Appendix of STCast

Hao Chen[1] Tao Han[1] Jie Zhang[1*] Song Guo[1*] Lei Bai[2]
[1]Hong Kong University of Science and Technology (HKUST) [2]Shanghai AI Laboratory
[*]Corresponding author {hchener, thanad}@connect.ust.hk {songguo, csejzhang}@ust.hk

1. Dataset Details

1.1. Dataset and Baselines

In this work, we conduct experiments on a popular weather dataset, *i.e.*, ERA5[1] [11], provided by the ECMWF [17]. As shown in Tab. 1, ERA5 dataset is a reanalysis atmospheric dataset, consisting of the atmospheric variables from 1979 to the present day with a 0.25°spatial resolution with 721×1440. The atmospheric variables include 5 upper-air variables (Z, Q, U, V, T) on 13 levels and 5 surface variables (T2M, U10, U10, MSL, SP). The low-resolution global forecasting is trained on 70 variables ($5 \times 13 + 5$) with 40 years atmospheric dataset from 1979 to 2019 with 1.4°spatial resolution. At the same time, the high-resolution regional weather forecasting task is trained in the same period with 5 surface variables on Eastern Asia (7.5°W114°E-36°W172.5°E) with 0.25°spatial resolution.

1.2. Data processing

To address disparities among variables, all model inputs are normalized to ensure consistency. Using the training dataset spanning 1979–2019, we compute the mean and standard deviation for each variable. Normalization is then performed by subtracting the respective mean and dividing by the corresponding standard deviation.

1.3. Implementation Details

The main structure of this work follows the backbone of Flash Attention [5]. In the global forecasting stage, we train the whole model. While in the regional forecasting stage, we only need to train the Spatial-Aligned Attention(SAA) module and freeze the main structure. The detailed training parameters are provided in Tab. 2.

1.4. Code Available and Training Logs

We provide some code and training logs in the Supplementary Materials for our STCast, OneForecast, Graphcast and related baselines. Baselines are collected from its respective official GitHub repository.

2. Additional Related Works

2.1. Time-series Methods

Before the emergence of recent data-driven forecasting methods on large-scale atmospheric datasets such as ERA5 [11], several time-series approaches had already been applied to weather forecasting. These earlier works typically framed regional forecasting as a video prediction task, employing spatio-temporal convolutional layers or Transformer blocks to process the input data. For instance, SimVP [8] and PastNet [21] employed spatio-temporal convolutions to forecast regional atmospheric images, while STTN [12] and PKD-STTN [10] utilized spatio-temporal Transformer blocks. Concurrently, HiSTGNN [15] and STCWF [9] introduced spatio-temporal graph neural networks and contrastive learning, respectively, to the weather forecasting domain. Notably, PastNet [21] further distinguished itself by incorporating physical principles into its neural network architecture.

Although the aforementioned time-series forecasting methods are commonly referred to as spatio-temporal approaches utilizing spatio-temporal neural networks (NNs), our proposed STCast fundamentally differs from them in several key aspects:

[1] https://cds.climate.copernicus.eu/

Table 1. A summary of atmospheric variables. The 13 levels are 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa. 'Single' denotes the variables under earth's surface.

Name	Description	Levels	Resolution	Lat-Lon Range	Time
Low-resolution Global Weather Forecasting					
Z	Geopotential	13	128×256	-90°S180°W-90°N180°E	1979-2020
Q	Specific humidity	13	128×256	-90°S180°W-90°N180°E	1979-2020
U	x-direction wind	13	128×256	-90°S180°W-90°N180°E	1979-2020
V	y-direction wind	13	128×256	-90°S180°W-90°N180°E	1979-2020
T	Temperature	13	128×256	-90°S180°W-90°N180°E	1979-2020
t2m	Temperature at 2m height	Single	128×256	-90°S180°W-90°N180°E	1979-2020
u10	x-direction wind at 10m height	Single	128×256	-90°S180°W-90°N180°E	1979-2020
v10	y-direction wind at 10m height	Single	128×256	-90°S180°W-90°N180°E	1979-2020
msl	Mean sea-level pressure	Single	128×256	-90°S180°W-90°N180°E	1979-2020
sp	Surface pressure	Single	128×256	-90°S180°W-90°N180°E	1979-2020
High-resolution Regional Weather Forecasting					
t2m	Temperature at 2m height	Single	721×1440	7.5°S114°W-36°N172.5°E	1979-2020
u10	x-direction wind at 10m height	Single	721×1440	7.5°S114°W-36°N172.5°E	1979-2020
v10	y-direction wind at 10m height	Single	721×1440	7.5°S114°W-36°N172.5°E	1979-2020
msl	Mean sea-level pressure	Single	721×1440	7.5°S114°W-36°N172.5°E	1979-2020
sp	Surface pressure	Single	721×1440	7.5°S114°W-36°N172.5°E	1979-2020

Table 2. Implementation details on low-resolution global forecasts.

Category	**Parameter**	**Value**	**Parameter**	**Value**
Model Architecture	Input Size	128×256	Input Channels	70
	Output Channels	70	Number of Blocks	24
	Embedding Dimension	768	Patch Size	2
Training Configuration	Optimizer	AdamW	Learning Rate Scheduler	CosineAnnealingLR
	Initial Learning Rate	2×10^{-4}	Minimum Learning Rate	1×10^{-7}
	Maximum Epochs	100	Loss Function	L2loss
Data Settings	Input Time Steps	1	Training Output Time Steps	1
	Testing Output Time Steps	40	Training Period	[1979, 2020]
	Validation Period	[2021, 2021]	Grid Resolution	1.4°
Experimental Setup	Batch Size	1	Global Batch Size	16
	Number of Data Workers	20	Mixed Precision Training	Enabled
	World Size	16		

(1) Motivation: Previous works aim to capture implicit temporal correlations among time-series inputs and spatial dependencies within the input domain using neural components such as convolutional layers. In contrast, STCast explicitly models temporal correlations across monthly atmospheric patterns and geographical relationships across the global domain. **(2) Architecture:** Prior works typically model spatial and temporal dimensions through convolution layers or Transformer blocks, where spatial modeling is performed via convolution kernels applied to image grids, and temporal modeling is achieved by extending these kernels across time steps. In contrast, STCast introduces a novel spatio-temporal modeling framework based on monthly Gaussian distributions and global–regional representations, enabling more structured and interpretable learning across both spatial and temporal domains. **(3) Benchmarks:** Unlike previous time-series methods that are typically evaluated on small-scale, low-resolution datasets such as SEVIR [19] and WD [20], STCast is the first to explore explicit spatio-temporal correlations within a realistic Earth system. It is evaluated on the ERA5 dataset [11], demonstrating its scalability and effectiveness in large-scale global weather forecasting.

3. Model Details

As shown in Algorithm 1, the global weather forecasting framework comprises three core components: Encoder, Processor, and Decoder, and two supplementary settings: Reconstruction and Loss Function. In this work, the Encoder and Decoder implementations follow Flash-Attention [5], while the Reconstruction setting and Loss Function design adopt the same setting of VA-MoE [2]. For the forecasting task, the AI model Φ predicts future atmospheric states $\mathbf{X}^{t+1}$ from historical fields $\mathbf{X}^t$ as $\mathbf{X}^{t+1} = \Phi(\mathbf{X}^t)$. Detailed configurations for all five elements are provided below.

3.1. Encoder

Atmospheric variables across pressure levels are organized into a 3D tensor $\mathbf{X}^t \in \mathbb{R}^{H \times W \times N}$, where H and W denote the global grid height and width, respectively, and N is the number of variables. This tensor is projected into patch embeddings via a convolutional layer Conv with stride p equal to the patch size:

$$\mathbf{X}^t = \text{Conv}_{p \times p}(\mathbf{X}^t), \tag{1}$$

where we set $p = 2$.

In addition to patch embedding, we incorporate a learnable positional embedding matrix, denoted as $\mathbf{P}$, to encode spatial information within the Encoder module. This matrix is initialized using a truncated normal distribution defined as:

$$f(x; \mu, \sigma, a, b) = \begin{cases} \frac{\phi(\frac{x-\mu}{\sigma})}{\sigma\left(\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right)} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where μ and σ represent the mean and standard deviation, and a and b denote the lower and upper bounds of the distribution, respectively. The positional embedding $\mathbf{P}$ is subsequently added to the input features prior to processing, yielding the updated representation $\mathbf{X}^t = \mathbf{X}^t + \mathbf{P}$.

3.2. Processor

The processor consists of a sequence of Transformer blocks, each comprising multi-head attention, a TMoE (Temporal Mixture-of-Experts) module, layer normalization, and residual connections. The operations within a single block can be formally expressed as:

$$\mathbf{A}^t = \text{LN}(\text{Attention}(\mathbf{X}^t)) + \mathbf{X}^t, \tag{3}$$

$$\mathbf{X}^{t+1} = \text{LN}(\text{TMoE}(\mathbf{A}^t)) + \mathbf{A}^t, \tag{4}$$

where Attention denotes the attention mechanism, LN represents layer normalization, and TMoE refers to the Temporal Mixture-of-Experts module as described in the main paper.

Following the design principles of FlashAttention [5] and VA-MoE [2], we adopt an alternating strategy that combines window-based attention and global self-attention. This hybrid approach enables the model to effectively capture both local and global dependencies in the input distribution.

3.3. Decoder

The Decoder module in this work is implemented as a multi-layer perceptron (MLP), which predicts the atmospheric variables $\mathbf{X}^{t+1}$ for the next timestep. The decoding operation is defined as:

$$\mathbf{X}^{t+1} = \text{MLP}(\mathbf{X}^{t+1}). \tag{5}$$

The MLP consists of two linear layers separated by a non-linear activation function. Specifically, the decoding process can be expressed as:

$$\mathbf{X}^{t+1} = \text{Linear}(\text{GELU}(\text{Linear}(\mathbf{X}^{t+1}))), \tag{6}$$

where GELU denotes the Gaussian Error Linear Unit activation function. This structure enables the decoder to model complex relationships in the input features and generate accurate predictions for the subsequent timestep.

Algorithm 1 STCast for Global Weather Forecasting

Input: Atmospheric variables $\mathbf{X}^t$ at timestep t
Output: Forecasted atmospheric variables $\mathbf{X}^{t+1}$ at timestep $t+1$

Encoder
Apply high-stride convolution for patch embedding: $\mathbf{X}^t = \text{Conv}_{p\times p}(\mathbf{X}^t)$
Add positional embedding: $\mathbf{X}^t = \mathbf{X}^t + \mathbf{P}$
Project to latent space via MLP: $\mathbf{X}^t = \text{MLP}(\mathbf{X}^t)$

Processor
for each $i \in$ [blocks' number] **do**
  if $i\%2 == 0$ **then**
    $j = i//2$
    if $j\%3 == 0$ **then**
      Windows Size = (8, 8)
    else if $j\%3 == 1$ **then**
      Windows Size = (4, 16)
    else
      Windows Size = (16, 4)
    end if
    Attention = Windows-Attention(Windows Size)
  else
    Attention = Self-Attention
  end if

  Apply multi-head attention with residual connection: $\mathbf{A}^t = \text{LN}(\text{Attention}(\mathbf{X}^t)) + \mathbf{X}^t$
  Apply TMoE with residual connection: $\mathbf{X}^{t+1} = \text{LN}(\text{TMoE}(\mathbf{A}^t)) + \mathbf{A}^t$
end for

Decoder
Reconstruct atmospheric variables to longitude-latitude grids: $\mathbf{X}^{t+1} = \text{MLP}(\mathbf{X}^{t+1})$

3.4. Reconstruction

To ensure training stability, we introduce an auxiliary reconstruction path that directly connects the Encoder and Decoder modules to reconstruct the input variables. This design complements the primary prediction path, which consists of the Encoder, Processor, and Decoder modules. Notably, the Encoder and Decoder are shared across both paths. The overall process is defined as:

$$\hat{\mathbf{X}}^t = \text{Decoder}(\text{Encoder}(\mathbf{X}^t)), \tag{7}$$

$$\hat{\mathbf{X}}^{t+1} = \text{Decoder}(\text{Processor}(\text{Encoder}(\mathbf{X}^t))), \tag{8}$$

where Encoder, Processor, and Decoder denote the respective modules in our framework. The reconstruction path facilitates the learning of robust representations by encouraging the model to preserve essential input information throughout the encoding and decoding stages.

Under this configuration, the Encoder and Decoder modules are dedicated to encoding and decoding the input variables, respectively, while the Processor is solely responsible for prediction. By excluding the Processor from the encoding and decoding stages, the framework avoids unnecessary computational overhead, thereby enhancing efficiency without compromising performance.

3.5. Loss function

To address both prediction and reconstruction tasks, we employ the L2 loss function to quantify point-wise errors between the predicted outputs and the ground truth. The prediction loss Obj_{pred} and reconstruction loss Obj_{recon} are defined as follows:

$$Obj_{pred} = \text{Mean}((\hat{\mathbf{X}}^{t+1} - \mathbf{X}^{t+1})^2), \tag{9}$$

$$Obj_{recon} = \text{Mean}((\hat{\mathbf{X}}^{t} - \mathbf{X}^{t})^2), \tag{10}$$

$$Obj_{final} = Obj_{pred} + \lambda * Obj_{recon}, \tag{11}$$

where $\hat{\mathbf{X}}^t$ and $\hat{\mathbf{X}}^{t+1}$ denote the reconstructed and predicted outputs, respectively. The operator Mean computes the average error across multiple dimensions. A weighting hyperparameter λ is introduced to balance the two objectives in the final loss function.

4. Experiments Details

4.1. Typhoon Tracking

This study focuses on two extreme cyclones: Typhoon Ewiniar and Typhoon Yinxing. Typhoon Ewiniar formed east of Mindanao on May 24, 2024, traversed the Philippine Sea, and recurved northeastward over the Okinawa–Ryukyu region. Typhoon Yinxing developed east of Yap Island on November 4, 2024, crossed the Philippine Sea, and entered the South China Sea. The initial conditions for these cyclones are set at 00:00 UTC on May 24, 2024, and 00:00 UTC on November 4, 2024, respectively. Ground-truth and ECMWF are obtained from TCData[2], and others are computed by us with official codes.

Following previous AI-based approaches [1, 16], we identify the eye of a tropical cyclone as the location of the local minimum in mean sea level pressure (MSLP). The typhoon track detection algorithm is a sophisticated numerical weather prediction post-processing tool specifically designed to automatically identify and track tropical cyclone paths from high-resolution meteorological model outputs. The algorithm employs a multi-constraint sequential tracking approach that combines sea-level pressure minimization with comprehensive physical validation to ensure robust and physically consistent trajectory estimation.

The specific steps are as followed. 1) Set current position to initial coordinates; 2) Extract latitude-longitude grids and atmospheric variables, including MSL, U10, and V10; 3) Locate the typhoon center by finding minimum sea-level pressure within a defined search radius, 278km; 4) Calculate the center pressure, wind speed, and moving distance; 5) Check the termination conditions by pressure threshold (101200 Pa), wind speed threshold (10.2 m/s), and max distance threshold (400 km); 6) If all validation criteria are satisfied, update current position and append to trajectory.

4.2. Evaluation Metric

In this work, we evaluate the forecasting performance between our STCast and other methods on RMSE (Root Mean Square Error) and ACC (Anomalous Correlation Coefficient), which can be defined as:

$$\text{RMSE}(t) = \sqrt{\frac{\sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} L_i (\hat{X}_{i,j}^t - X_{i,j}^t)^2}{N_{lat} \times N_{lon}}}, \tag{12}$$

$$\text{ACC}(t) = \sqrt{\frac{\sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} L_i \hat{X}_{i,j}^t X_{i,j}^t}{\sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} L_i (\hat{X}_{i,j}^t)^2 \times \sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} L_i (X_{i,j}^t)^2}}, \tag{13}$$

where $\hat{X}_{i,j}^t$ and $X_{i,j}^t$ denotes the predicted variables and ground-truth at the horizontal coordinate (i, j) and time t; N_{lat} and N_{lon} denote the length of latitude and longitude in the global region.

Considering the difference in the distribution of atmospheric variables at latitudes, we introduce the latitude-dependent function L_i to weight the atmospheric variables. The function is formulated as:

$$L_i = N_{lat} \times \frac{\cos\phi_i}{\sum_{j=1}^{N_{lat}} \cos\phi_j}, \tag{14}$$

[2] tcdata.typhoon.org.cn

Table 3. Comparative Analysis of Training Times and Hardware Specifications for Deep Learning Models. $*$ is trained by ourselves. Some data is collected from KARINA [4]

Model	Params(M)	MACs(G)	GPUs	Training Time
Fengwu [3]	153.49	132.83	32 A100	17 days
FourCastNet [13]	-	-	64 A100	16 hrs
Graphcast [14]	28.95	1639.26	32 TPUv4	4 weeks
Pangu-Weather [1]	23.83	142.39	192 V100	64 days
VA-MoE [2]	665.37	-	32 A100	6 days
OneForecast [7]	24.76	509.27	16 A100*	8 days*
Ours (STCast)	654.82	436.12	16 A100	5 days

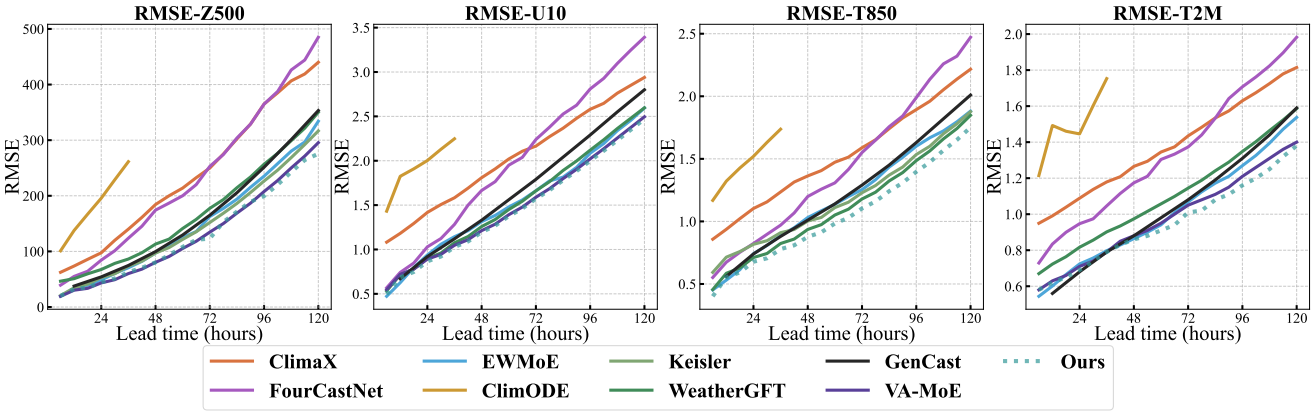


Figure 1. 120-hour comparative analysis of **RMSE** ↓ across 10 data-driven models for four variables, including Z500, T850, T2M, and U10. Results are collected from EWMoE [6], WeatherGFT [22] and WeatherBench [18] in https://sites.research.google/gr/weatherbench/deterministic-scores.

where ϕ_i and ϕ_j denote the latitude at index i and j, respectively.

For typhoon track prediction, we evaluate model performance using two metrics: Mean Distance Error (MDE) and Haversine Distance. First, the Haversine Distance is computed between the predicted typhoon center and the ground-truth location to account for the curvature of the Earth. Subsequently, the MDE is used to quantify the average positional error across all time steps. These metrics are defined as follows:

$$\text{MDE} = \frac{1}{N}\sum_{i=1}^{N} d(P_{\text{pred}}, P_{\text{obs}}), \tag{15}$$

$$d(P_1, P_2) = 2R \cdot \arcsin(\sqrt{a}), \tag{16}$$

$$a = \sin^2(\frac{\Delta\phi}{2}) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right), \tag{17}$$

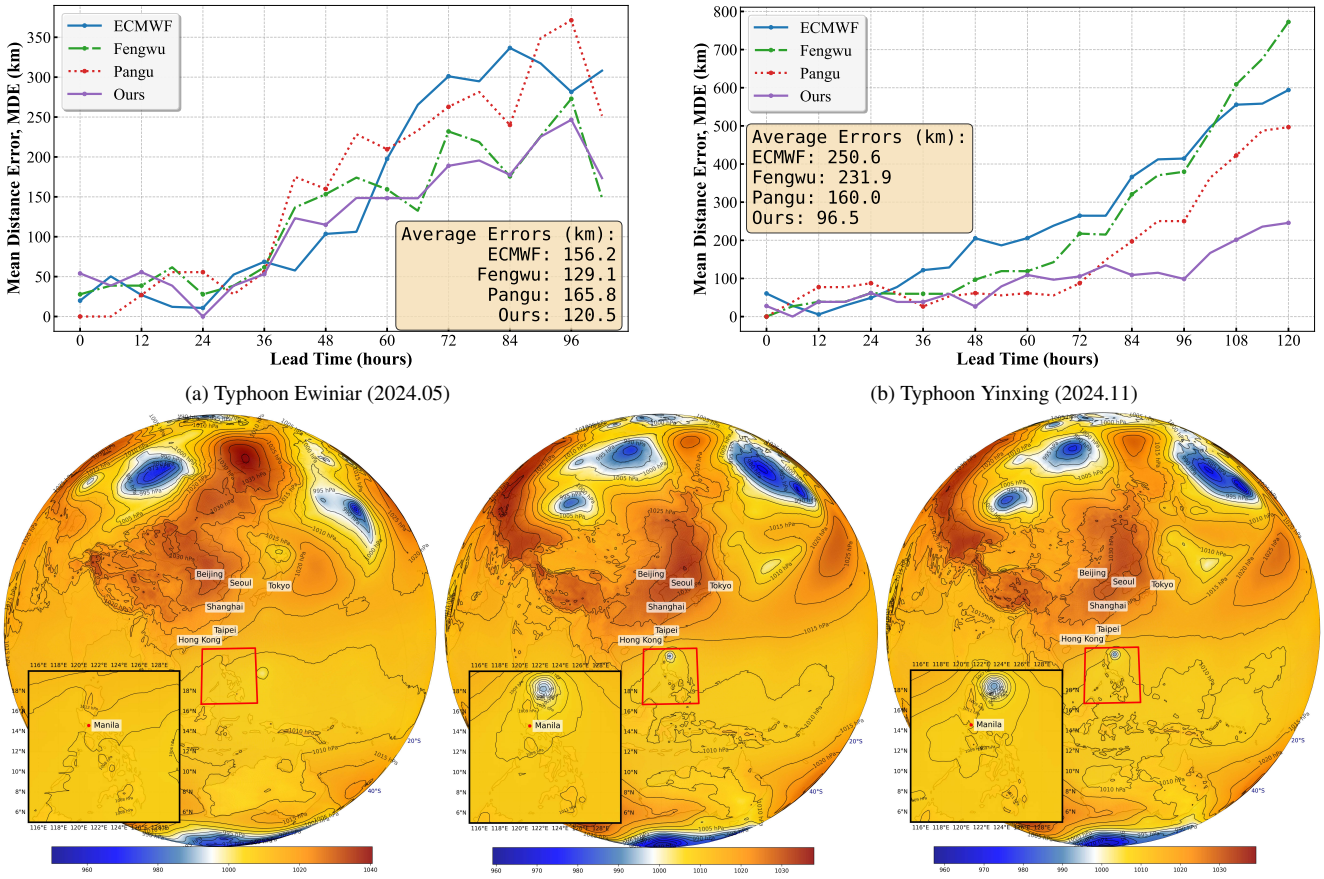
$$\Delta\phi = \phi_2 - \phi_1, \tag{18}$$

$$\Delta\lambda = \lambda_2 - \lambda_1, \tag{19}$$

where $R = 6371$km is the average radius of the earth, ϕ_1, λ_1 and ϕ_2, λ_2 are the latitude and longitude of two points in earth system. P_{pred} and P_{obs} are the latitude-longitude coordinates of predicted and real points.

Table 4. Ablation Studies on the processor's structure with normalized mean RMSE.

Low-resolution Global Forecasts (RMSE↓)					
Processor Structure	6-hour	1-day	4-day	7-day	10-day
Self-Attn	0.0814	0.1363	0.2797	0.4576	0.5998
Self-Attn + Window-Attn (Fixed window size)	0.0811	0.1348	0.2705	0.4424	0.5838
Self-Attn + Window-Attn (Various window size)	0.0617	0.1197	0.2578	0.4348	0.5763



(a) Typhoon Ewiniar (2024.05)

(b) Typhoon Yinxing (2024.11)

(c) Initial MSL
2024.11.04, T+0

(d) Predicted MSL
2024.11.06, T+12

(e) Real MSL
2024.11.06, T+12

Figure 2. Typhoon Track Assessment. (a) and (b) present a five-day comparative analysis of Mean Distance Error (MDE, in kilometers ↓) for Typhoons Ewiniar and Yinxing, respectively. (c), (d), and (e) visualize the evolution of Mean Sea-Level Pressure (MSL) across three temporal stages: initial conditions, 60-hour predictions, and corresponding ground-truth observations.

5. Additional Results

5.1. Efficiency Analysis

As shown in Tab. 3, we compare the number of parameters, multiply–accumulate operations (MACs), GPU usage, and training duration across six baseline models. Although STCast contains more parameters and MACs than GNN-based methods such as Graphcast and OneForecast, its overall computational cost remains significantly lower than that of previous models, particularly Fengwu, Pangu-Weather, and Graphcast. These results demonstrate that STCast achieves superior performance while maintaining comparable computational efficiency.

5.2. Additional Global Forecasting Analysis

As illustrated in Fig. 1, we compare STCast with several state-of-the-art models across forecasting horizons ranging from 6 to 120 hours, evaluated on four key atmospheric variables. Experimental results show that STCast performs comparably to VA-MoE and Stormer in predicting Z500 and U10, while outperforming all baselines in T850 and T2M. These findings highlight the effectiveness of STCast and its integrated TMoE architecture in global weather forecasting, demonstrating its ability to capture both temporal correlations and seasonal variability across diverse meteorological variables.

5.3. Additional Typhoon Track Prediction

In addition to the typhoon track visualizations presented in the main paper, we conduct a quantitative evaluation of forecasting performance using Mean Distance Error (MDE, in kilometers) across four methods: ECMWF [17], FengWu [3], Pangu-Weather [1], and our STCast. As illustrated in Figure 2a-b, experimental results demonstrate that STCast achieves competitive accuracy in short-term forecasts while exhibiting superior performance in long-term predictions, particularly for Typhoon Yinxing, with mean errors of 96.5 km and 128.5 km for Typhoons Yinxing and Ewiniar respectively—the lowest among all compared algorithms. Complementary visualizations in Figure 2c-e further dissect Typhoon Yinxing's evolutionary process by contrasting initial conditions, STCast predictions, and ground-truth observations. Collectively, these findings substantiate STCast's robust capability in extreme weather event assessment, demonstrating notable advantages in accurately forecasting tropical cyclone trajectories over extended temporal horizons.

5.4. Additional Ablation Study

To further assess the impact of the Processor's architectural design, we conduct an ablation study examining three attention configurations: 1) full self-attention, 2) self-attention + fixed-size window attention, and 3) self-attention + various-size window attention. Compared to configuration 1, configuration 2 demonstrates substantial improvements in long-term forecasting performance while yielding limited gains in short-term predictions. Configuration 3 achieves an additional average improvement of approximately 0.02 across all evaluation metrics, underscoring the efficacy of the proposed processor design. Through this adaptive attention mechanism, STCast effectively captures both global patterns and regional characteristics across different geographical areas, thereby enhancing its capacity to learn complex atmospheric dynamics across Earth's surface.

5.5. Additional Visualization

We provide more visualization about regional weather forecasting of 6-hour, 0.5-day, 1-day, 1.5-day, 2-day, 2.5-day, 3-day, 3.5-day, 4-day, 4.5-day, 5-day, 5.5-day, 6-day, 6.5-day, 7-day, 7.5-day, 8-day, 8.5-day, 9-day, 9.5-day, and 10-day in Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17, Fig. 18, Fig. 19, Fig. 20, Fig. 21, Fig. 22, and Fig. 23.

We also provide more visualization about global weather forecasting of 6-hour, 0.5-day, 1-day, 1.5-day, 2-day, 2.5-day, 3-day, 3.5-day, 4-day, 4.5-day, 5-day, 5.5-day, 6-day, 6.5-day, 7-day, 7.5-day, 8-day, 8.5-day, 9-day, 9.5-day, and 10-day in Fig. 24, Fig. 25, Fig. 26, Fig. 27, Fig. 28, Fig. 29, Fig. 30, Fig. 31, Fig. 32, Fig. 33, Fig. 34, Fig. 35, Fig. 36, Fig. 37, Fig. 38, Fig. 39, Fig. 40, Fig. 41, Fig. 42, Fig. 43, and Fig. 44.

References

[1] 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(7970):533–538, 2023. 5, 6, 8

[2] Hao Chen, Han Tao, Guo Song, Jie Zhang, Yunlong Yu, Yonghan Dong, and Lei Bai. Va-moe: Variables-adaptive mixture of experts for incremental weather forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3, 6

[3] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023. 6, 8

[4] Minjong Cheon, Yo-Hwan Choi, Seon-Yu Kang, Yumi Choi, Jeong-Gil Lee, and Daehyun Kang. Karina: An efficient deep learning model for global weather forecast. *arXiv preprint arXiv:2403.10555*, 2024. 6

[5] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, pages 16344–16359, 2022. 1, 3

[6] Lihao Gan, Xin Man, Chenghong Zhang, and Jie Shao. Ewmoe: An effective model for global weather forecasting with mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 210–218, 2025. 6

[7] Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, et al. Oneforecast: A universal framework for global and regional weather forecasting. In *Proceedings of the 42th International Conference on Machine Learning*, 2025. 6

[8] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180, 2022. 1

[9] Yongshun Gong, Tiantian He, Meng Chen, Bin Wang, Liqiang Nie, and Yilong Yin. Spatio-temporal enhanced contrastive and contextual learning for weather forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):4260–4274, 2024. 1

[10] Jing He, Junzhong Ji, and Minglong Lei. Spatio-temporal transformer network with physical knowledge distillation for weather forecasting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 819–828, 2024. 1

[11] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. 1, 2

[12] Junzhong Ji, Jing He, Minglong Lei, Muhua Wang, and Wei Tang. Spatio-temporal transformer network for weather forecasting. *IEEE Transactions on Big Data*, 11(2):372–387, 2024. 1

[13] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, 2023. 6

[14] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. 6

[15] Minbo Ma, Peng Xie, Fei Teng, Bin Wang, Shenggong Ji, Junbo Zhang, and Tianrui Li. Histgnn: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648:119580, 2023. 1

[16] Linus Magnusson, Sharanya Majumdar, Rebecca Emerton, David Richardson, Magdalena Alonso-Balmaseda, Calum Baugh, Peter Bechtold, Jean Bidlot, Antonino Bonanni, Massimo Bonavita, et al. Tropical cyclone activities at ecmwf. *ECMWF Technical Memoranda*, 2021. 5

[17] Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529):73–119, 1996. 1, 8

[18] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, 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):e2023MS004019, 2024. 6

[19] Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, pages 22009–22019, 2020. 2

[20] Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, and Guangquan Zhang. Deep uncertainty quantification: A machine learning approach for weather forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2087–2095, 2019. 2

[21] Hao Wu, Fan Xu, Chong Chen, Xian-Sheng Hua, Xiao Luo, and Haixin Wang. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 2917–2926, 2024. 1

[22] Wanghan Xu, Fenghua Ling, Wenlong Zhang, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Generalizing weather forecast to fine-grained temporal scales via physics-ai hybrid modeling. In *Advances in Neural Information Processing Systems*, 2024. 6

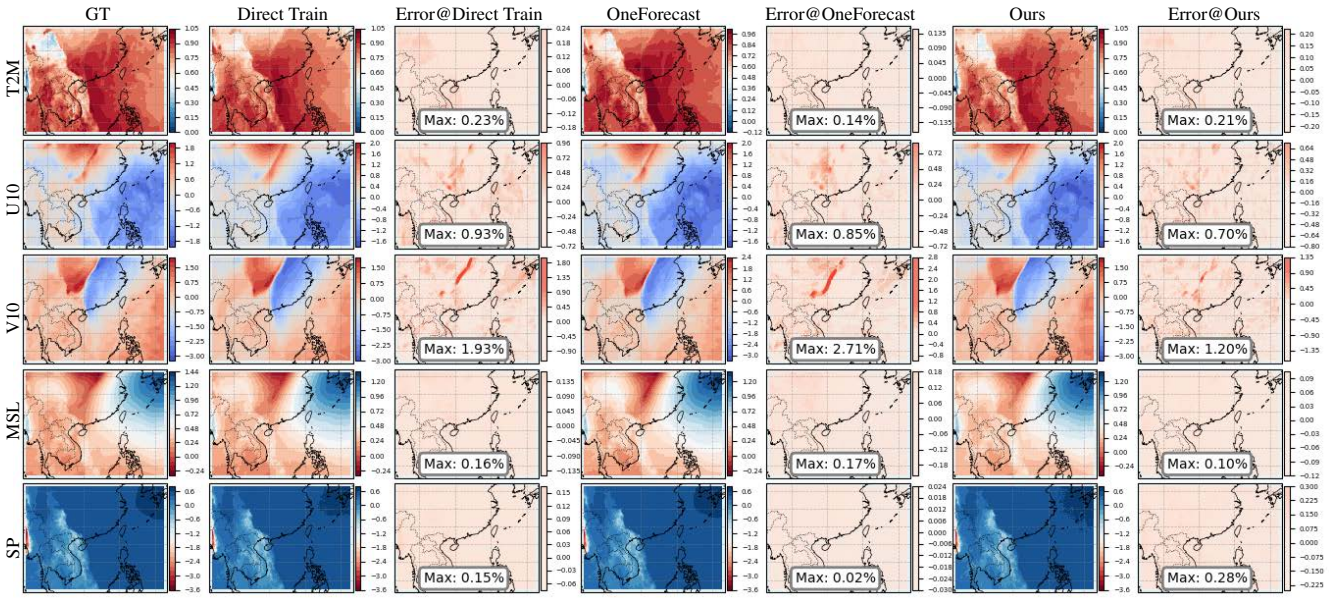


Figure 3. 6-hour forecast results of regional weather among different models.

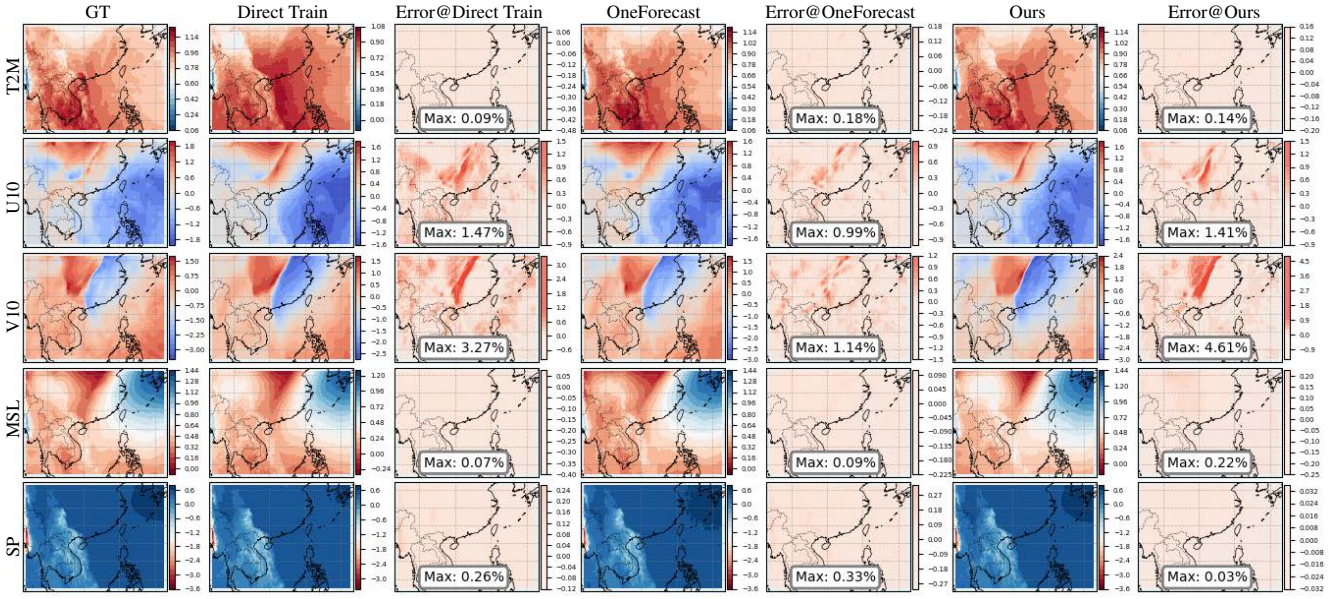


Figure 4. 0.5-day forecast results of regional weather among different models.

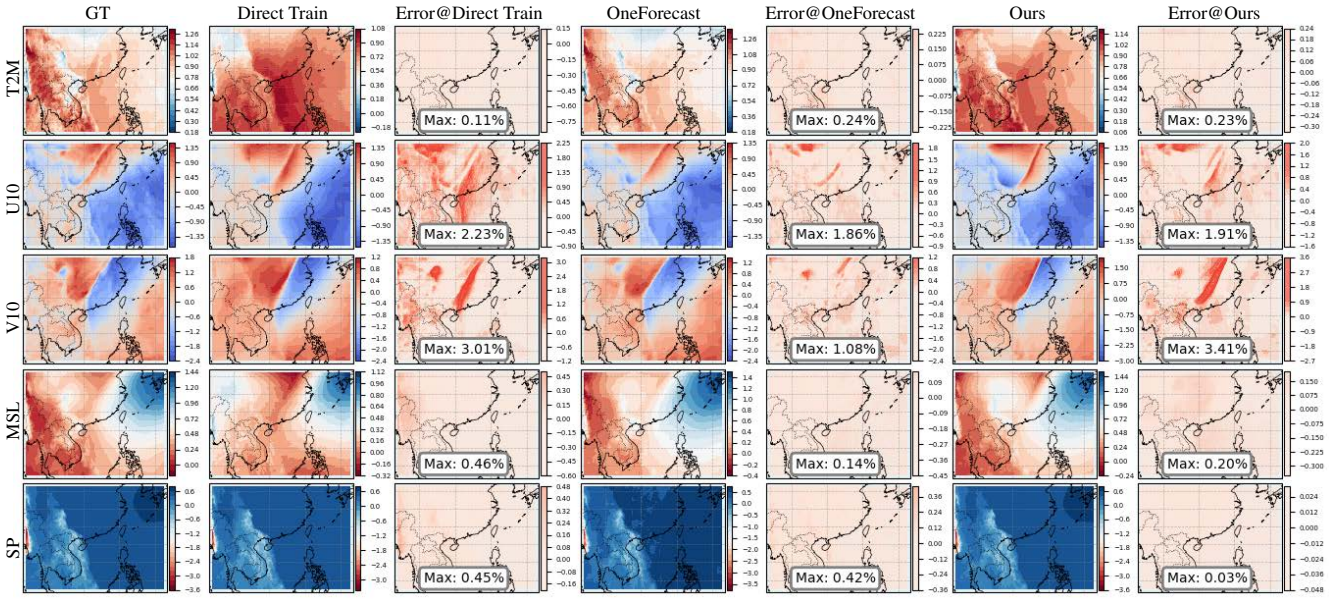


Figure 5. 1-day forecast results of regional weather among different models.

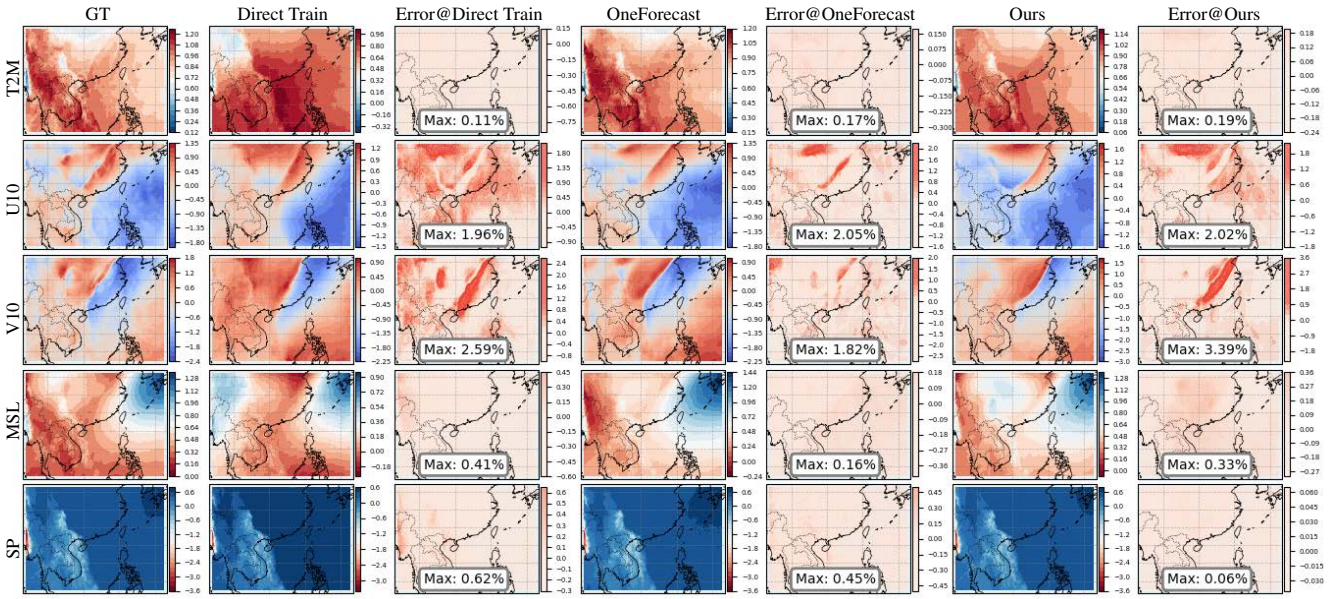


Figure 6. 1.5-day forecast results of regional weather among different models.

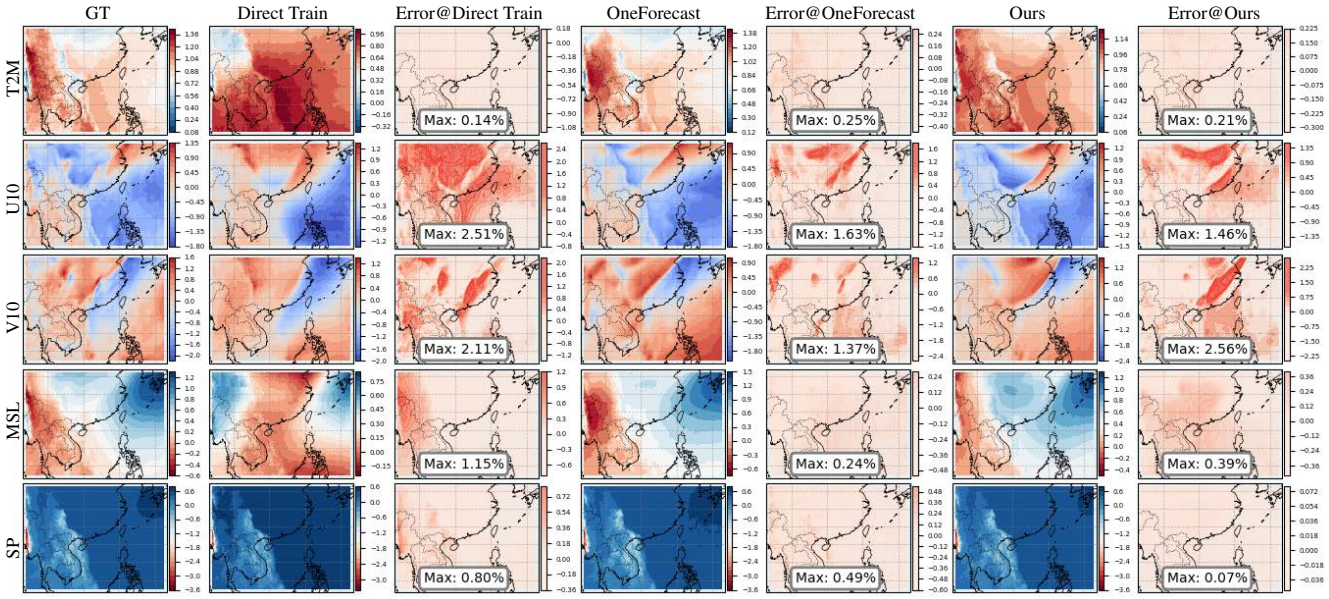


Figure 7. 2-day forecast results of regional weather among different models.

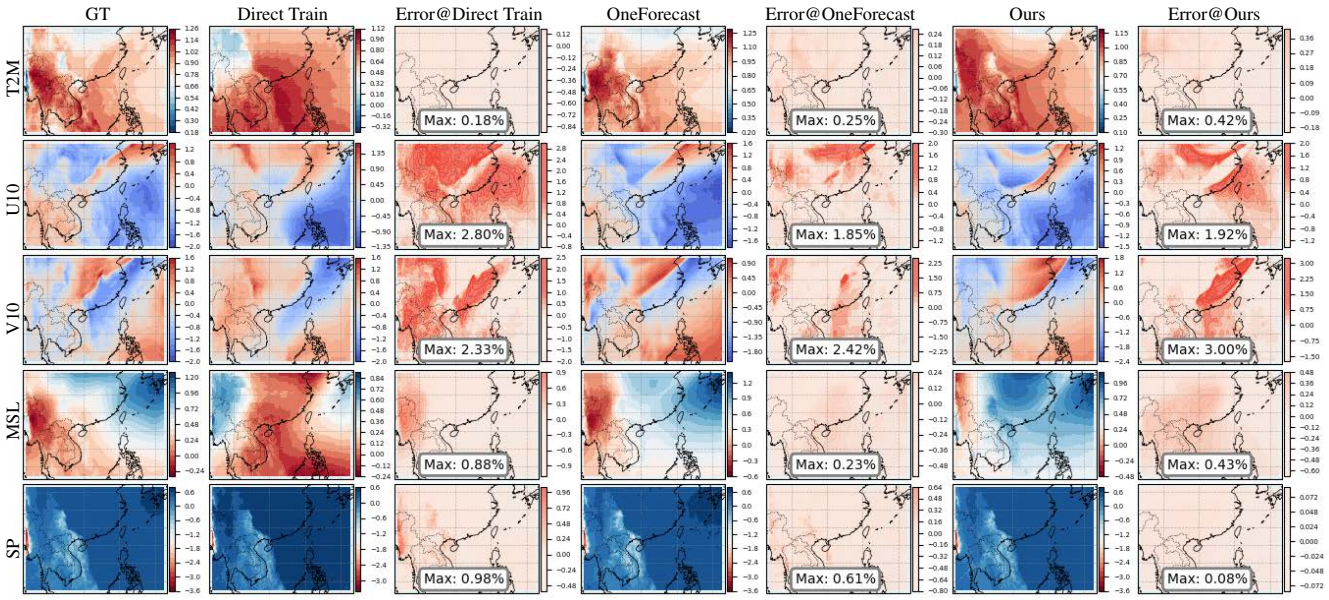


Figure 8. 2.5-day forecast results of regional weather among different models.

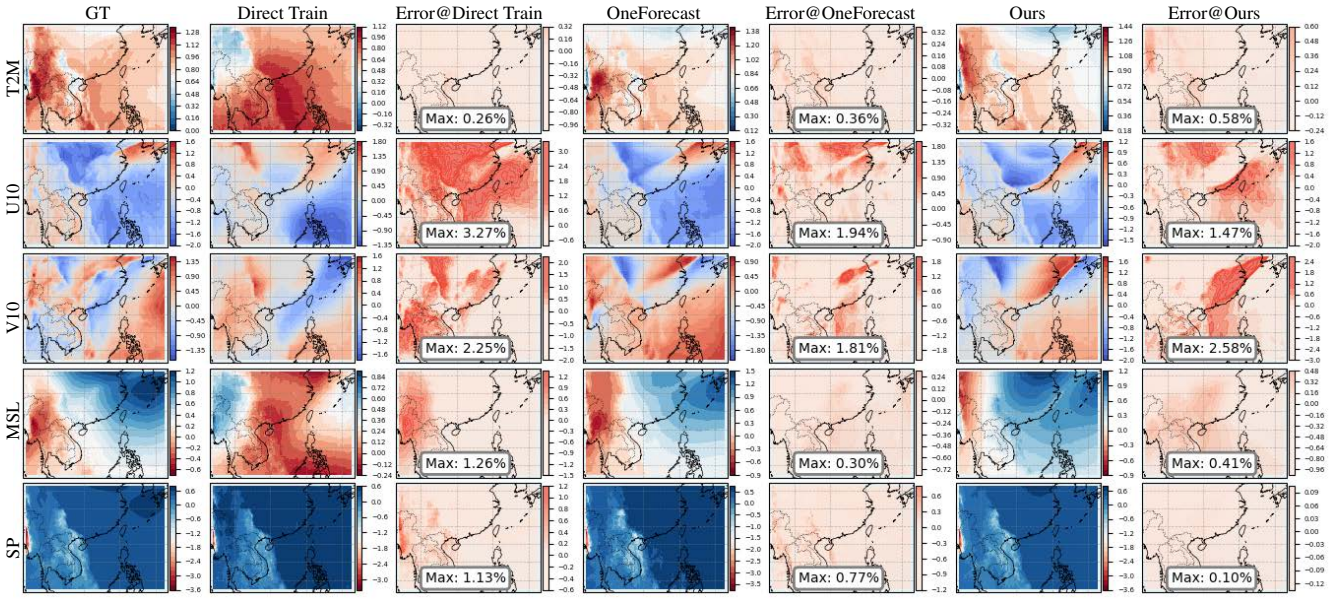


Figure 9. 3-day forecast results of regional weather among different models.

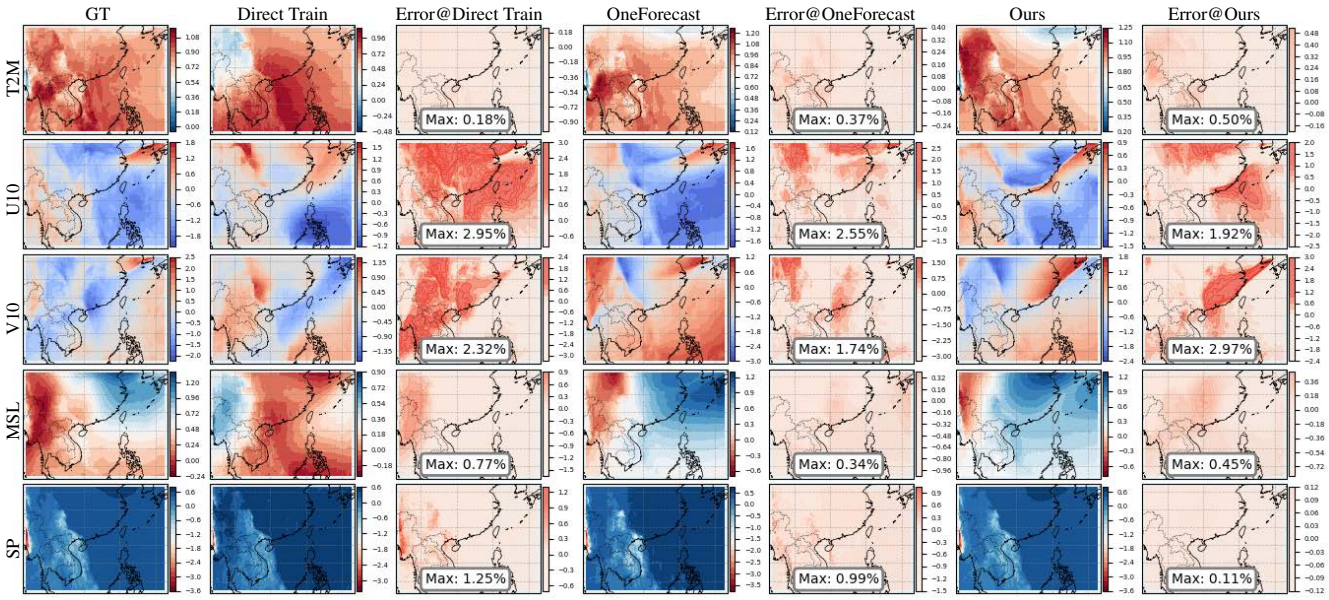


Figure 10. 3.5-day forecast results of regional weather among different models.

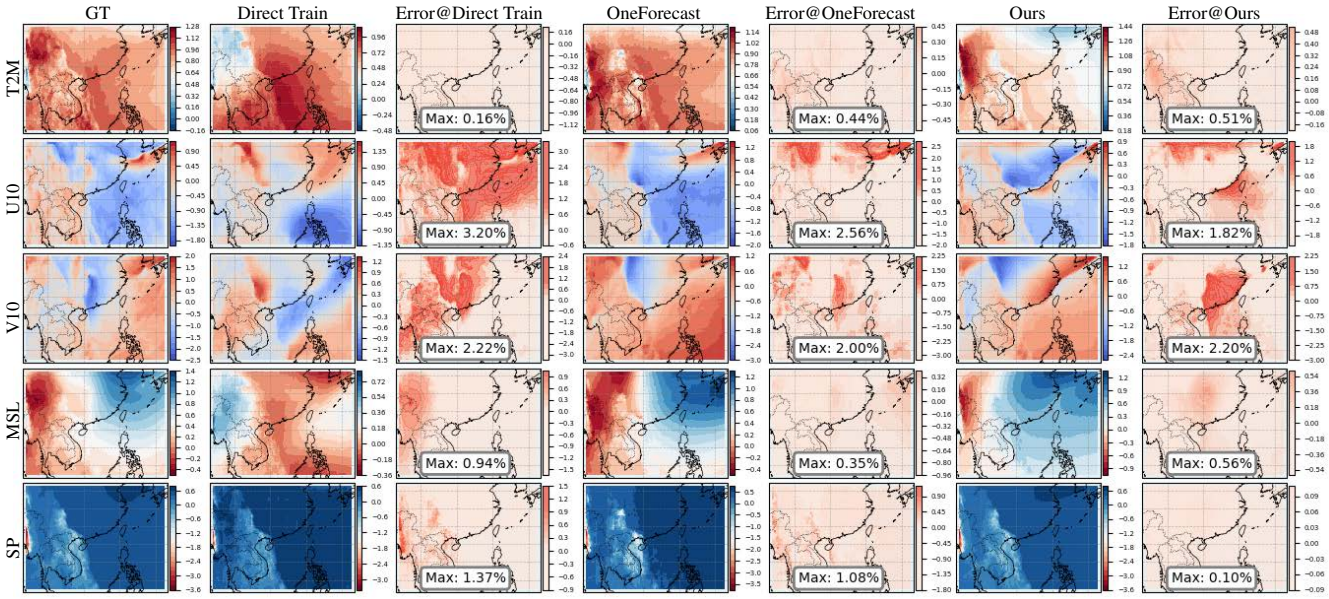


Figure 11. 4-day forecast results of regional weather among different models.

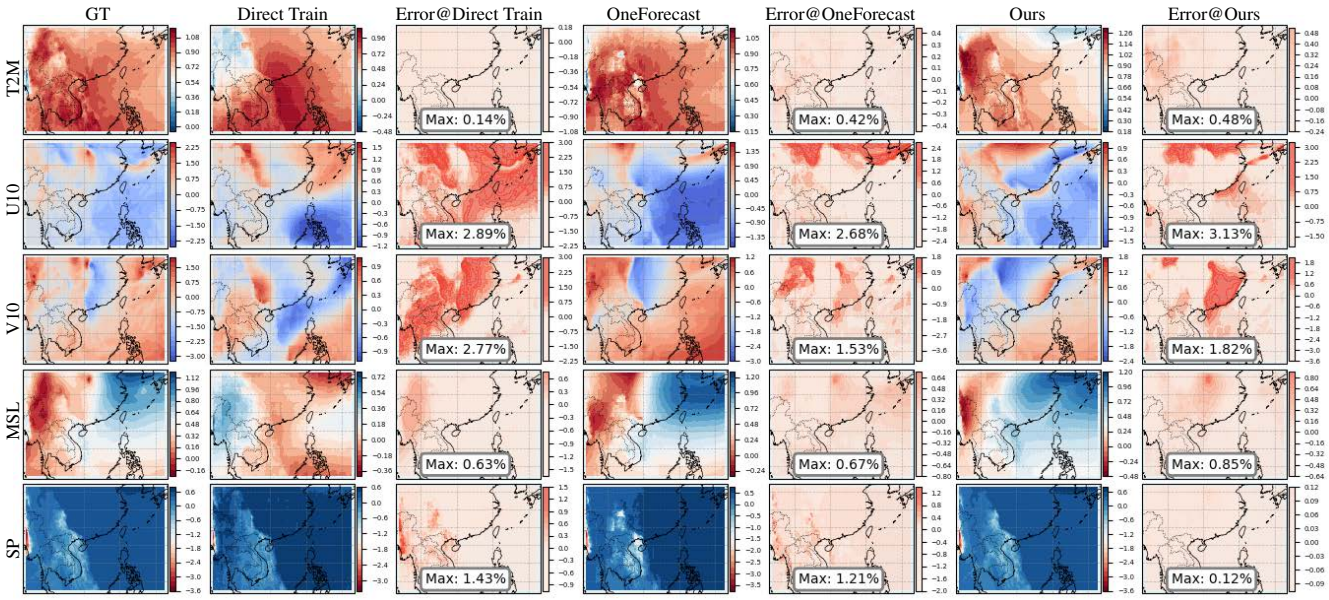


Figure 12. 4.5-day forecast results of regional weather among different models.

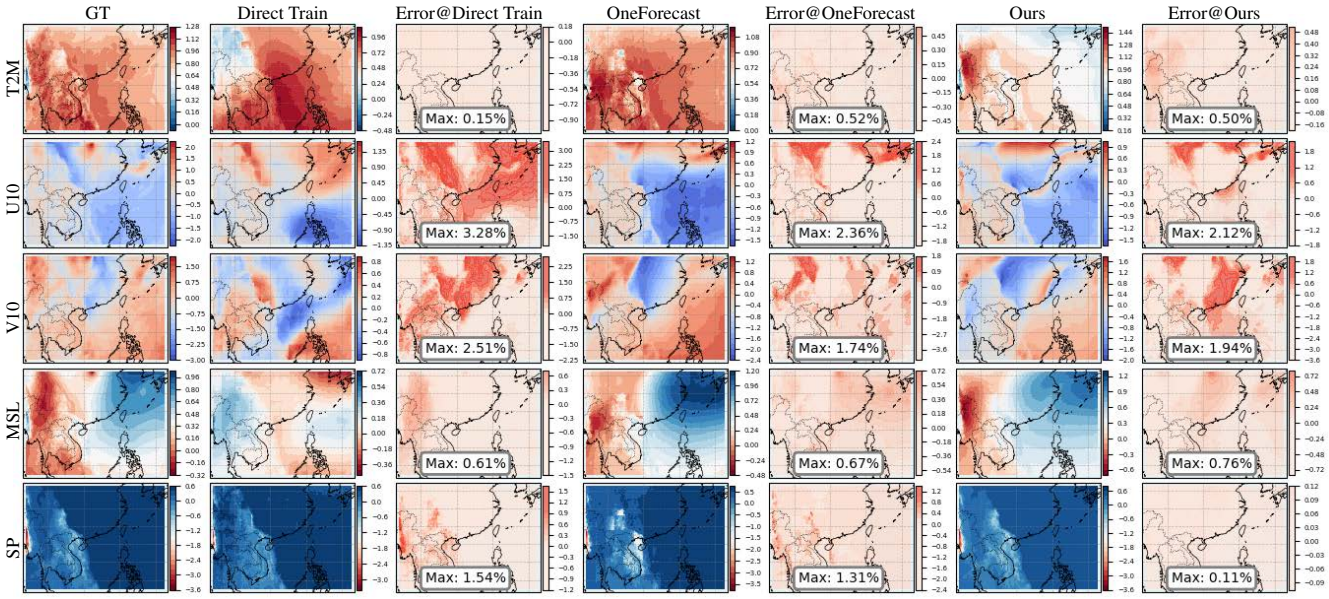


Figure 13. 5-day forecast results of regional weather among different models.

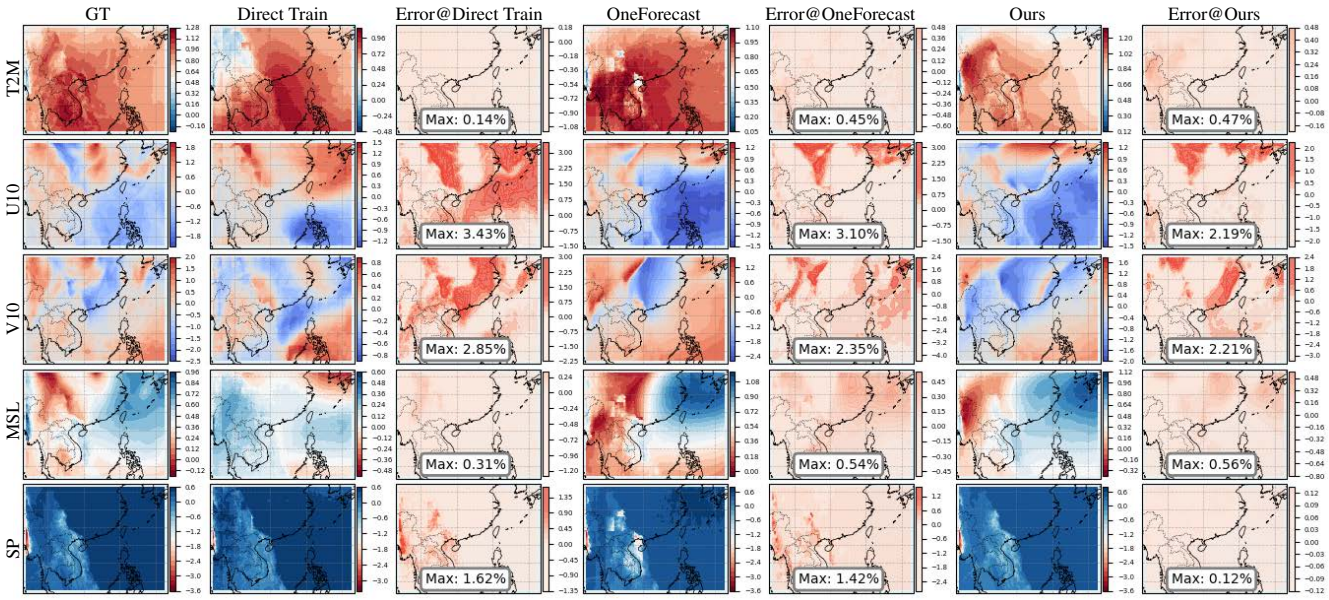


Figure 14. 5.5-day forecast results of regional weather among different models.

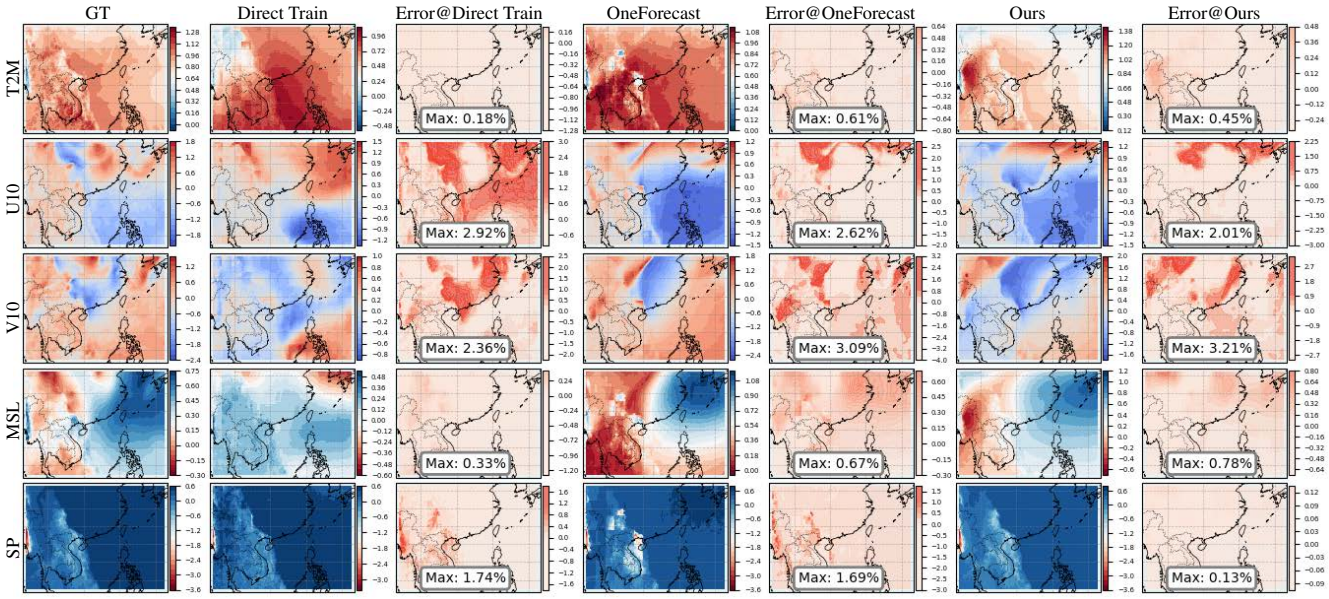


Figure 15. 6-day forecast results of regional weather among different models.

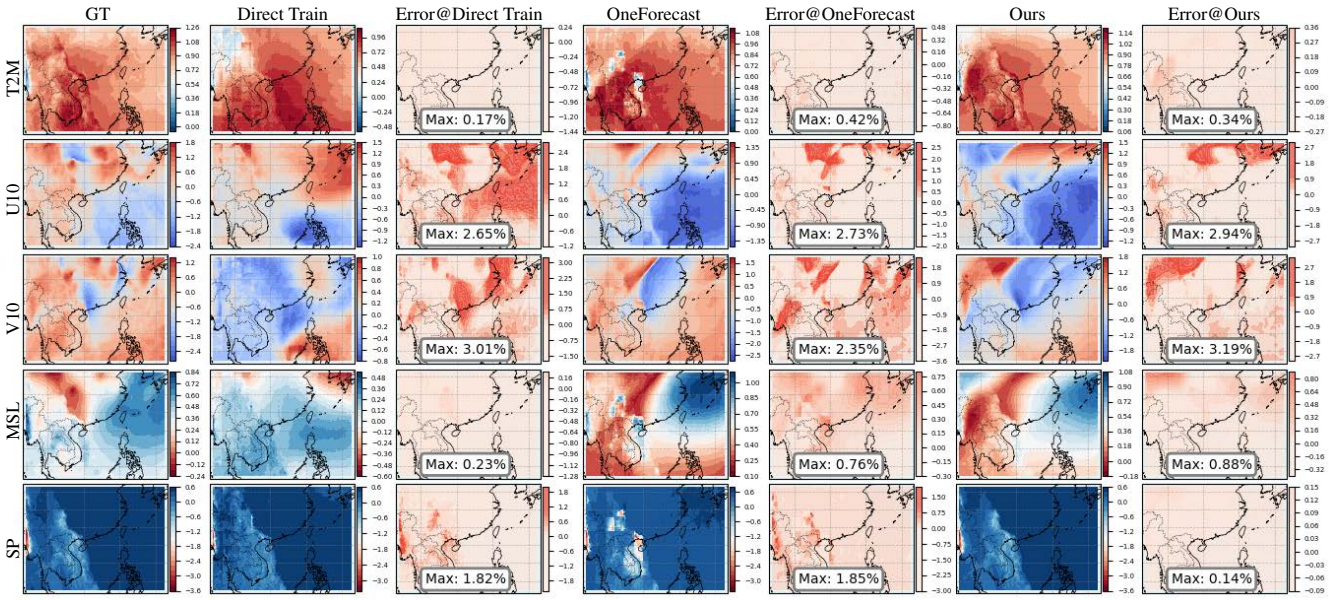


Figure 16. 6.5-day forecast results of regional weather among different models.

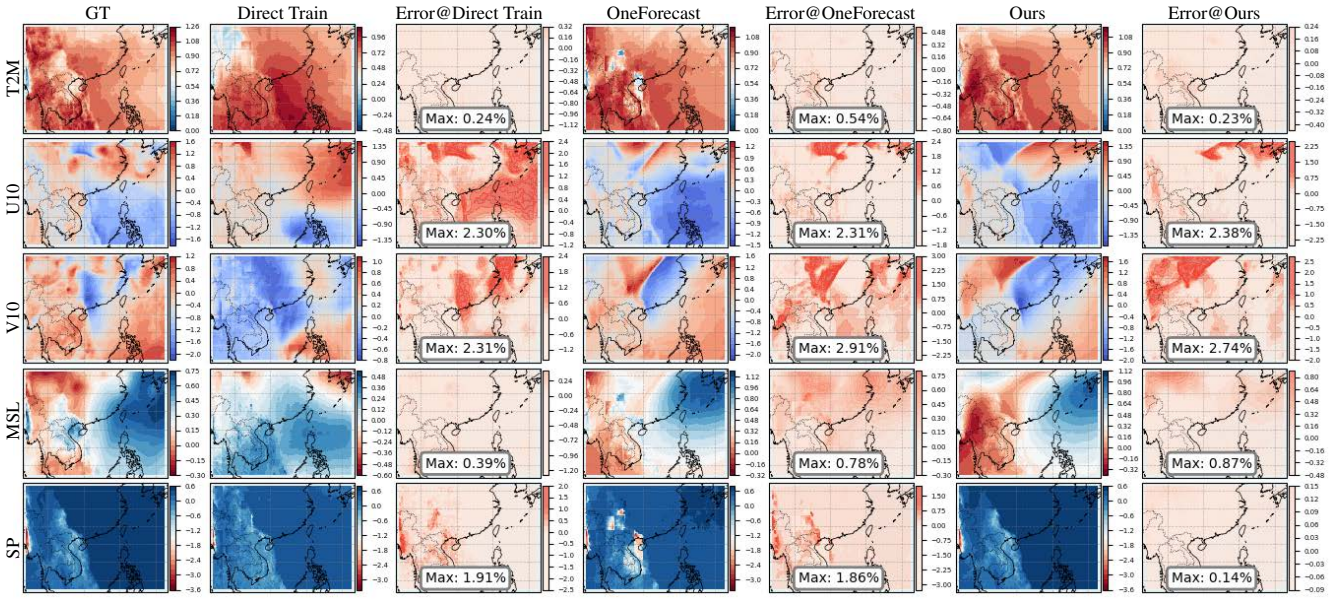


Figure 17. 7-day forecast results of regional weather among different models.

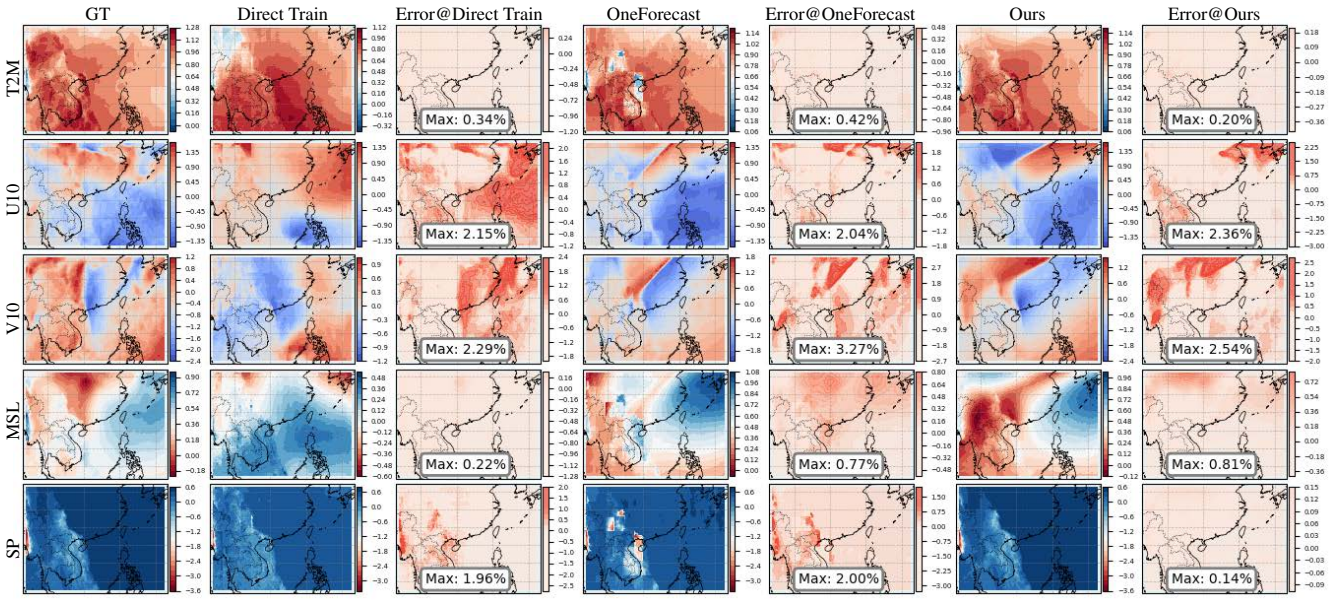


Figure 18. 7.5-day forecast results of regional weather among different models.

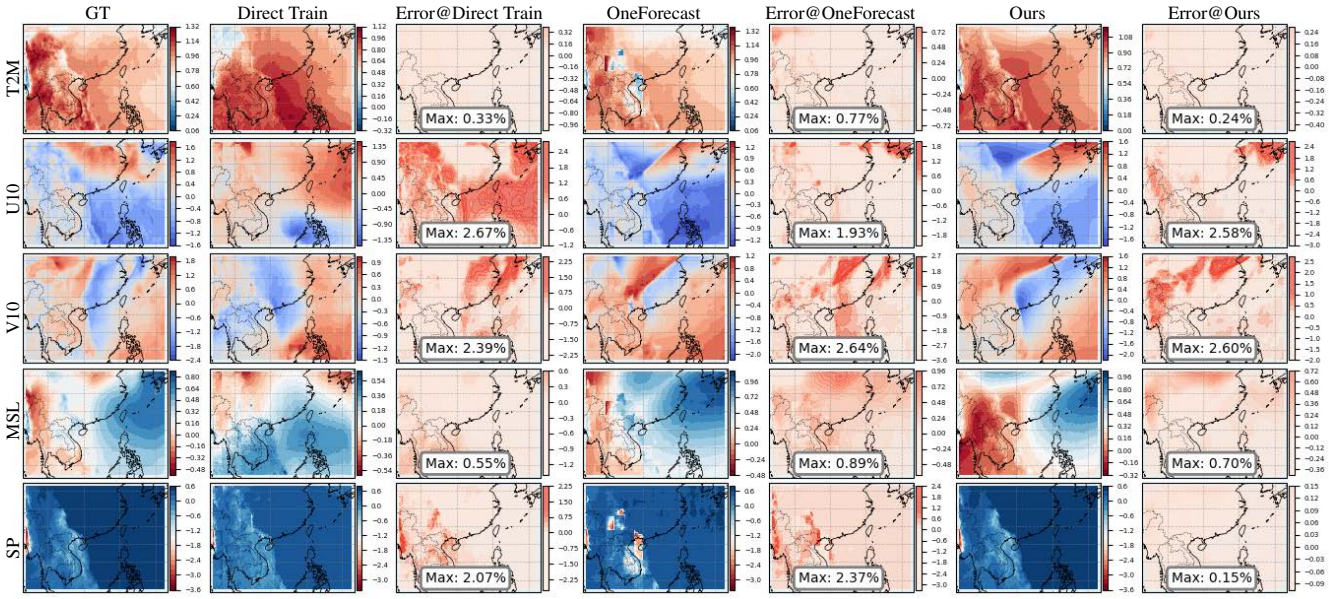


Figure 19. 8-day forecast results of regional weather among different models.

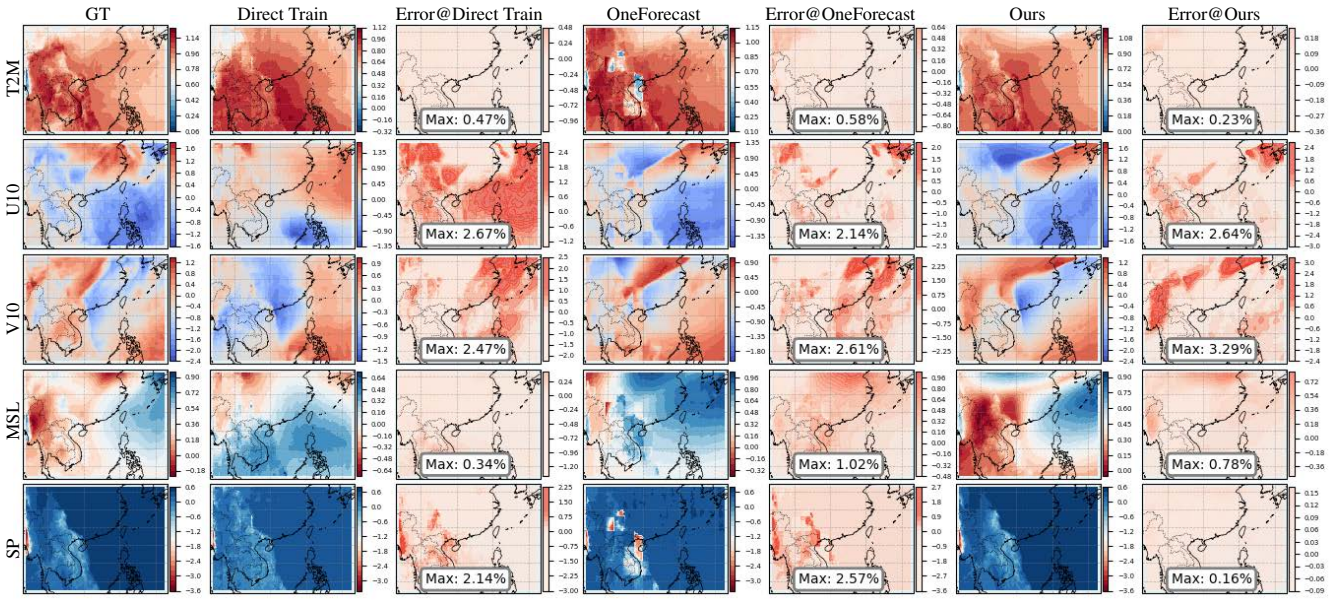


Figure 20. 8.5-day forecast results of regional weather among different models.

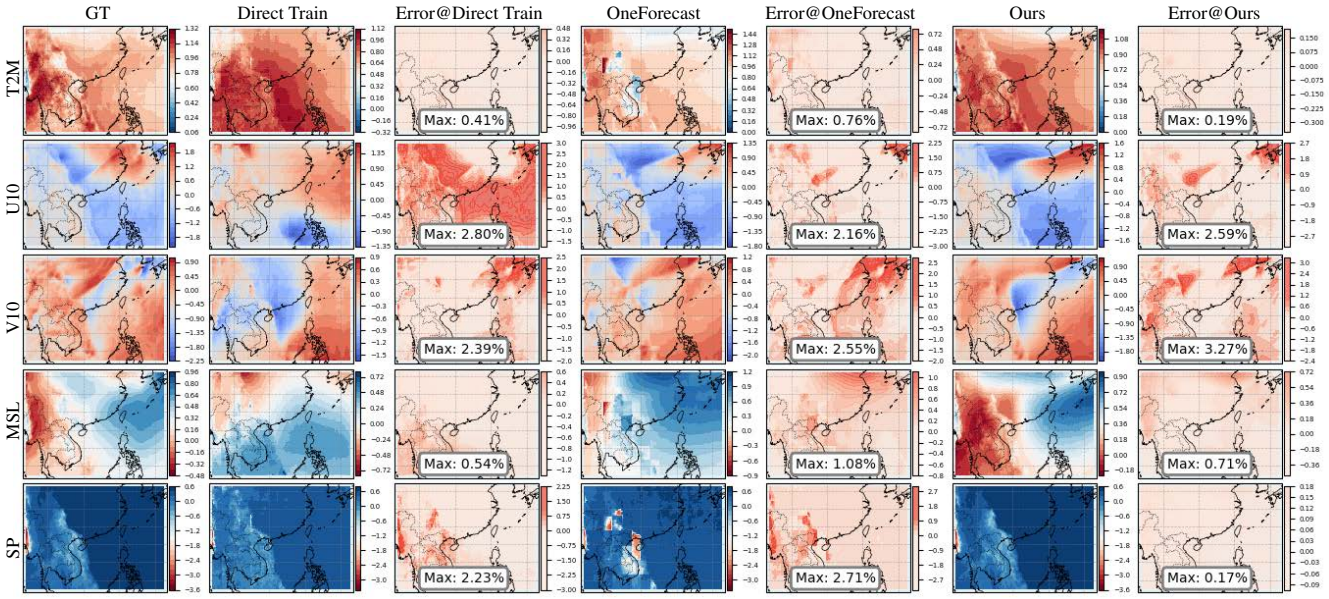


Figure 21. 9-day forecast results of regional weather among different models.

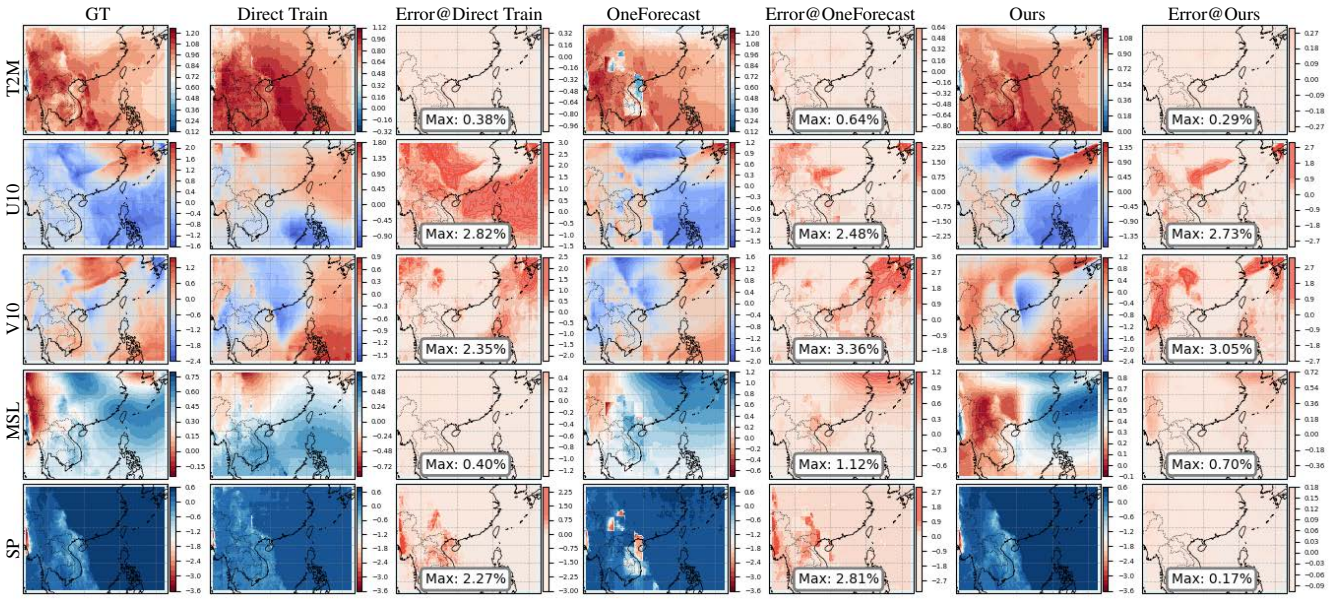


Figure 22. 9.5-day forecast results of regional weather among different models.

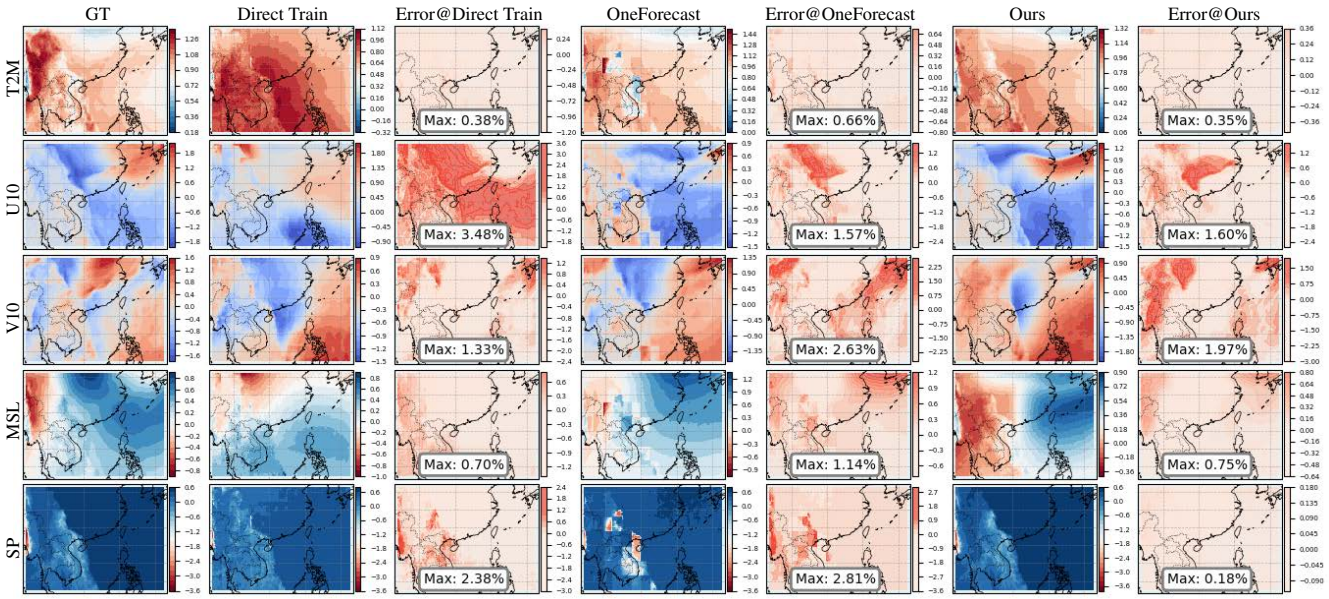


Figure 23. 10-day forecast results of regional weather among different models.

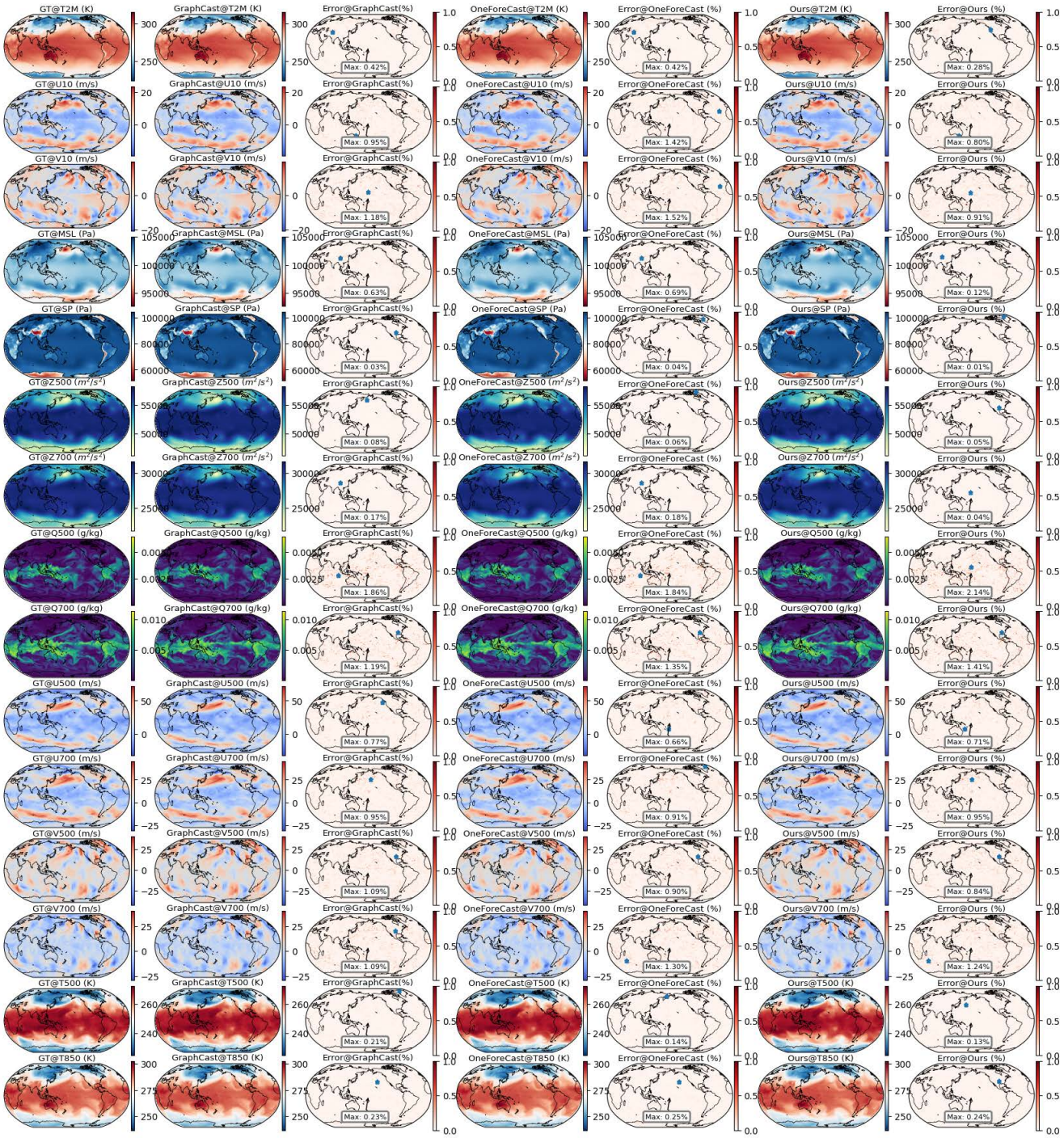


Figure 24. 6-hour forecast results of global weather among different models.

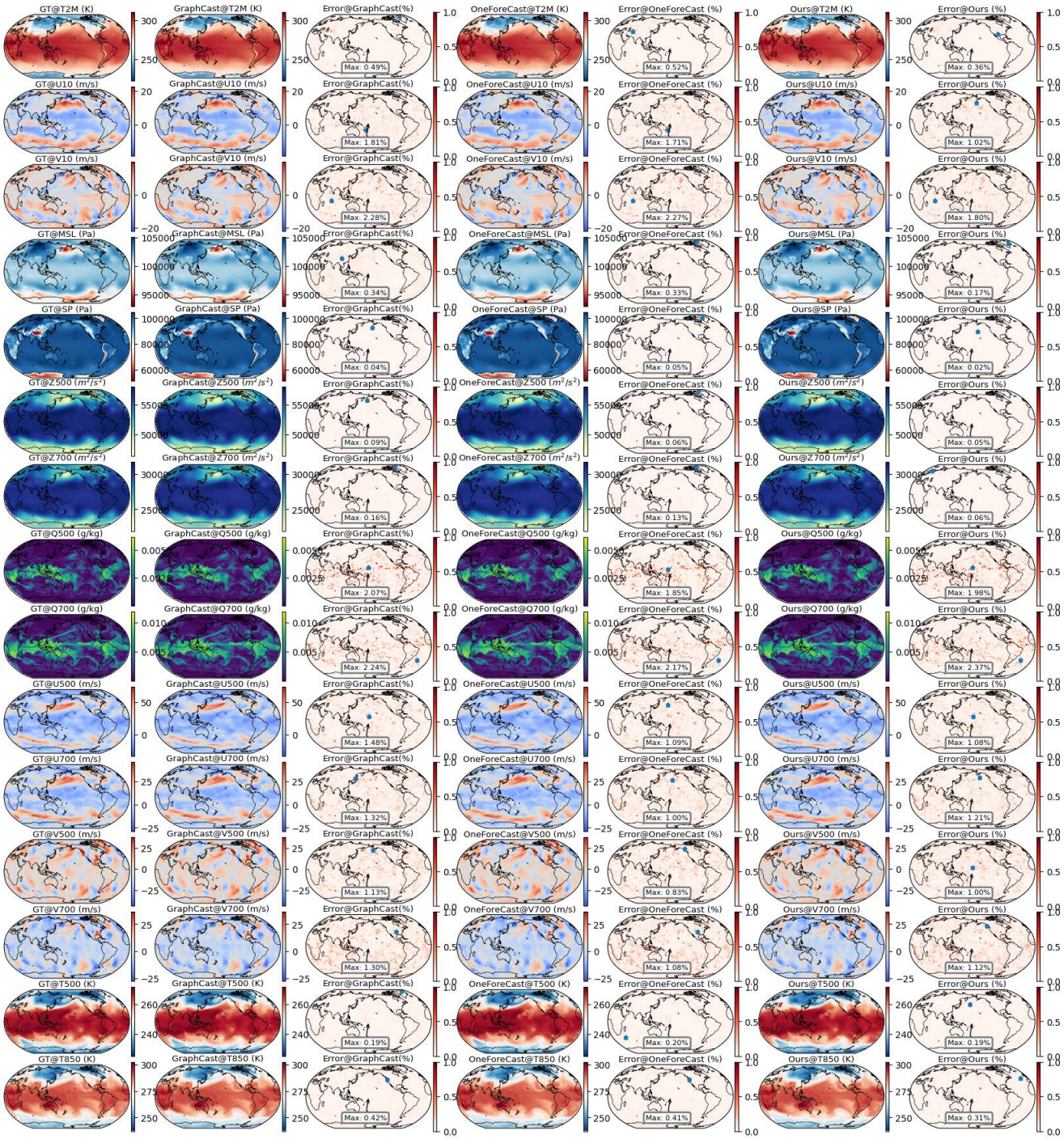


Figure 25. 0.5-day forecast results of global weather among different models.

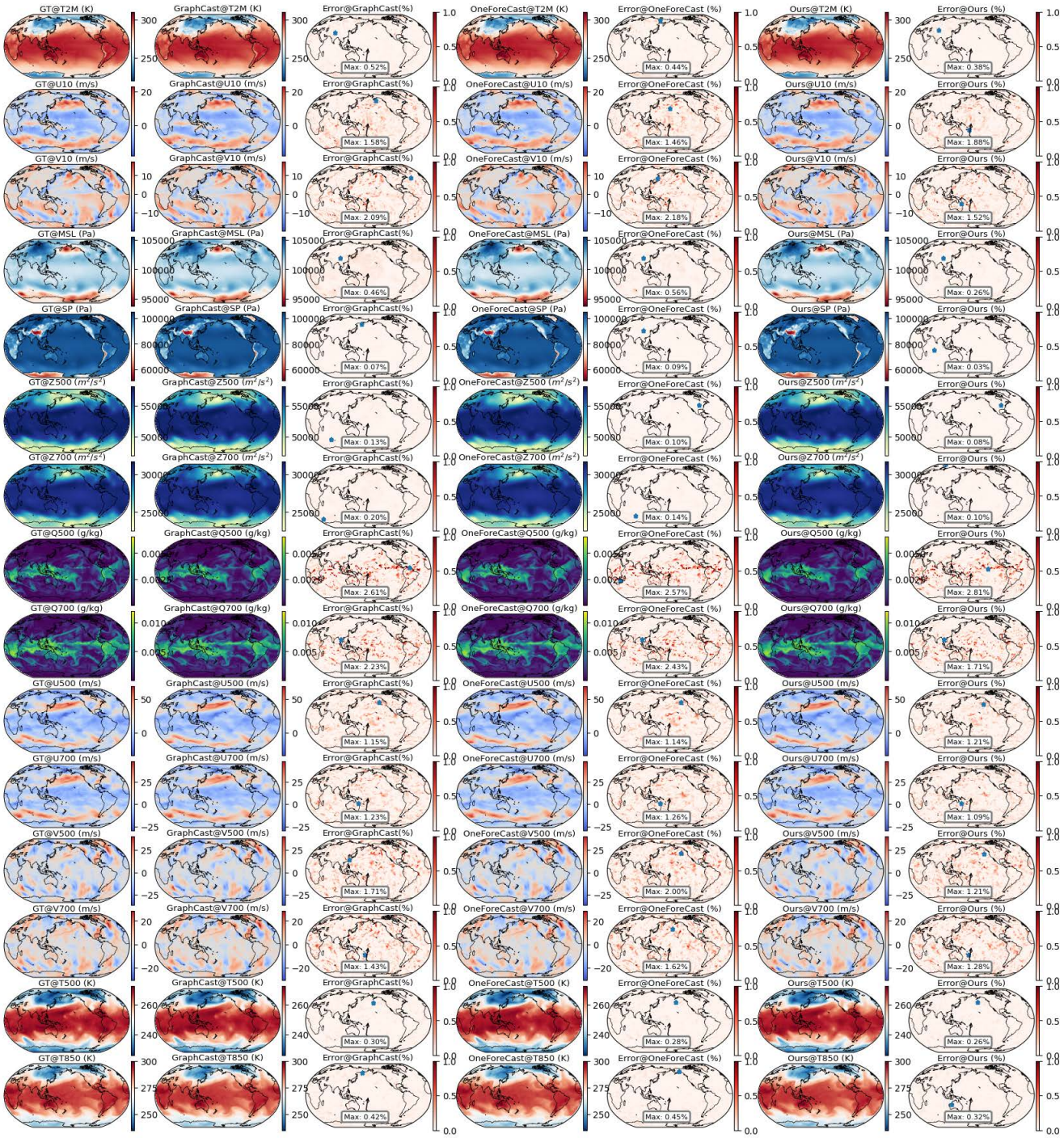


Figure 26. 1-day forecast results of global weather among different models.

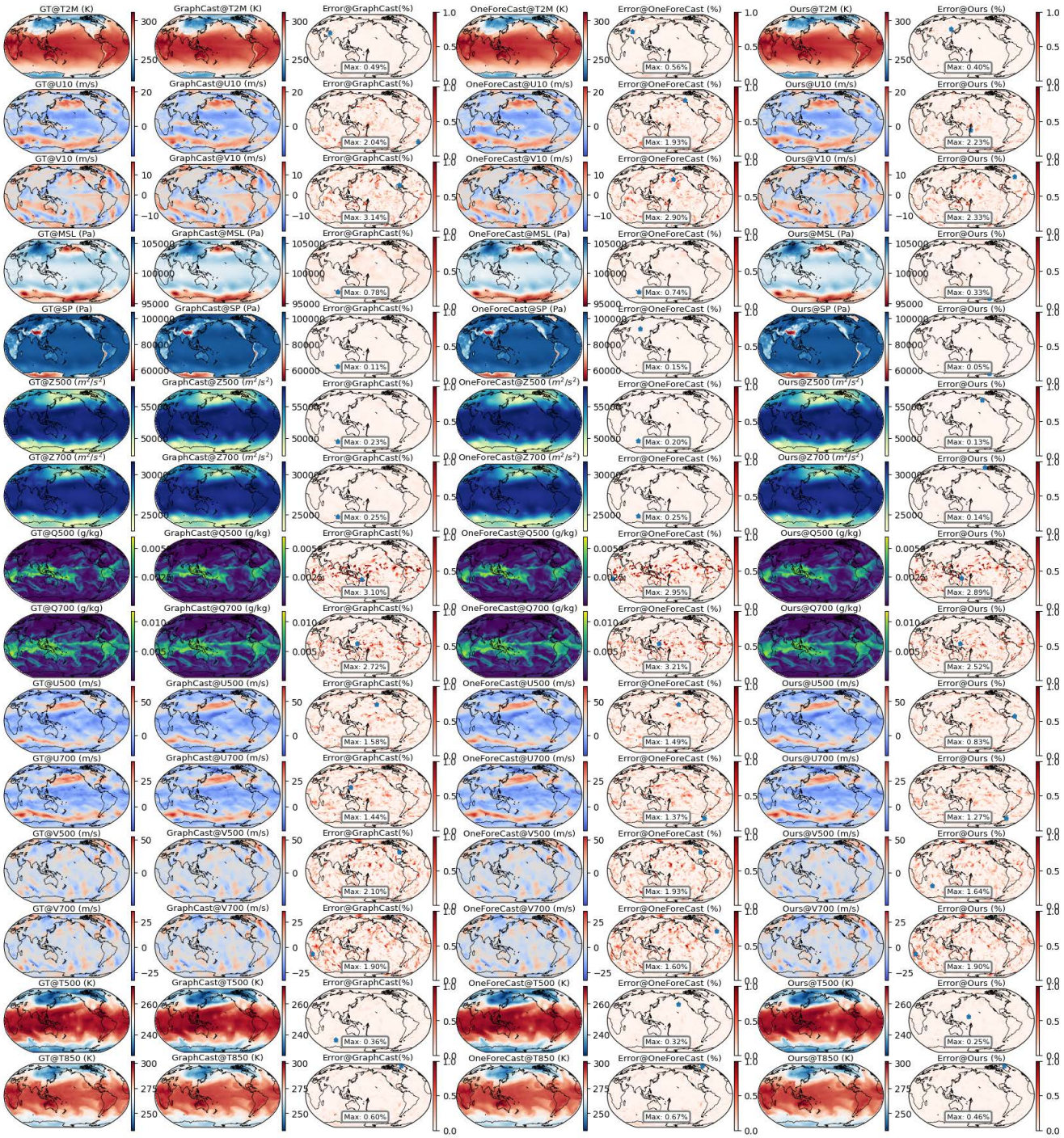


Figure 27. 1.5-day forecast results of global weather among different models.

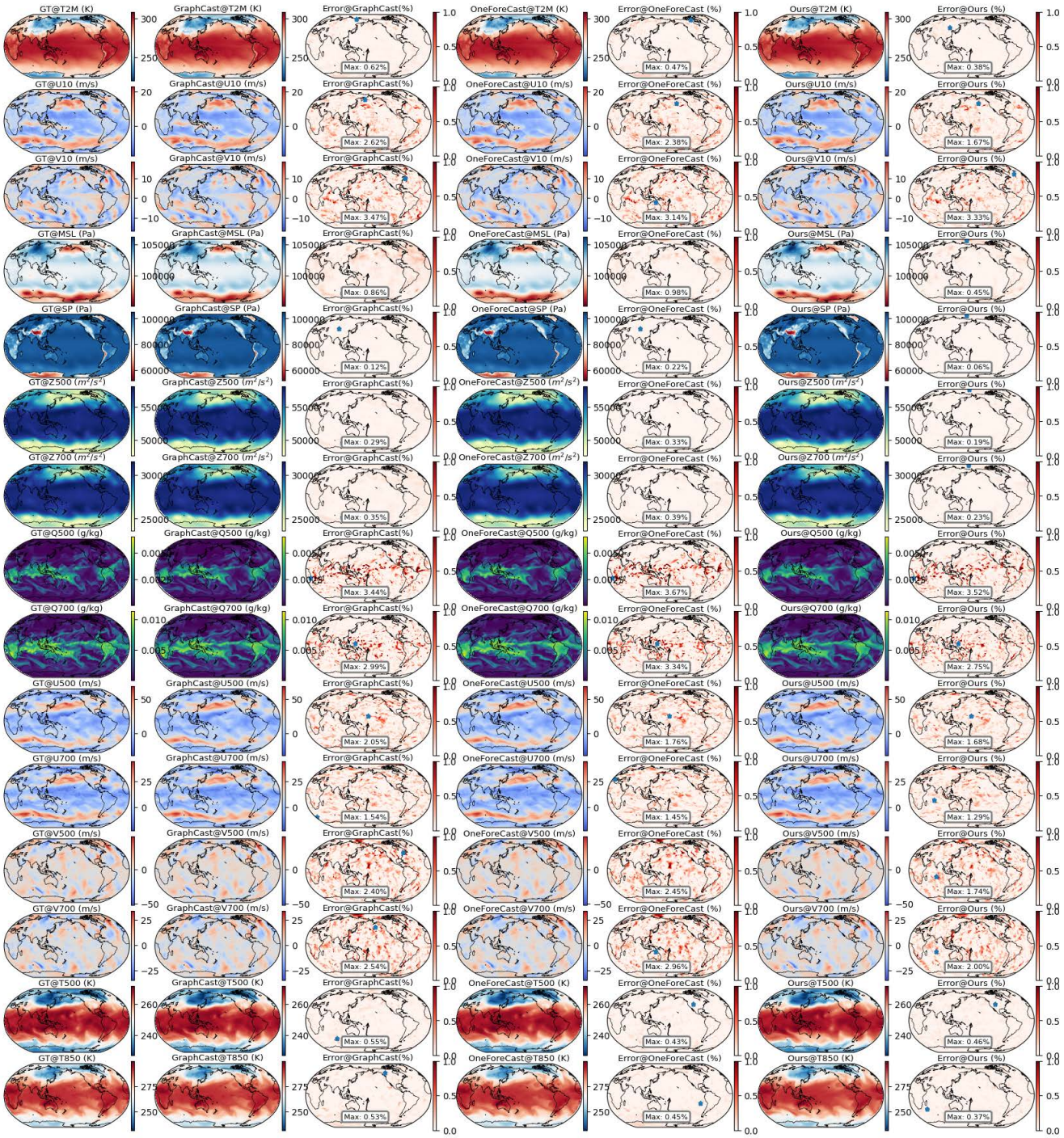


Figure 28. 2-day forecast results of global weather among different models.

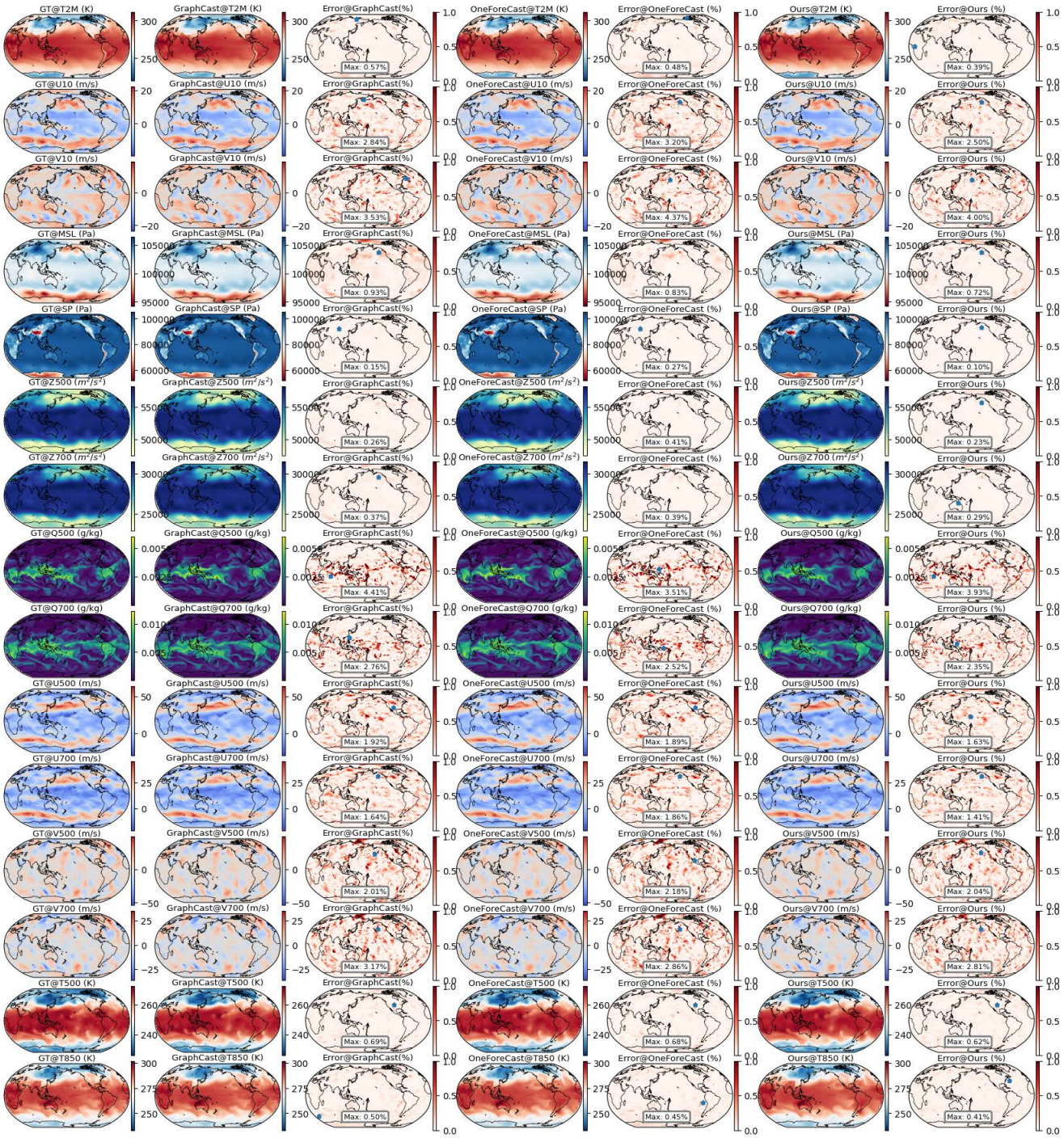


Figure 29. 2.5-day forecast results of global weather among different models.

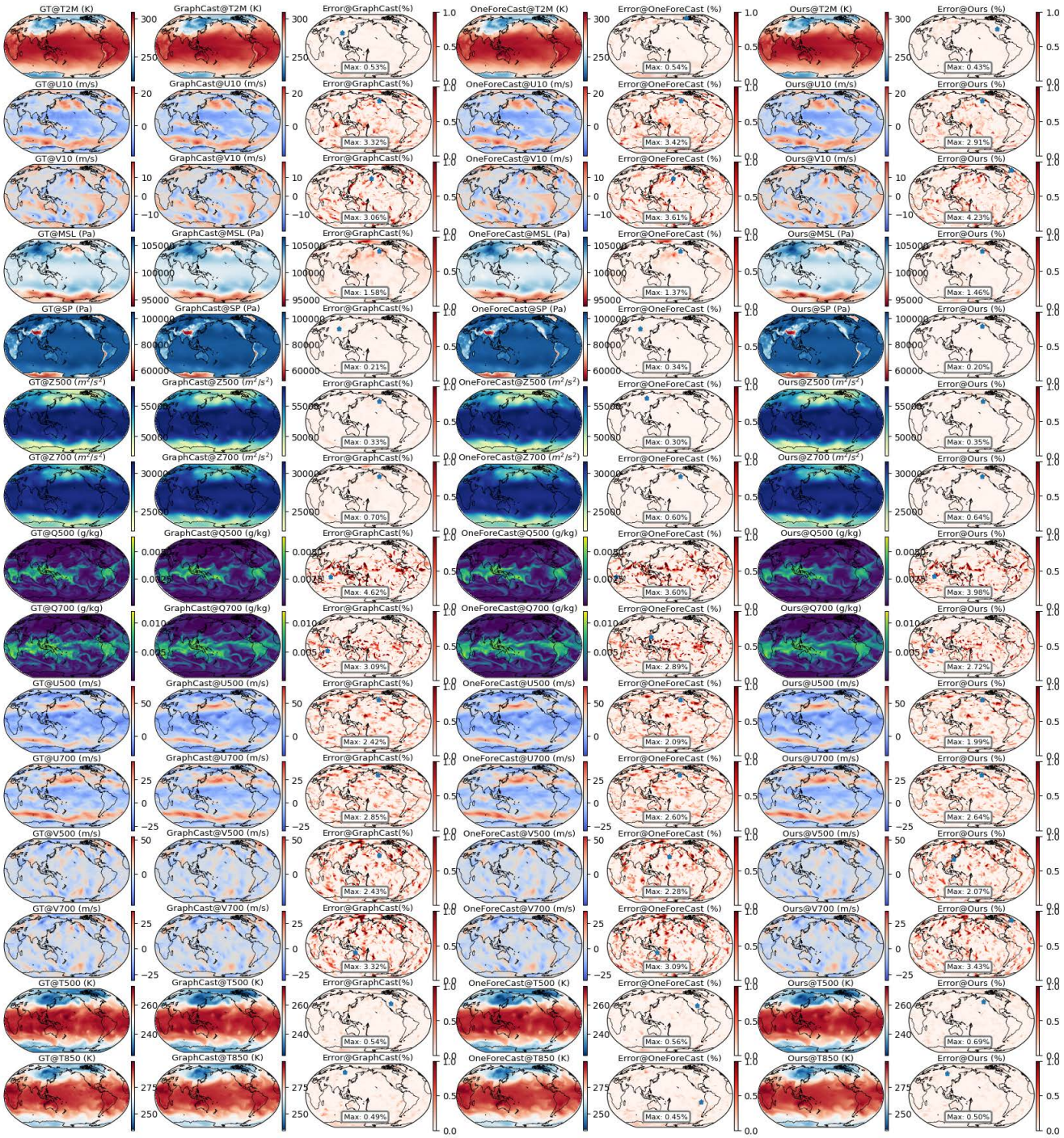


Figure 30. 3-day forecast results of global weather among different models.

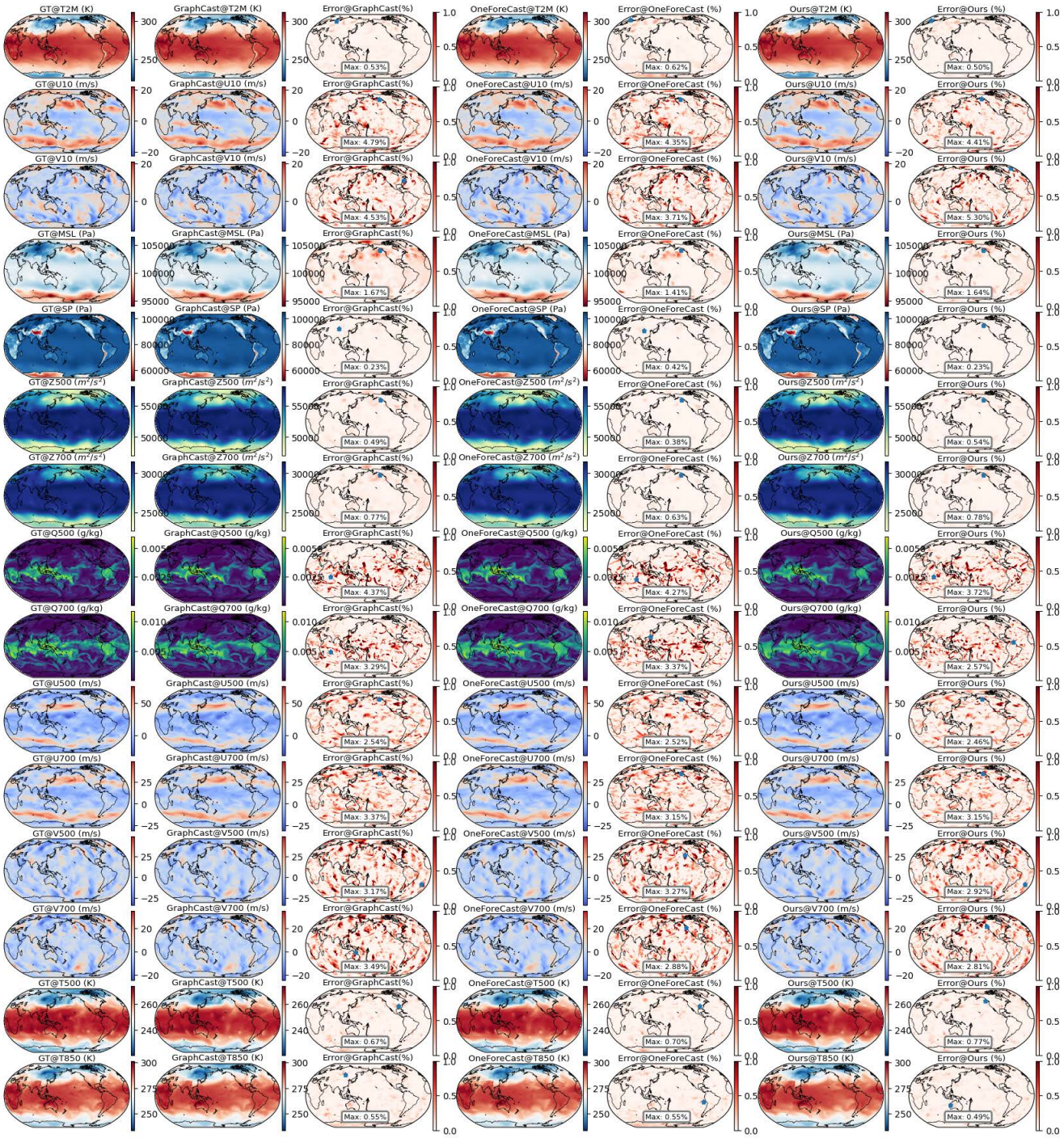


Figure 31. 3.5-day forecast results of global weather among different models.

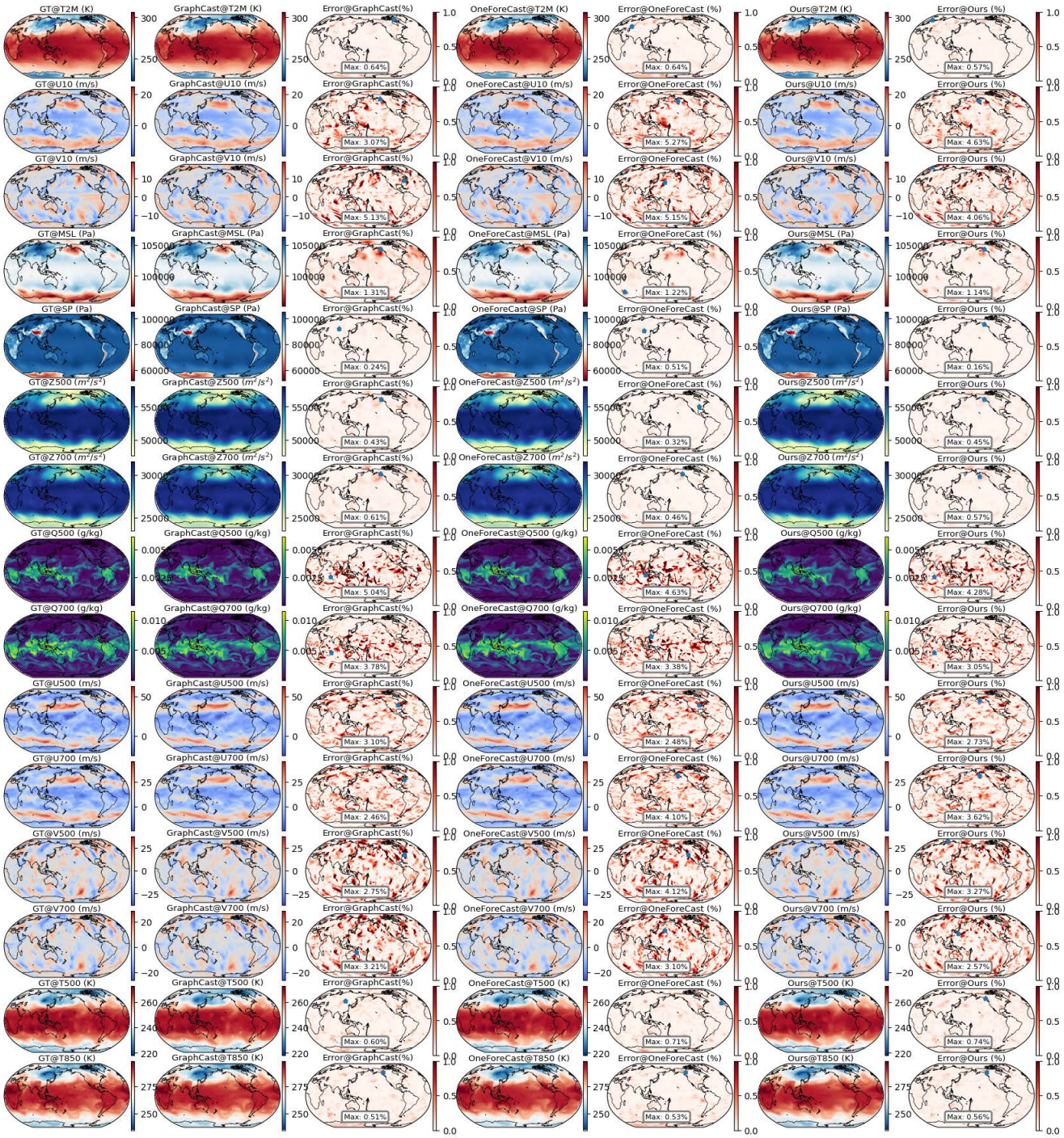


Figure 32. 4-day forecast results of global weather among different models.

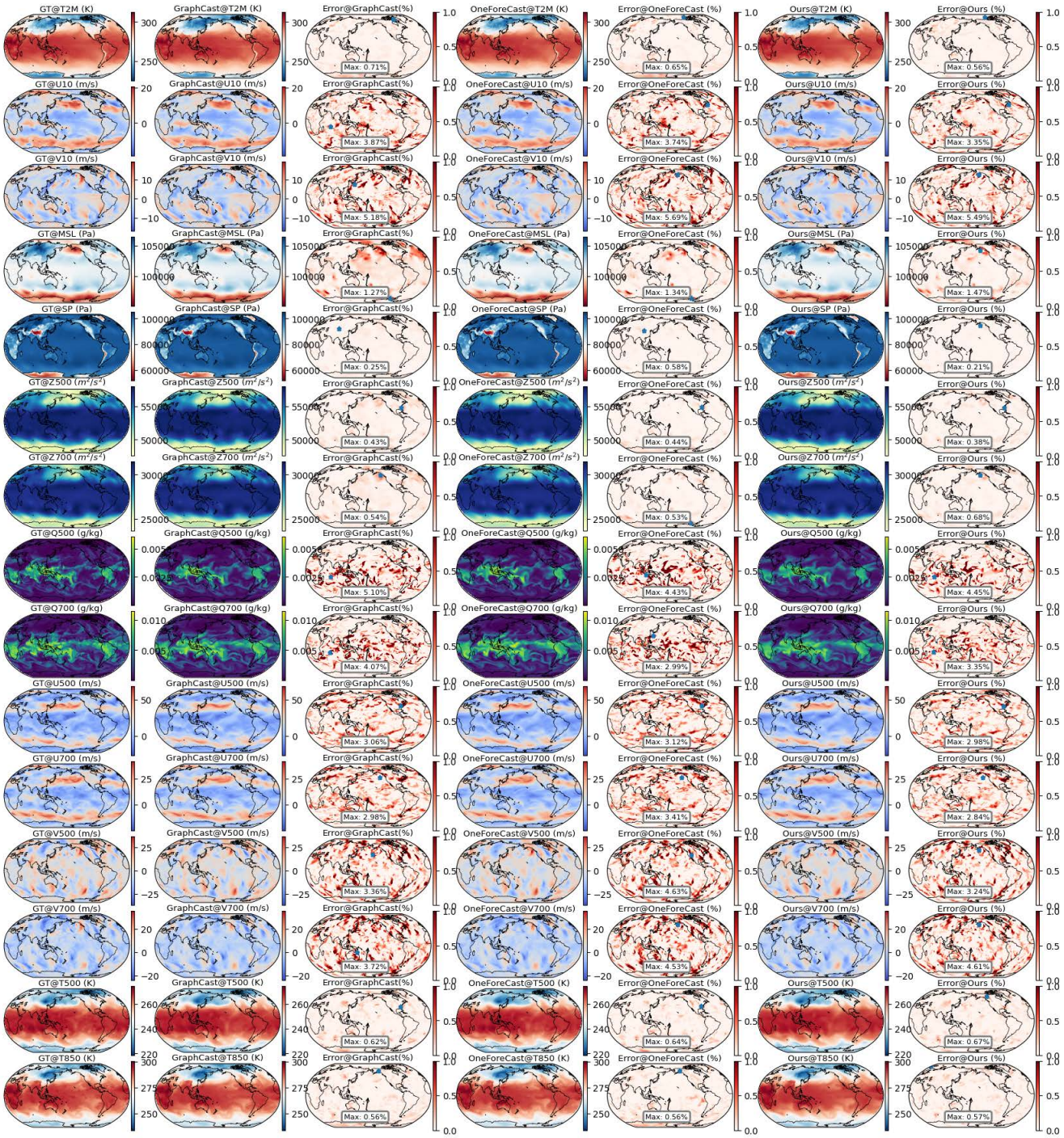


Figure 33. 4.5-day forecast results of global weather among different models.

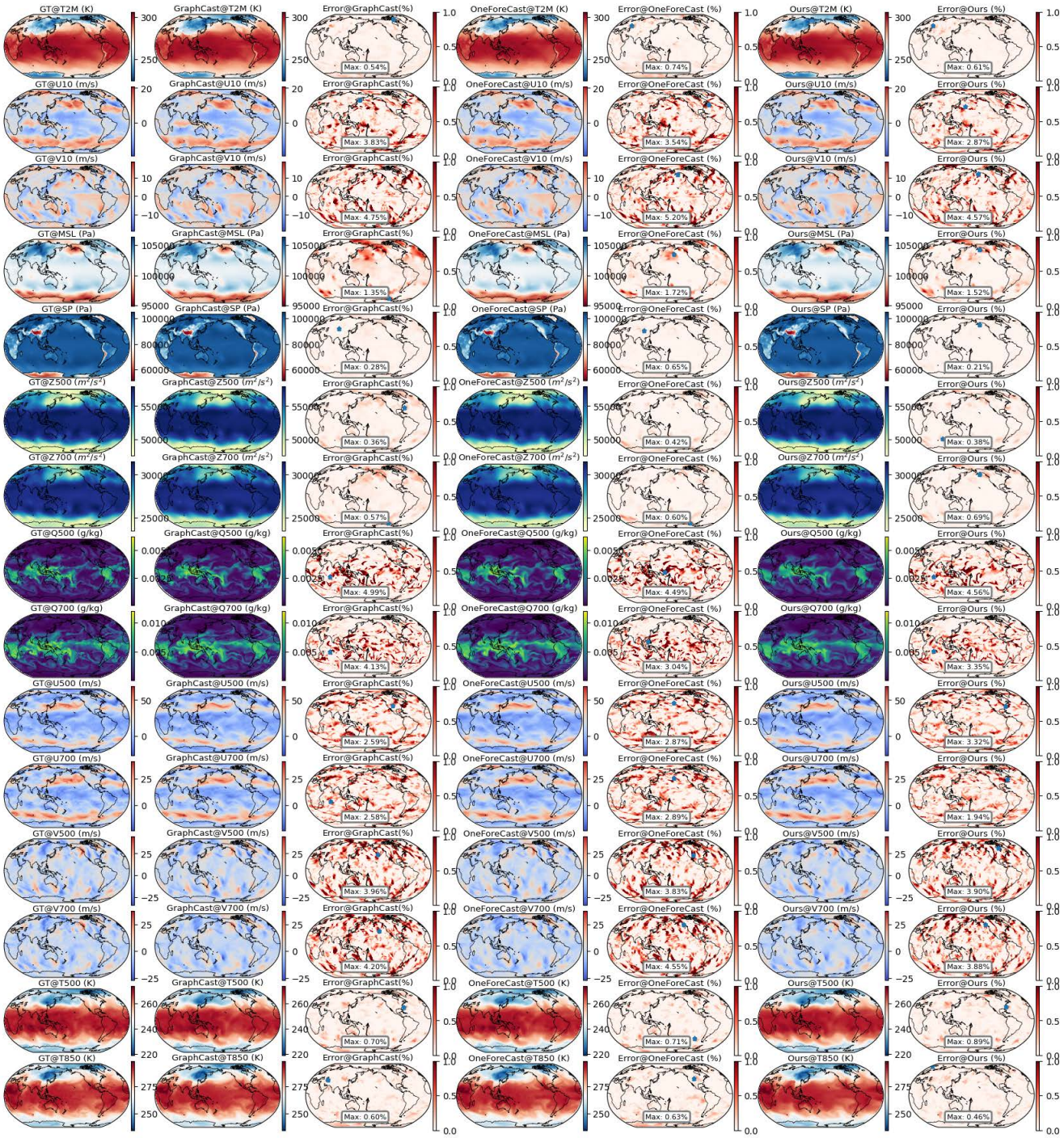


Figure 34. 5-day forecast results of global weather among different models.

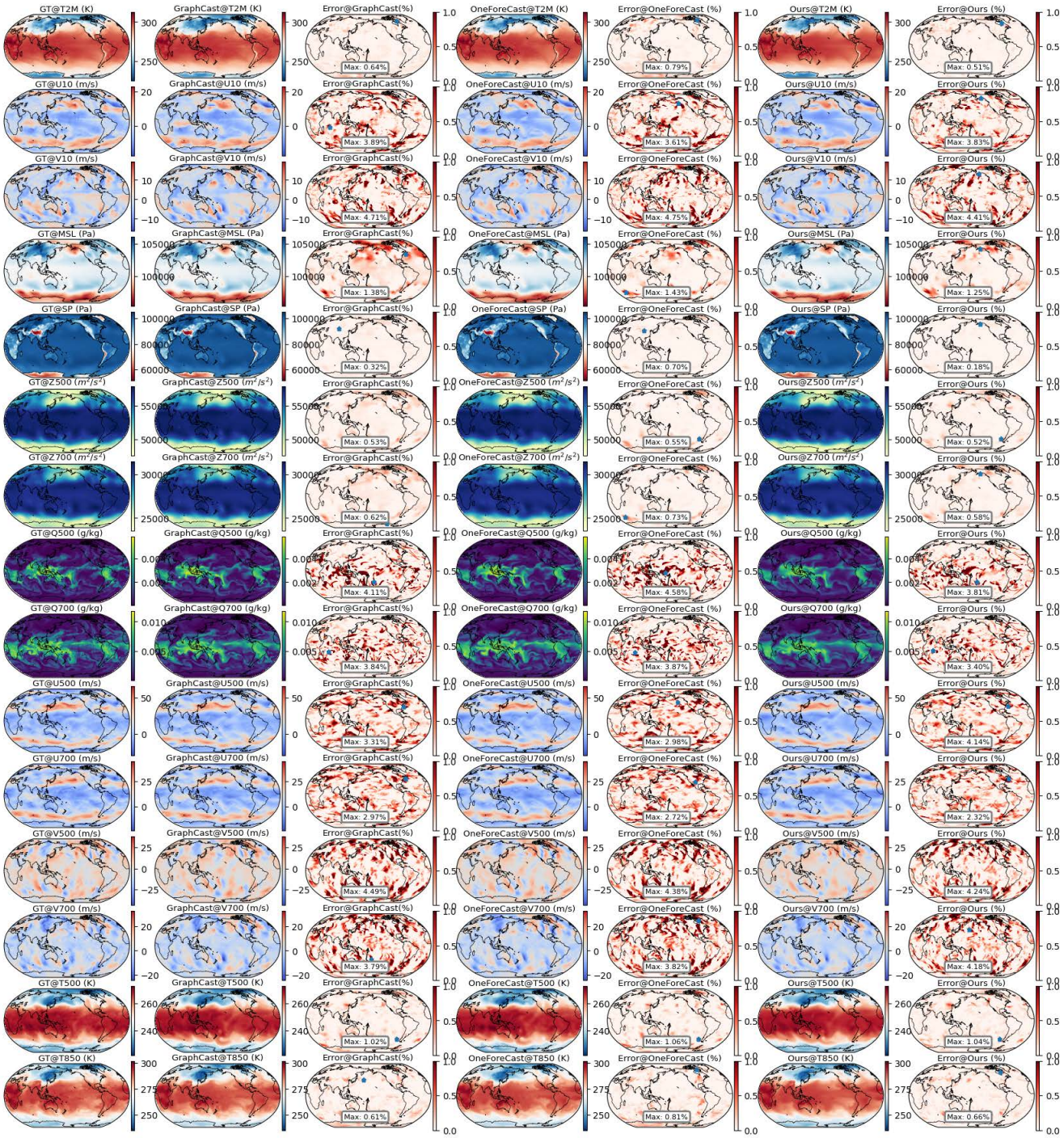


Figure 35. 5.5-day forecast results of global weather among different models.

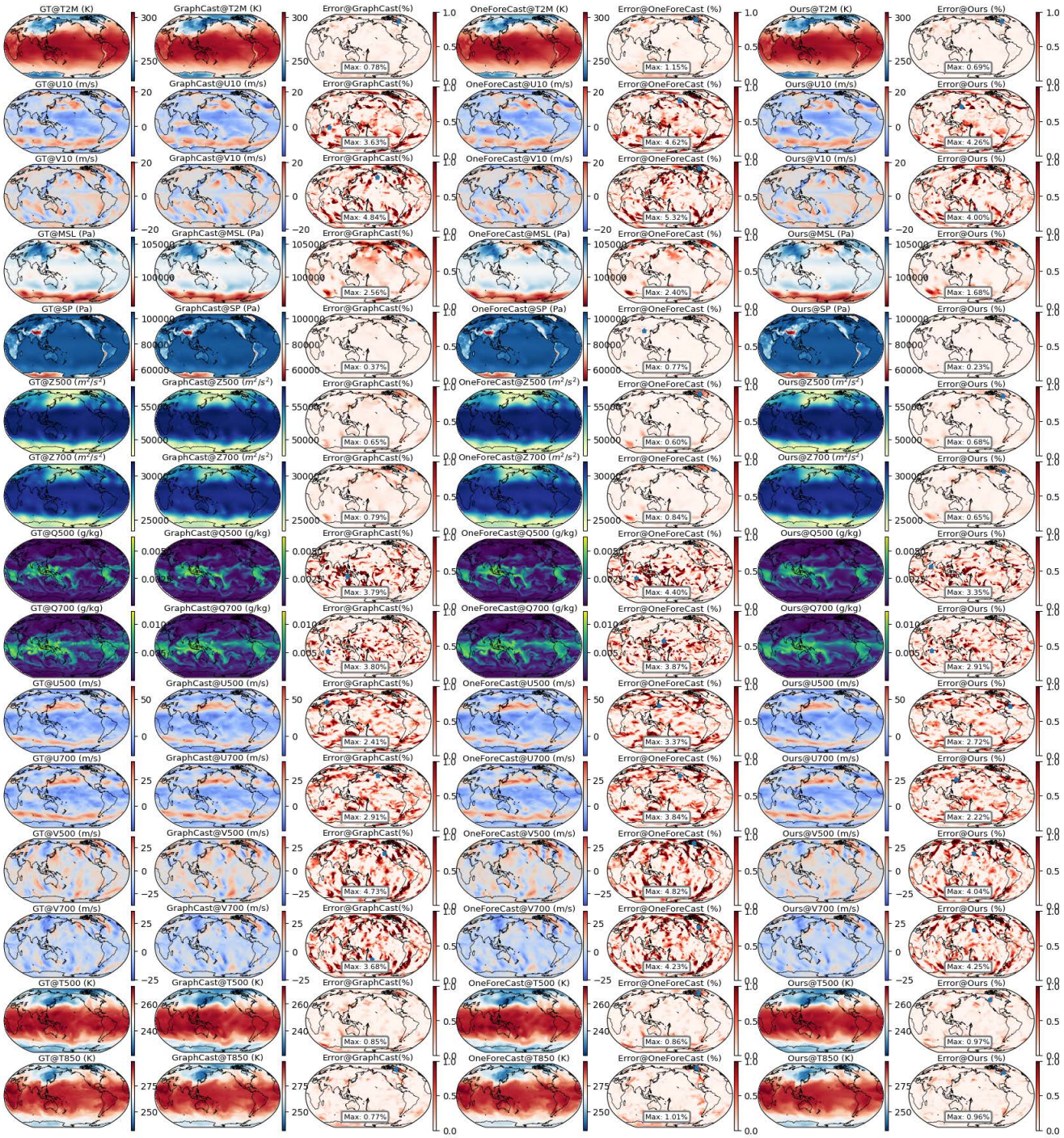


Figure 36. 6-day forecast results of global weather among different models.

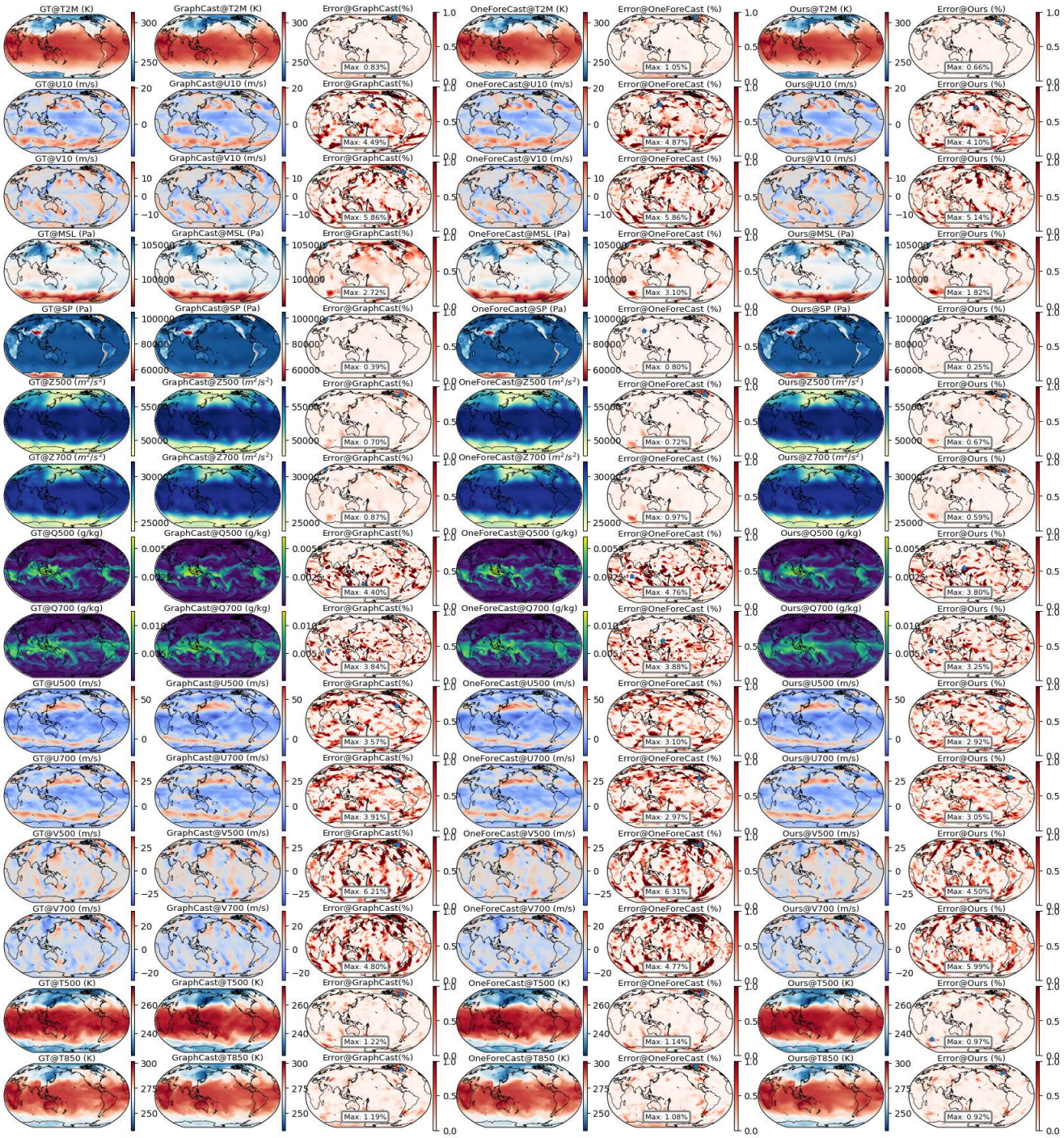


Figure 37. 6.5-day forecast results of global weather among different models.

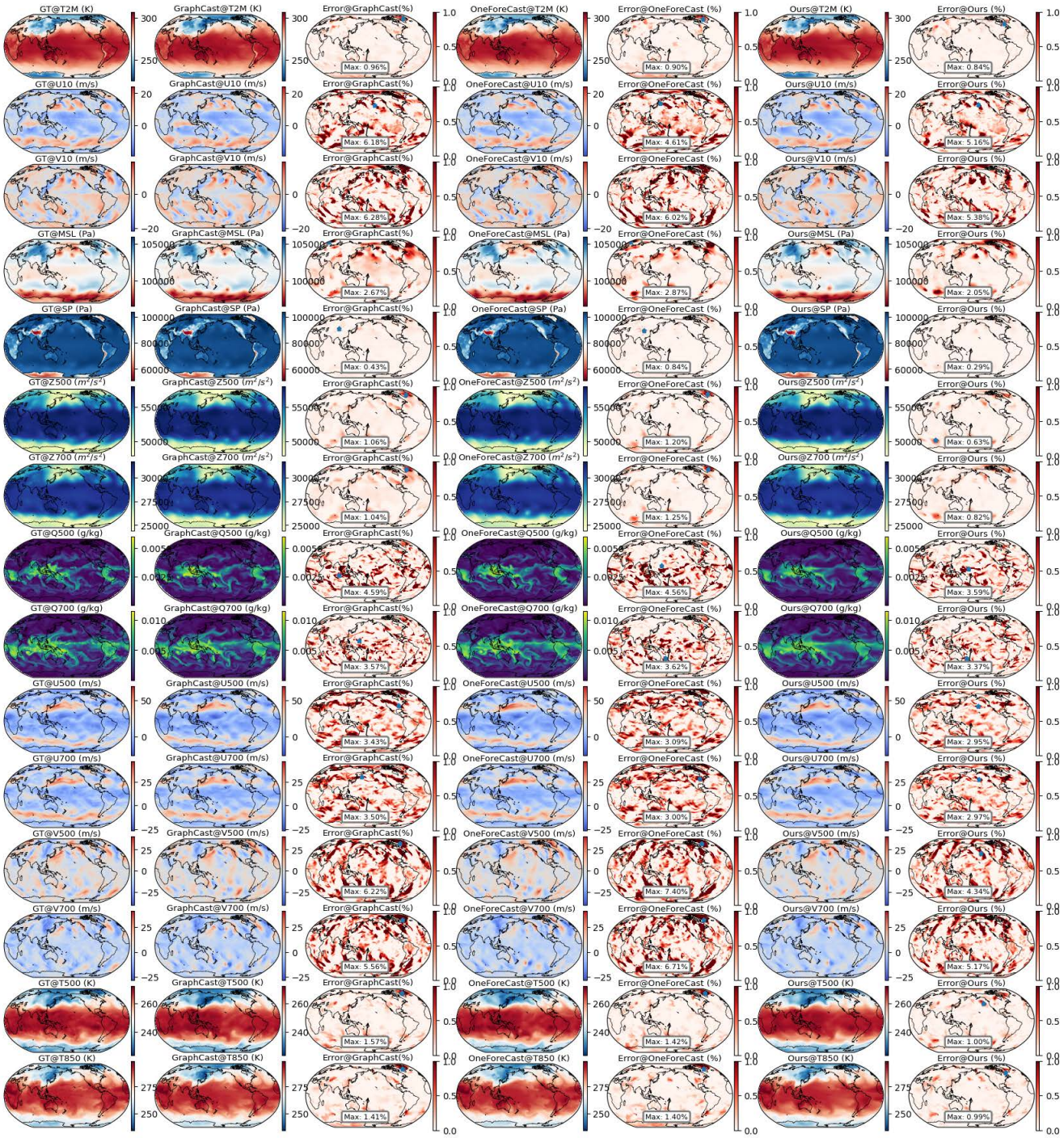


Figure 38. 7-day forecast results of global weather among different models.

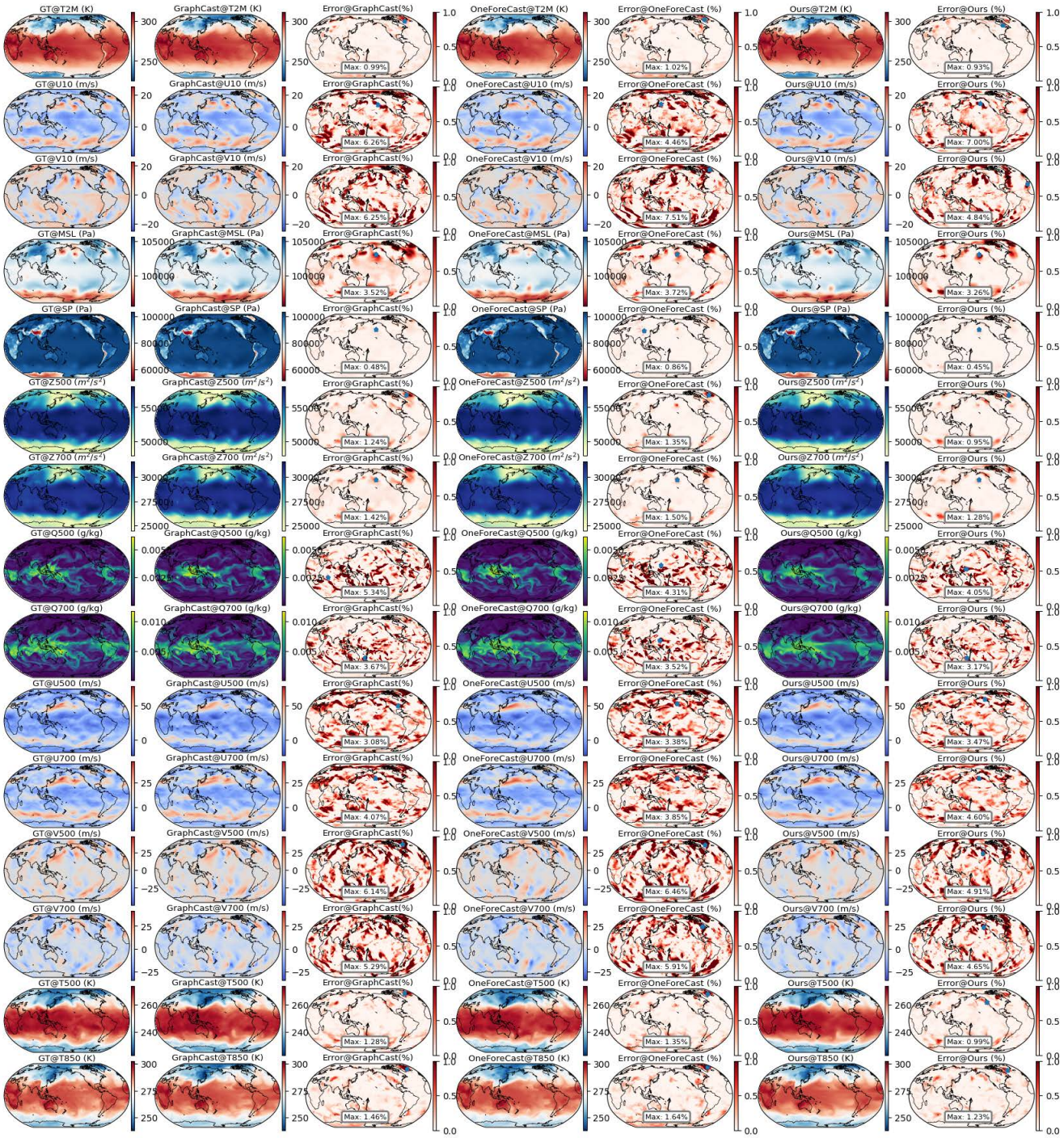


Figure 39. 7.5-day forecast results of global weather among different models.

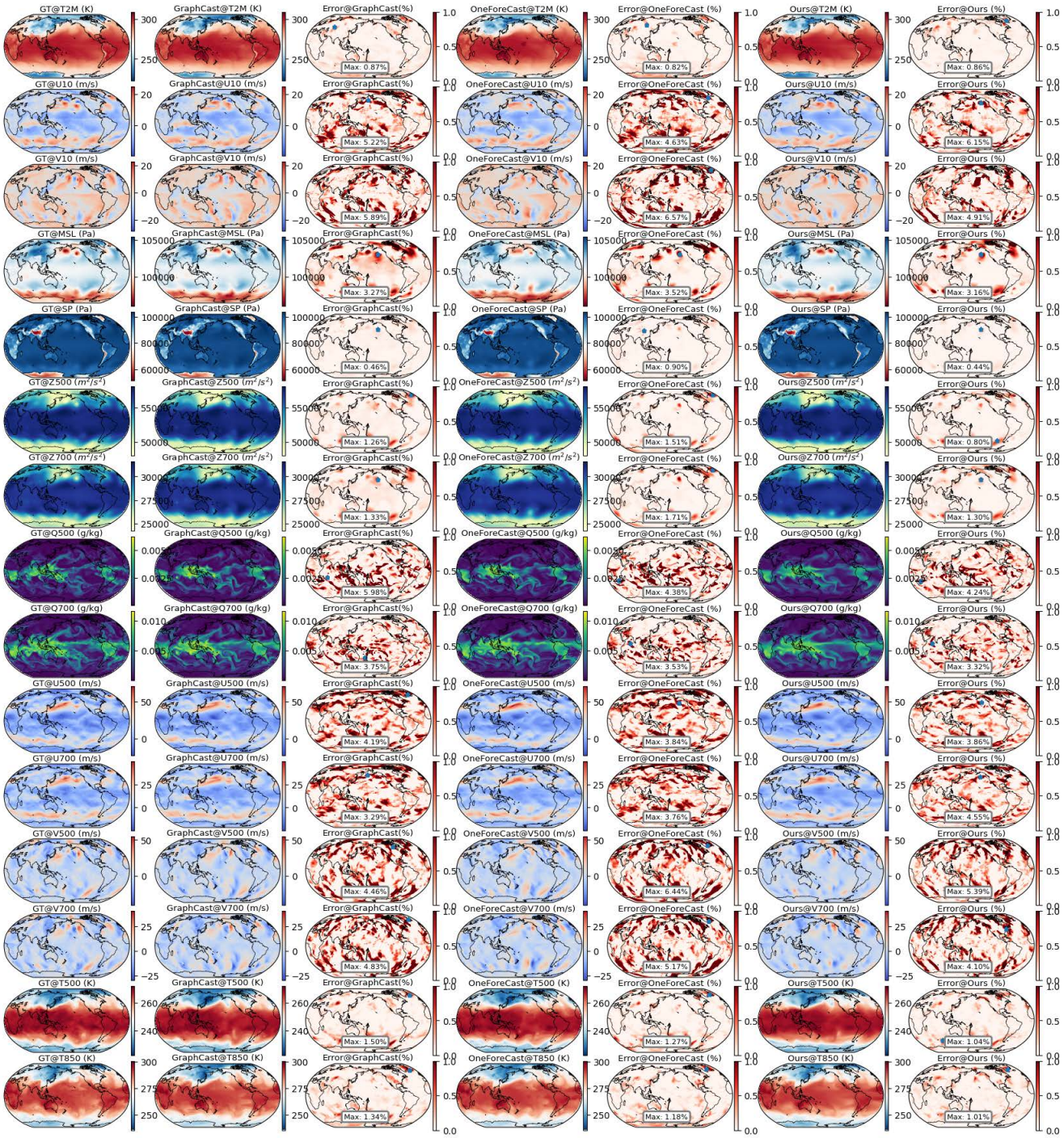


Figure 40. 8-day forecast results of global weather among different models.

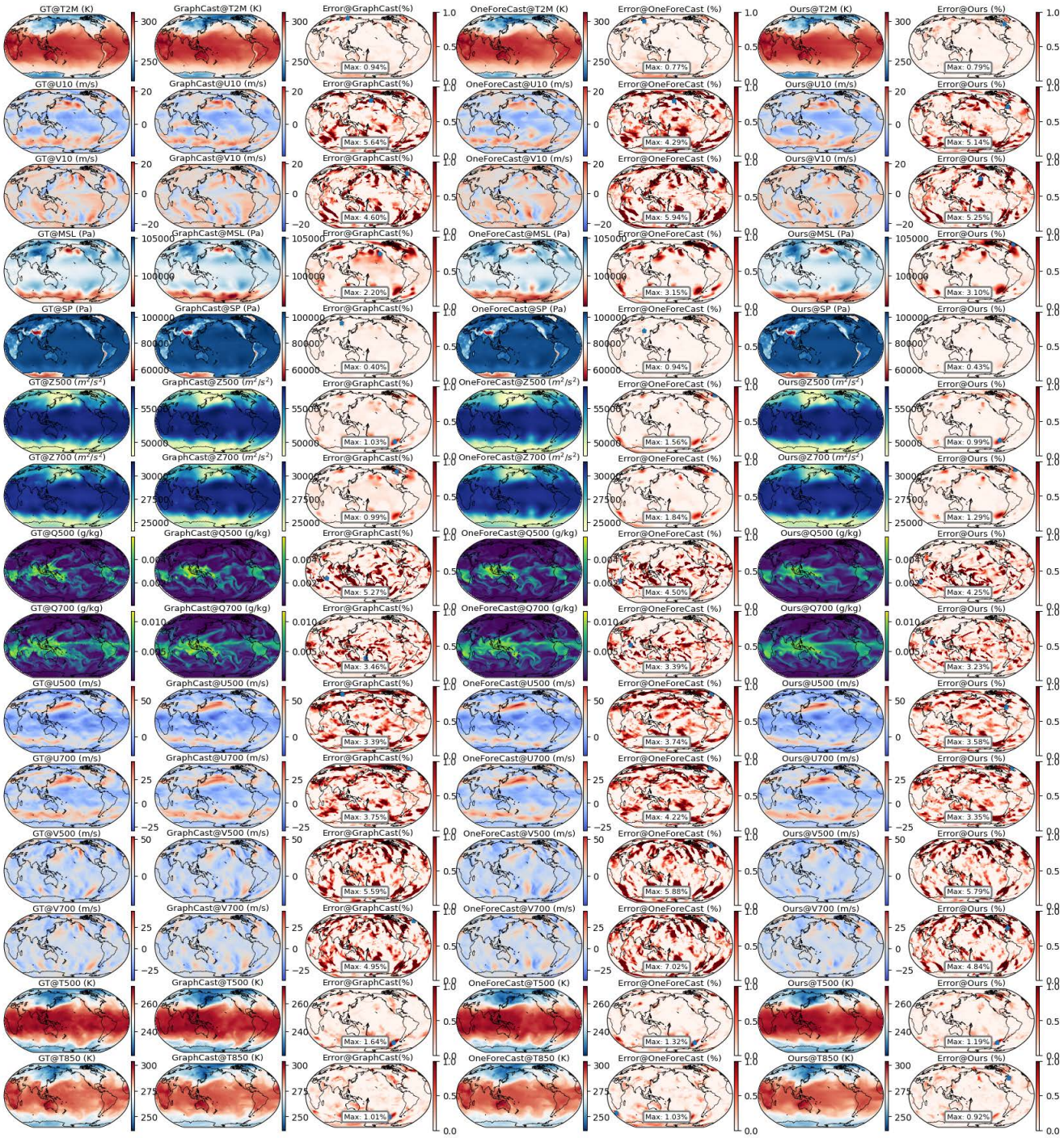


Figure 41. 8.5-day forecast results of global weather among different models.

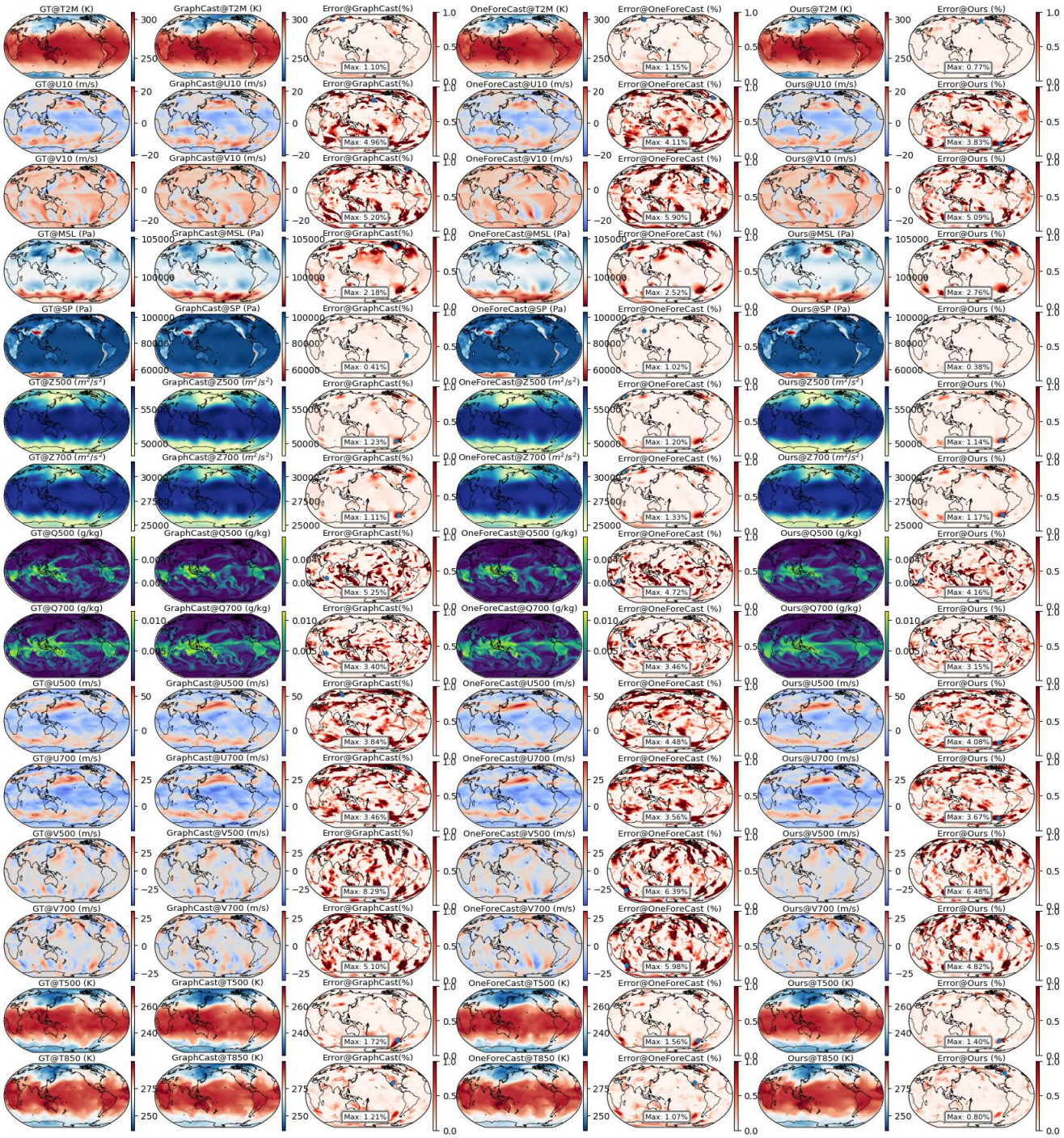


Figure 42. 9-day forecast results of global weather among different models.

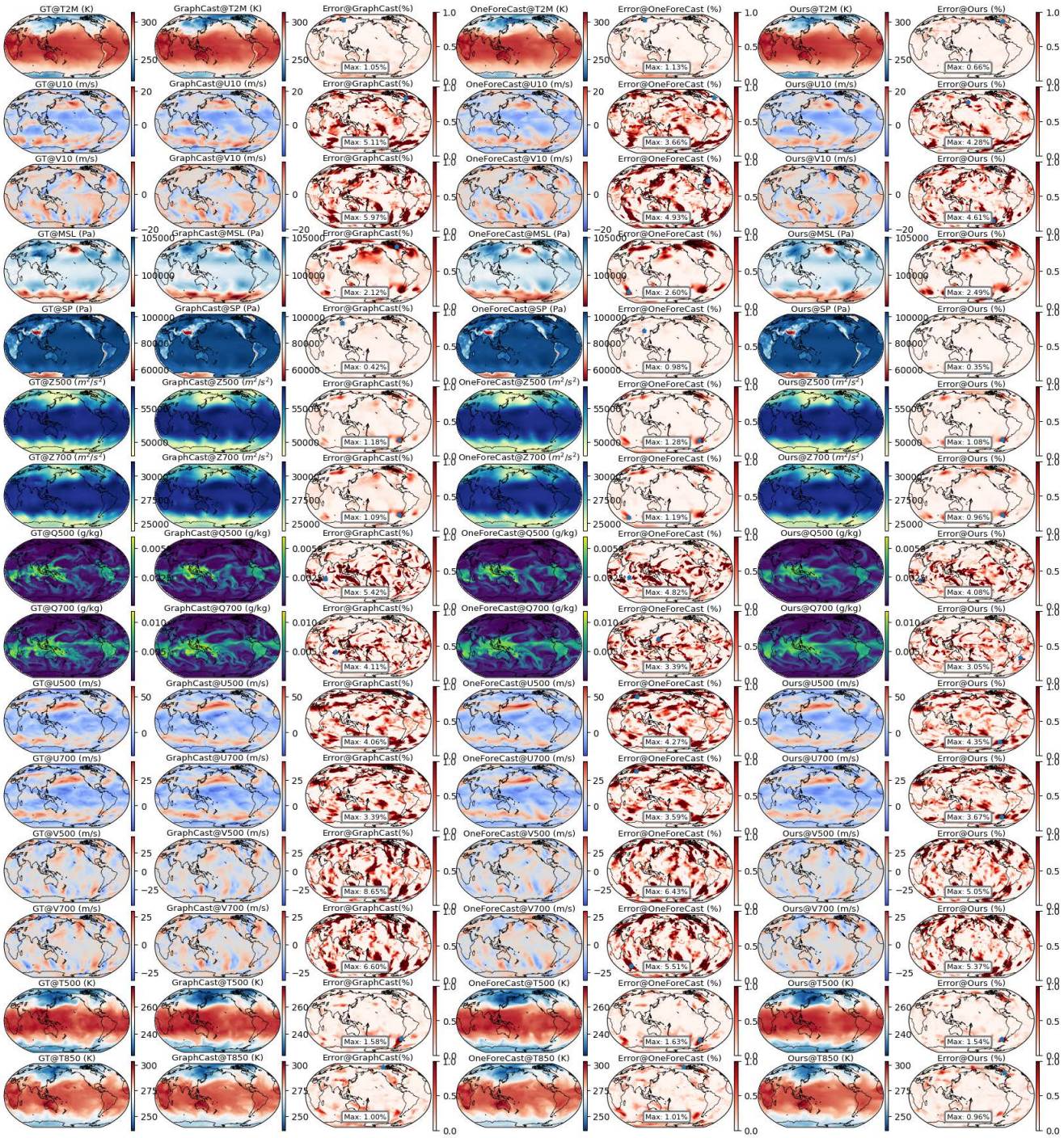


Figure 43. 9.5-day forecast results of global weather among different models.

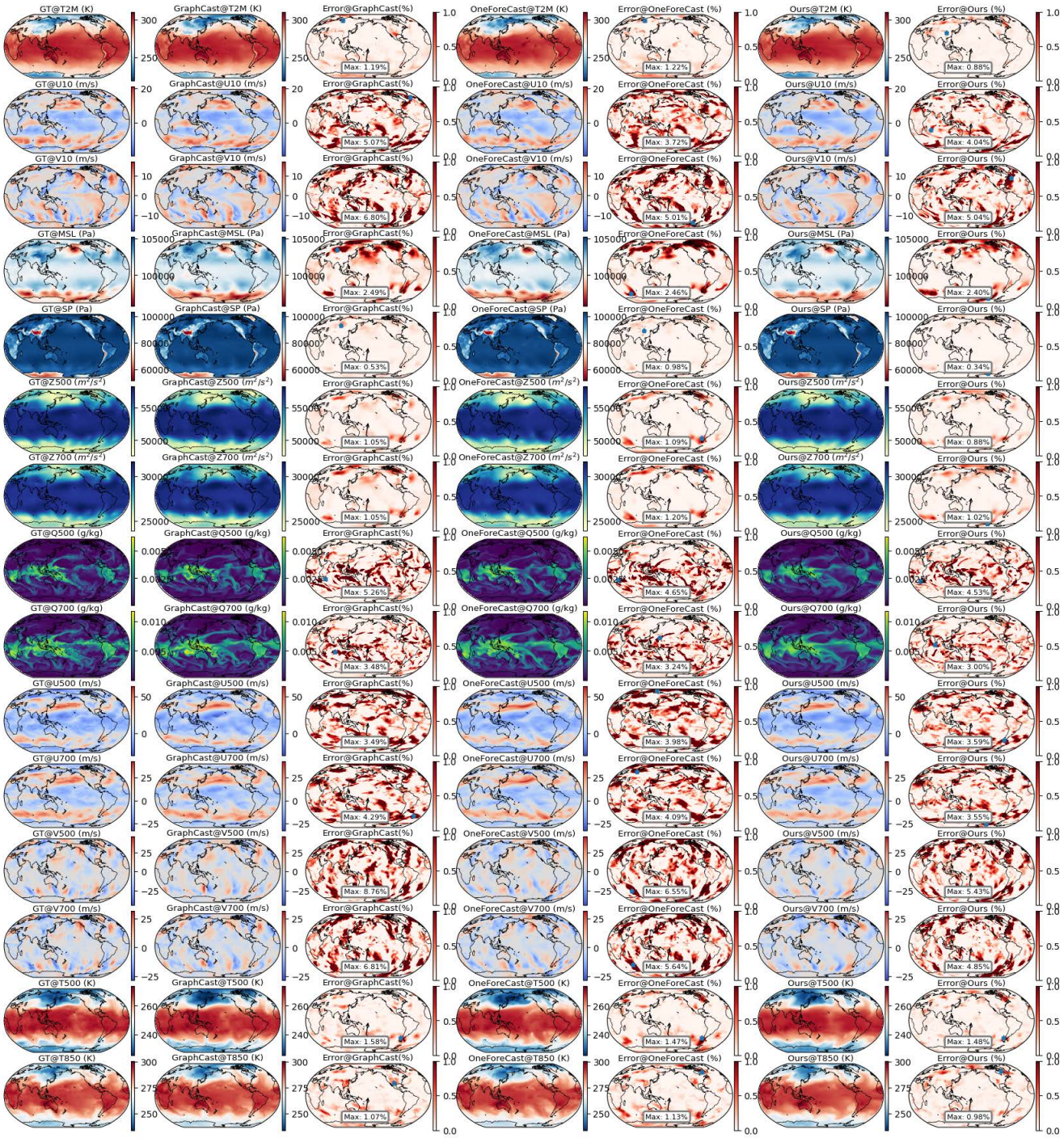


Figure 44. 10-day forecast results of global weather among different models.