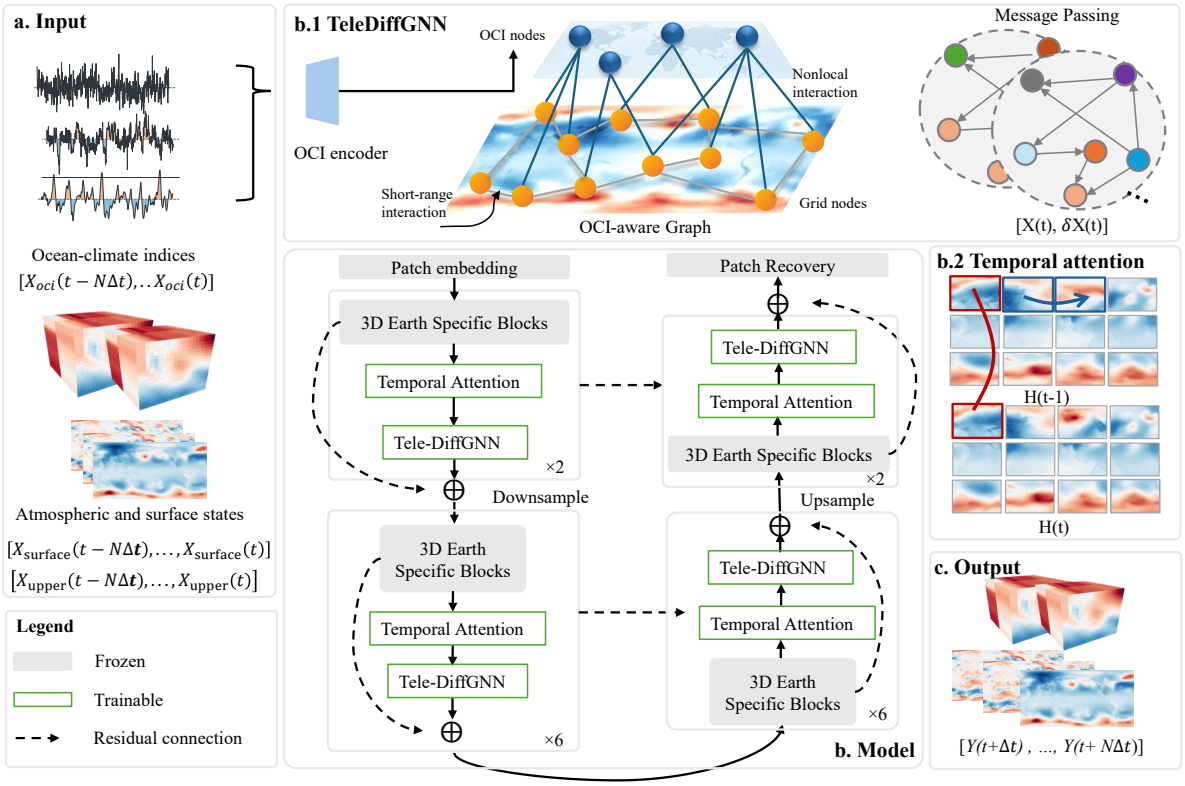


Fig. 8: Model architecture. The model uses multi-frame atmospheric and surface states and ocean-climate indices series (a). Tele-DiffGNN constructs an OCI-aware graph in which OCI nodes encode climate-index histories and grid nodes represent temporal changes in atmospheric spatial features. Message passing between OCI and grid nodes encourages the nonlocal, climate-informed information exchange (b.1). The temporal attention figure illustrate the difference between the cross-window spatial information propagation (blue) and temporal attention between successive hidden states (red) (b.2). Gray and green blocks denote frozen and trainable components, respectively. The hierarchical encoder-decoder (b) produces multi-step forecasts of atmospheric and surface states (c).

5.3 Physics-Motivated Teleconnection Difference Graph

The multi-frame setting provides longer temporal context that the model can leverage, and enable it to capture the temporal transitions and evolving changes across a sequence.

5.3.1 Physical motivation

We motivate the proposed graph formulation from the classical diffusion equation. Let $\mathbf{u}(\mathbf{s}, t)$ denote a physical state variable at location $\mathbf{s}$ and time t. A diffusion process can be expressed as

$$\frac{\partial \mathbf{u}(\mathbf{s}, t)}{\partial t} = \kappa \nabla^2 \mathbf{u}(\mathbf{s}, t), \tag{3}$$

where κ is the diffusion coefficient and ∇^2 is the spatial Laplacian. The Laplacian measures the difference between the state at a location and its surrounding states. For example, under a standard discrete spatial approximation,

$$\nabla^2 \mathbf{u}_i \approx \sum_{j \in \mathcal{N}(i)} w_{ij} \left(\mathbf{u}_j - \mathbf{u}_i\right), \tag{4}$$

where $\mathcal{N}(i)$ denotes the neighbourhood of location i and w_{ij} controls the interaction strength. This formulation can naturally extend to a graph domain[26]. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$, the graph Laplacian is

$$\mathbf{L}_G = \mathbf{D} - \mathbf{A}. \tag{5}$$

For graph signal $\mathbf{X}$, its Laplacian operation can be written node-wise as

$$-(\mathbf{L}_G\mathbf{X})_i = \sum_{j\in\mathcal{N}(i)} w_{ij}\left(\mathbf{x}_j - \mathbf{x}_i\right). \tag{6}$$

Consequently, Eq. 3 using discrete graph update is

$$\mathbf{x}_i' = \mathbf{x}_i + \alpha \sum_{j\in\mathcal{N}(i)} w_{ij}\left(\mathbf{x}_j - \mathbf{x}_i\right). \tag{7}$$

Equations 6 and 7 establish the connection between a physical Laplacian and graph message passing: physical spatial gradients are represented by differences between connected nodes, while the graph connectivity determines the path along which information can propagate. However, this formulation assumes a linear and local interactions, while the teleconnection responses can have diverse or even opposite response across different regions and in remote areas. We therefore generalize the fixed kernel in Eq. 7 using a learnable nonlinear message function.

5.3.2 Teleconnection-aware Difference Graph Neural Network

This part introduces the Teleconnection-aware Difference Graph Neural Network (TeleDiffGNN). In our task, the heterogeneous graph $\mathcal{G}$ contains two sets of nodes

$$\mathcal{G} = (\mathcal{V}_g \cup \mathcal{V}_c, \mathcal{E}), \tag{8}$$

where $\mathcal{V}_g$ contains geographical grid nodes and $\mathcal{V}_c$ contains teleconnection nodes corresponding to large-scale OCIs.

The initial connectivity is constructed from prior knowledge of the potential geographical influence region of each teleconnection pattern, as shown in Appendix Fig. 12. The cross-type edge exist if the teleconnection mode has a potential influence on the geographical grid cell. Each teleconnection is represented by a 24-dimensional historical sequence and embedded into feature space by simple linear layer, and the node feature of geographical nodes are initialized with the difference of the status temporal attention evolved features. The GNN enables message passing between two types of nodes and performed by aggregating the information from all neighboring nodes ($\mathcal{N}$) and nonlinear transformation. Finally, the updated temporal evolvement is added back to the base-states to have the final representation:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \sum_{j\in\mathcal{N}(i)} \phi_\theta\left([\Delta_t\mathbf{h}_i^t \| \Delta_t\mathbf{h}_j^t]\right), \tag{9}$$

where $\|$ is the concatenating operation and $\mathbf{h}_i$ is the feature of node i. The proposed formulation can therefore be viewed as replacing the predefined linear diffusion interaction in Eq. 7 by more generalized transformation and can be interpreted as learning how large-scale climate modes modulate local atmospheric changes. The prior graph specifies where such effects are possible. Candidate influences are identified from relationships reported in the literature[17,18,27,28], supplemented by additional relationships suggested through an Large Language Model-assisted knowledge search[29]. Since both sources may be incomplete or uncertain, the learned edge weights provide flexibility to adjust the strength of the corresponding teleconnection influences during training.

5.4 Experiment

The model was trained using a weighted mean absolute error (MAE) over both pressure-level and surface variables. The data normalization and variable-specific weights are set following Bi et al.[9]. To better preserve the spatial granularity of the predicted fields, we added a spatial spectral loss[30] that penalizes discrepancies between the spatial spectra of the prediction and the ground truth. Specifically, we compute the spatial Fourier transform of both fields and minimize the ℓ^1 difference between their spectra:

$$\mathcal{L}_{\text{spectral}} = \sum_k \left|\mathcal{F}_{\text{spat}}[\mathbf{Y}](k) - \mathcal{F}_{\text{spat}}[\hat{\mathbf{Y}}](k)\right|, \tag{10}$$

where $\mathcal{F}_{\text{spat}}$ denotes the spatial Fourier transform and k indexes the spatial frequencies. To focus on learning the small-scale spatial patterns, we only include the highest 80% of the spatial frequency bands into the loss term. This encourages the model to better preserve high-frequency information which associated with small-scale and localized features. The total loss was defined as

$$\mathcal{L} = \mathcal{L}_{\text{upper}} + \lambda_{\text{surface}}\mathcal{L}_{\text{surface}} + \lambda_{\text{spectral}}L_{\text{spectral}}, \tag{11}$$

where $\mathcal{L}_{\text{surface}}$ and $\mathcal{L}_{\text{upper}}$ correspond to surface and upper-level weighted MAE losses. λ_{surface} is set as 0.25 and $\lambda_{\text{spectral}}$ is set as 0.1.

We adopt the 3D Earth-specific Transformer backbone and spatial window configuration from PanguWeather[9]. Temporal attention is integrated into the four backbone stages using $[6, 12, 12, 6]$ attention heads, respectively. For the TeleDiffGNN module, at each stage, the teleconnection time series are projected to the corresponding spatial feature dimension $[192, 384, 384, 192]$, and exchanged with the spatial features through a message-passing network consisting of a linear transformation, LayerNorm, GELU activation, and dropout (0.1), followed by a final linear projection.

The model was initialized either from the pretrained 24-h PanguWeather checkpoint or from random weights for the pretraining ablation. The original PanguWeather release provides the pretrained 24-hour model only in ONNX[31] format which supports the inference but not the training nor finetuning. We therefore reconstructed the model in PyTorch following the published model description and pseudocode, and transferred the ONNX parameters to the corresponding PyTorch tensors by matching the model components and parameter shapes. The reconstructed model was validated against the original ONNX file using ERA5 data. The implementation, parameter-mapping table, and validation results of 24-h PanguWeather are publicly available at https://github.com/zhaoshan2/pangu-pytorch. For the parameter-efficient setting, the pretrained input projection, downsampling and upsampling modules, together with the Earth-specific spatial attention blocks, were frozen. Newly introduced modules, including the temporal embedding, temporal-attention, TeleDiffGNN module were optimized using the base learning rate of $1e-3$. Remaining trainable pretrained parameters were updated using a learning rate reduced by a factor of ten. We used AdamW[32] as optimizer, and a learning rate scheduler with linear warm-up followed by cosine annealing. The learning rate was increased from 1% to 100% of the target learning rate during the warm-up phase and subsequently decayed using a cosine schedule to 10% of the initial learning rate. The scheduler was updated after every optimization step.

The model uses gridded atmospheric states with a set of OCIs as input. The atmospheric variables are taken from ERA5[33] in WeatherBench2[34] with a spatial resolution of 0.25°, while the OCIs are obtained from NOAA climate-index archives. We use 15 teleconnection and oceanic climate indices covering major modes of variability over the Pacific, Atlantic, Arctic, and Southern Hemisphere. These include ENSO-related indices as well as extratropical atmospheric circulation patterns.

Training data is from 1987 to 2017 and were sampled at a temporal interval of 720 hours, while 2019 and 2022 were used for validation and testing, respectively. Training samples were shuffled. The model takes four consecutive frames as input and predicts four future frames corresponding to forecast horizons of 168, 336, 504, and 672 hours.

5.5 Evaluation

The model is trained and selected using the ℓ_1 loss on the validation set. For deterministic forecast evaluation, we report latitude-weighted RMSE and latitude-weighted ACC defined in Bi et al.[9] and Nathaniel et al.[20] for global evaluation. Performance is compared against the subseasonal forecasting baselines reported in ChaosBench under the corresponding forecast horizons.

Because ChaosBench reports geopotential height rather than geopotential, we convert the predicted geopotential Z to geopotential height by dividing the standard gravitational acceleration[35] $9.80665\ \mathrm{m\,s^{-2}}$ when comparing the results with baselines in ChaosBench.

Author contributions. S.Z: Methodology, Data curation, Visualization, Investigation, Writing- Original draft preparation. Z.X: Conceptualization, Methodology, Supervision, Writing- Original draft preparation. X.X.Z: Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing, Funding acquisition.

Competing interests. The authors declare no competing interests.

Data availability. The ERA5 reanalysis data used in this study are publicly available from the Copernicus Climate Data Store and can also be accessed through WeatherBench2[34]. The ocean-climate indices are obtained from the NOAA Physical Sciences Laboratory Climate Indices database: https://psl.noaa.gov/data/climateindices/list/.

Code availability. The code will be made publicly available upon publication at https://github.com/zhu-xlab/weather-torch.

Acknowledgments. The work is jointly supported by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" (grant number: 50EE2201C), by the Excellence Strategy of the Federal Government and the Länder through the TUM Innovation Network EarthCare, by Munich Center for Machine Learning, and by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) based on a resolution of the German Bundestag (grant number: 67KI32002B; Acronym: EKAPEx), by the European Commission through the project "ThinkingEarth-Copernicus Foundation Models for a Thinking Earth" under the Horizon 2020 Research and Innovation program (Grant Agreement No.101130544), and by the Munich Center for Machine Learning.

References

[1] Shamir, E.: The value and skill of seasonal forecasts for water resources management in the upper santa cruz river basin, southern arizona. Journal of Arid Environments **137**, 35–45 (2017)

[2] Goddard, L., Aitchellouche, Y., Baethgen, W., Dettinger, M., Graham, R., Hayman, P., Kadi, M., Martínez, R., Meinke, H., *et al.*: Providing seasonal-to-interannual climate information for risk management and decision-making. Procedia Environmental Sciences **1**, 81–101 (2010)

[3] Robertson, A.W., Vitart, F., Camargo, S.J.: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. Journal of Geophysical Research: Atmospheres **125**(6), 2018–029375 (2020)

[4] Bauer, P., Thorpe, A., Brunet, G.: The quiet revolution of numerical weather prediction. Nature **525**(7567), 47–55 (2015)

[5] Wedi, N., Bauer, P., Denoninck, W., Diamantakis, M., Hamrud, M., Kuhnlein, C., Malardel, S., Mogensen, K., Mozdzynski, G., Smolarkiewicz, P.: The modelling infrastructure of the integrated forecasting system: Recent advances and future challenges (2015)

[6] Moosavi, A., Rao, V., Sandu, A.: Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. Journal of Computational Science **50**, 101295 (2021)

[7] Clyne, G., Couairon, G., Gastineau, G., Monteleoni, C., Charantonis, A.: Archesclimate: Probabilistic decadal ensemble generation with flow matching. arXiv preprint arXiv:2509.15942 (2025)

[8] Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., Anandkumar, A.: Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In: Proceedings of the Platform for Advanced Scientific Computing Conference, pp. 1–11 (2023)

[9] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3d neural networks. Nature **619**(7970), 533–538 (2023)

[10] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., *et al.*: Learning skillful medium-range global weather forecasting. Science **382**(6677), 1416–1421 (2023)

[11] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A.: Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343 (2023)

[12] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al.: Aurora: A foundation model of the atmosphere. arXiv preprint arXiv:2405.13063 **1**(8) (2024)

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). Ieee

[14] Vitart, F., Robertson, A.W., Anderson, D.L.: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. Bulletin of the World Meteorological Organization **61**(2), 23 (2012)

[15] Krishnamurthy, V.: Predictability of weather and climate. Earth and Space Science **6**(7), 1043–1056 (2019)

[16] Lorenz, E.N.: Atmospheric predictability as revealed by naturally occurring analogues. Journal of Atmospheric Sciences **26**(4), 636–646 (1969)

[17] Ashok, K., Behera, S.K., Rao, S.A., Weng, H., Yamagata, T.: El niño modoki and its possible teleconnection. Journal of Geophysical Research: Oceans **112**(C11) (2007)

[18] Domeisen, D.I., Garfinkel, C.I., Butler, A.H.: The teleconnection of el niño southern oscillation to the stratosphere. Reviews of Geophysics **57**(1), 5–47 (2019)

[19] Nguyen, T., Pham, T., Arcomano, T., Kotamarthi, R., Foster, I., Madireddy, S., Grover, A.: Omnicast: A masked latent diffusion model for weather forecasting across time scales. Advances in Neural Information Processing Systems **38**, 172182–172212 (2026)

[20] Nathaniel, J., Qu, Y., Nguyen, T., Yu, S., Busecke, J., Grover, A., Gentine, P.: Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. Advances in Neural Information Processing Systems **37**, 43715–43729 (2024)

[21] DOCUMENTATION, I.: Part v: The ensemble prediction

[22] Zhu, X.X., Xiong, Z., Wang, Y., Stewart, A.J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., Xu, Q., Shi, Y.: On the foundations of earth foundation models. Communications Earth & Environment **7**(1), 103 (2026)

[23] Leutbecher, M., Palmer, T.N.: Ensemble forecasting. Journal of computational physics **227**(7), 3515–3539 (2008)

[24] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021)

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[26] Seo, S., Meng, C., Liu, Y.: Physics-aware difference graph networks for sparsely-observed dynamics. In: International Conference on Learning Representations (2019)

[27] Cohen, J., Barlow, M.: The nao, the ao, and global warming: how closely related? Journal of Climate **18**(21), 4498–4513 (2005)

[28] Li, X., Cai, W., Meehl, G.A., Chen, D., Yuan, X., Raphael, M., Holland, D.M., Ding, Q., Fogt, R.L., Markle, B.R., *et al.*: Tropical teleconnection impacts on antarctic climate changes. Nature Reviews Earth & Environment **2**(10), 680–698 (2021)

[29] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt,

J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[30] Hickman, S.H., Trajković, I., Kaltenborn, J., Pelletier, F., Archibald, A., Gurwicz, Y., Nowack, P., Rolnick, D., Boussard, J.: Causal climate emulation with bayesian filtering. Advances in Neural Information Processing Systems **38**, 89418–89461 (2026)

[31] Bai, J., Lu, F., Zhang, K., et al.: ONNX: Open Neural Network Exchange. https://github.com/onnx/onnx (2019)

[32] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[33] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., *et al.*: The era5 global reanalysis. Quarterly journal of the royal meteorological society **146**(730), 1999–2049 (2020)

[34] Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., *et al.*: Weatherbench 2: A benchmark for the next generation of data-driven global weather models. Journal of Advances in Modeling Earth Systems **16**(6), 2023–004019 (2024)

[35] Moritz, H.: Geodetic reference system 1980. Journal of geodesy **74**(1), 128–133 (2000)

Table 3: Variables used in the model. We include both the atmospheric variables at surface and 13 pressure levels, as well as some large-ocean climate indices into the model input.

Index	Full name	Associated Area
AAO	Antarctic Oscillation	Southern Hemisphere annular circulation
AO	Arctic Oscillation	Northern Hemisphere annular circulation
EA	East Atlantic Pattern	North Atlantic circulation variability
EPO	Eastern Pacific Oscillation	Eastern North Pacific circulation and jet variability
NAO	North Atlantic Oscillation	North Atlantic pressure dipole and westerly circulation
Niño3	Niño3 SST Anomaly	Eastern tropical Pacific SST / ENSO variability
ONI	Oceanic Ni no Index	Central-eastern tropical Pacific ENSO variability
PDO	Pacific Decadal Oscillation	Decadal-scale North Pacific SST variability
PNA	Pacific–North American Pattern	North Pacific–North American circulation
SOI	Southern Oscillation Index	Atmospheric component of ENSO
TNA	Tropical Northern Atlantic Index	Northern tropical Atlantic SST variability
TNI	Trans-Ni no Index	Tropical Pacific zonal SST-gradient variability
TSA	Tropical Southern Atlantic Index	Southern tropical Atlantic SST variability
WHWP	Western Hemisphere Warm Pool	Tropical Atlantic–eastern Pacific warm-pool variability
WP	Western Pacific Pattern	Western North Pacific atmospheric circulation
MSL	Mean sea-level pressure	Surface
U10	10-m U wind component	Surface
V10	10-m V wind component	Surface
T2M	2-m temperature	Surface
Z	Geopotential	13 pressure levels
Q	Specific humidity	13 pressure levels
T	Temperature	13 pressure levels
U	U component of wind	13 pressure levels
V	V component of wind	13 pressure levels

A Data

Table 3 summarizes the input variables used in this study.

B Additional Figures

Additional subseasonal performance evaluation is provided in Fig 9, Fig 10, and Fig 11.

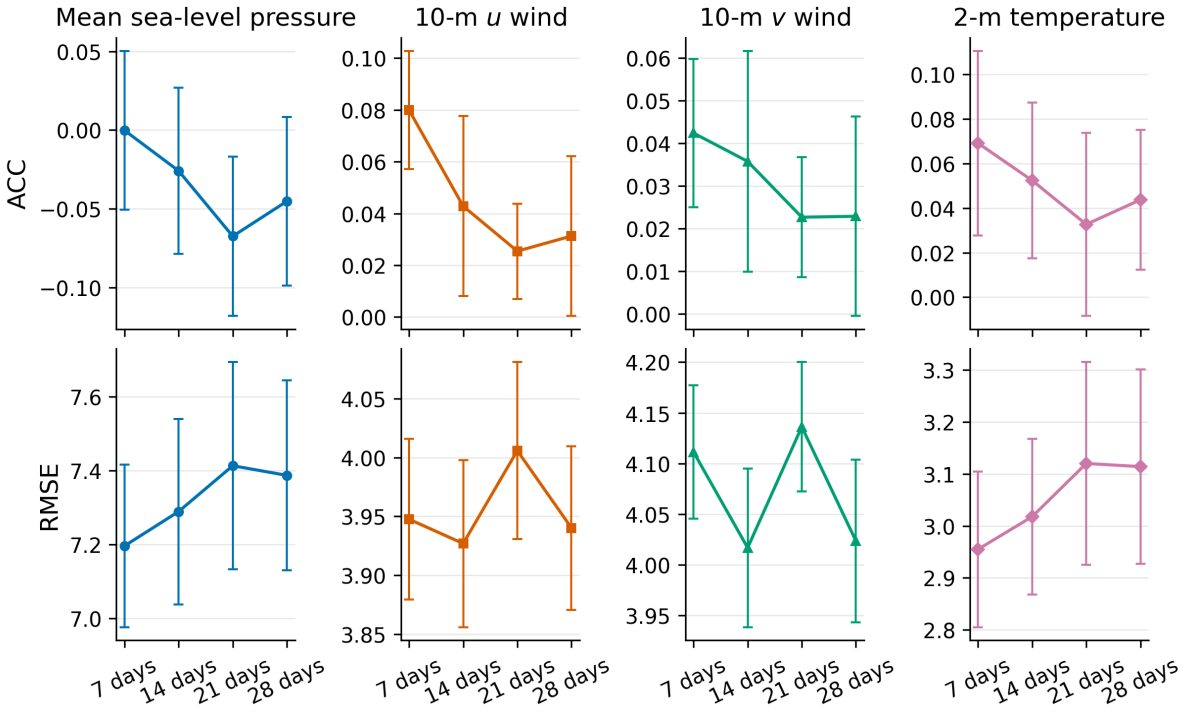


Fig. 9: ACC and RMSE of the surface variables across forecasting horizon up to four weeks. Markers indicate the mean performance at each forecast horizon, and error bars denote the variability across test samples

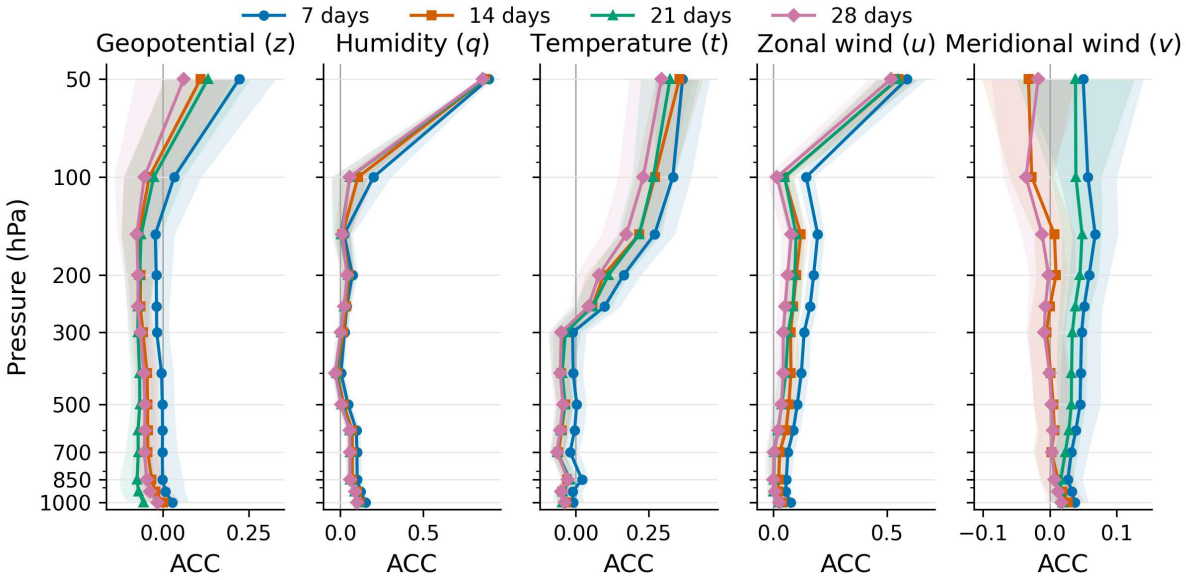


Fig. 10: ACC of the upper-level variables at forecasting horizon up to four weeks. Markers indicate the mean performance at each forecast horizon, and dashed area denotes the variability across test samples.

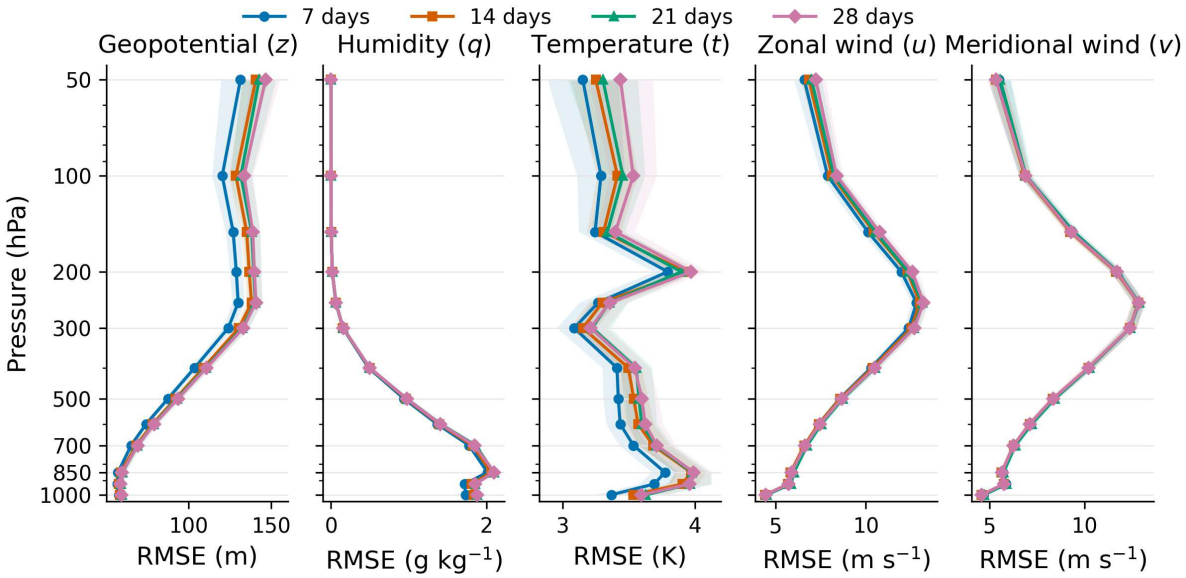


Fig. 11: RMSE of the upper-level variables at forecasting horizon up to four weeks. Markers indicate the mean performance at each forecast horizon, and dashed area denotes the variability across test samples.

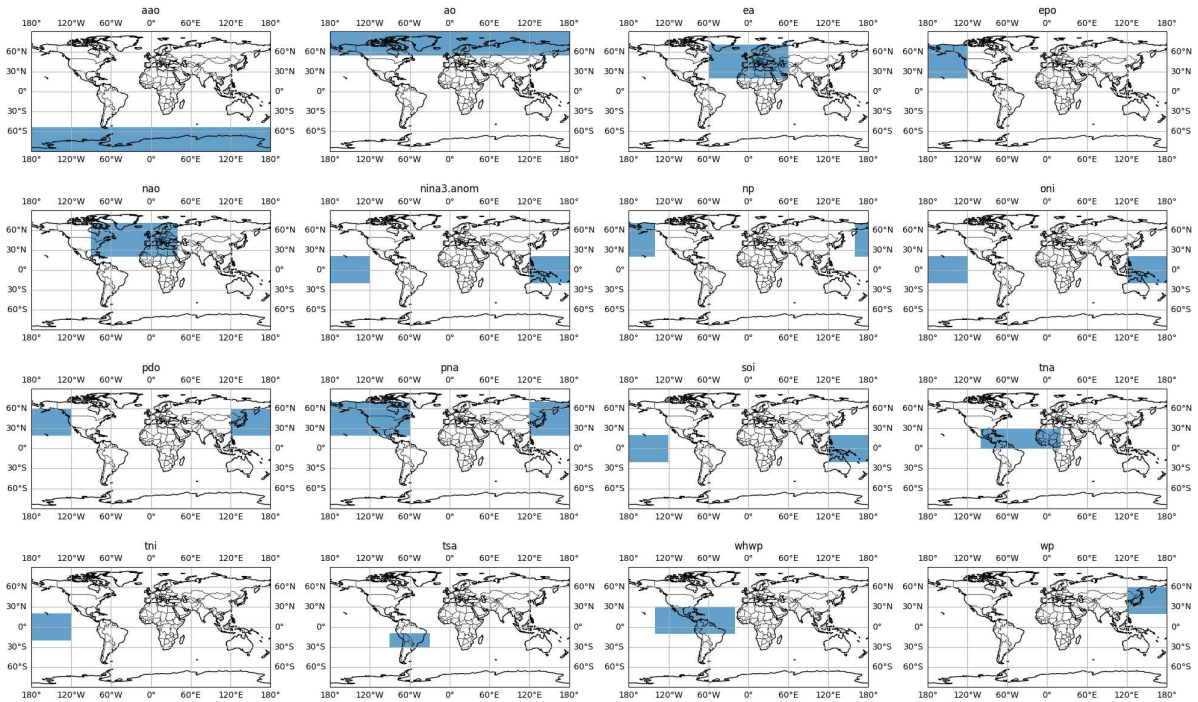


Fig. 12: The initial potential impact areas or the origination area (blue patches) of the teleconnection patters and ocean-climate indices. Due to the incomplete of the knowledge, we only use it as an initialization of the graph connectivity.