Variables-Adaptive Mixture of Experts for Incremental Weather Forecasting Supplementary Material

I. Overview

We provide additional analyses and experiments of our proposed method. In particular:

1. In Sec. II, we introduce the metric of this work.
2. We compare the training times with several top-tier methods in Sec. III.
3. In Sec. IV, we provide the ablation analysis of our VA-MoE.
4. We visualize some predicted cases in Sec. V.

II. metric

This method uses Root-Mean-Square Error (RMSE) to evaluate the performance of the proposed method. RMSE calculates the deviation between the predicted results and the ground truth, with lower values indicating better performance. The metric is formulated as:

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\mathrm{X}}^{t+1} - \mathrm{X}^{t+1})^2}{n}}, \tag{1}$$

where $\hat{\mathrm{X}}^{t+1}$ and X^{t+1} denote the predicted result and the ground-truth, respectively.

III. Training Time

In Tab. 1, we present a comparison of the training times among various public methods, namely Fengwu [2], FourCastNet [3], GraphCast [4], and Pangu-Weather [1]. Notably, when contrasted with numerous top-tier methods, the computational cost of training VA-MoE is significantly lower than that of FengWu [2], GraphCastNet [4], and Pangu-Weather [1], while remaining on par with FourCastNet [3].

During the incremental stage, VA-MoE incurs only one-third of the computational cost required during the training phase, offering a substantial reduction in the computational workload associated with training from scratch.

Model	GPUs	Training Time
Fengwu[2]	32 A100s	17 days
FourCastNet[3]	64 A100s	16 hrs
GraphCast[4]	32 TPUv4s	4 weeks
Pangu-Weather[1]	192 V100s	64 days
VA-MoE	32 A100s	6 days
VA-MoE(IL)	32 A100s	2 days

Table 1. Comparative Analysis of Training Times and Hardware Specifications for Deep Learning Models.

IV. Ablation Study

In this section, we introduce various ablation studies conducted on our VA-MoE model, focusing on training paradigm, loss function, and channel number in the TopK selection layer.

Training Paradigm. During the incremental training stage, our model transitions from learning upper-air variables over a 40-year period to incorporating surface variables from a subsequent 20-year dataset. As shown in Sec. III, we compare the performance of VA-MoE with incremental learning (VA-MoE(IL)) against a baseline VA-MoE trained solely on the 20-year dataset across five surface variables. Despite both models utilizing the same 20-year dataset, VA-MoE(IL) consistently outperforms the baseline VA-MoE across all evaluation metrics. This demonstrates the effectiveness of our incremental learning framework in generalizing to new datasets while maintaining robust performance.

Loss Function. Our method employs a composite loss design in the loss function, incorporating both reconstruction loss and predicted loss. As shown in Tab. 3, by combining these losses and introducing a hyper-parameter λ to balance them, we achieve optimal performance with $\lambda = 0.1$, surpassing alternatives such as $\lambda = 0$ and $\lambda = 0.2$ across all settings, including 6h, 72h, and 120h predictions for the 5 upper-air variables. The experimental results demonstrate that the reconstruction loss makes sense in our structure. For the effect of this loss, we think that the reconstruction branch adjusts the Encoder and Decoder modules for variable reconstruction, while the transformer blocks focus on learning the distribution of weather variables.

In terms of dynamic loss weight, a comparison between L2Loss and Dynamic L2Loss in Tab. 3 indicates that Dy-

	Dataset (years)	**T2M** $(K)\downarrow$ 6h	72h	120h	**U10** $(m/s)\downarrow$ 6h	72h	120h	**V10** $(m/s)\downarrow$ 6h	72h	120h	**MSL** $(Pa)\downarrow$ 6h	72h	120h	**SP** $(Pa)\downarrow$ 6h	72h	120h
Plain Training on **70 variables (00-20)**																
VA-MoE	2000-2020	0.77	1.38	1.88	0.72	1.89	2.93	0.70	1.95	3.03	52.7	194.9	378.1	112.6	298.7	474.8
Incremental Training from **65 Upper-Air Variables (79-20)** to **5 Surface Variables (00-20)**																
VA-MoE(IL)	2000-2020	0.73	1.17	1.57	0.54	1.58	2.49	0.55	1.63	2.57	30.0	148.8	304.7	60.6	171.4	314.8

Table 2. Ablation Study on Incremental Training with 5 surface variables, i.e., T2M, U10, V10, MSL, and SP. **All experiments are in 0.25° with** 721×1440 **resolutions.**

λ	**Z500**$(m^2/s^2)\downarrow$ 6h	72h	120h	**Q500**$(\times e^{-3}, g/kg)\downarrow$ 6h	72h	120h	**U500**$(m/s)\downarrow$ 6h	72h	120h	**V500**$(m/s)\downarrow$ 6h	72h	120h	**T500**$(K)\downarrow$ 6h	72h	120h
	L2Loss														
0.1	34.22	236.4	457.7	0.18	0.61	0.80	0.92	4.09	6.41	0.90	4.26	6.75	0.34	1.35	2.23
	Dynamic L2Loss														
0	30.36	197.0	406.0	0.22	0.60	0.78	1.08	3.71	5.91	1.09	3.83	6.21	0.32	1.19	2.02
0.1	20.59	139.02	302.13	0.18	0.49	0.62	0.91	3.02	4.76	0.91	3.08	4.97	0.27	0.92	1.59
0.2	29.94	224.2	435.7	0.19	0.60	0.79	0.97	3.86	6.07	0.97	3.92	6.32	0.32	1.35	2.22

Table 3. The ablation results of loss function on the ERA5 dataset with the VA-MoE method for upper-air variables prediction task. **All experiments are in 1.5° with** 128×256 **resolutions.**

namic L2Loss consistently outperforms L2Loss across all 15 metrics, highlighting the significant performance enhancement brought about by dynamic loss weight in the VA-MoE model.

Number of Selected Channels. Within the TopK selection layer, the number of selected channels serves as a crucial hyper-parameter governing correlated information and computational complexity. Comparing different channel numbers, including the original $channel$, $channel/2$, and $channel/4$, reveals that when $K = channel/2$, VA-MoE excels in predicting the 5 upper-air variables at 6h, 72h, and 120h intervals. Setting $K = channel/2$ as our baseline model, reducing the selected channels to $channel/4$ results in a performance decline due to the loss of information. Moreover, when $K = channel$, the model's performance remains subpar compared to the baseline, indicating that irrelevant information negatively impacts the original model's efficacy.

V. Visualization

As illustrated in Fig. 1, we visualize the gate weights in the last block of six experts, which include the variables Z, Q, U, V, T, and SV. The visualization reveals that different channels are activated in different experts, indicating that the experts develop affinities for distinct variables throughout the training process.

This section also includes the visualizations of the prediction results for surface variables in Fig. 2 and upper-air variables in Fig. 3 at 6h, 72h, and 120h periods. Among the surface variables, namely T2M, U10, V10, MSL, and SP, the absolute errors for the 6h, 72h, and 120h predictions do not exceed 1.7%, 3.7%, and 5.0%, respectively, across

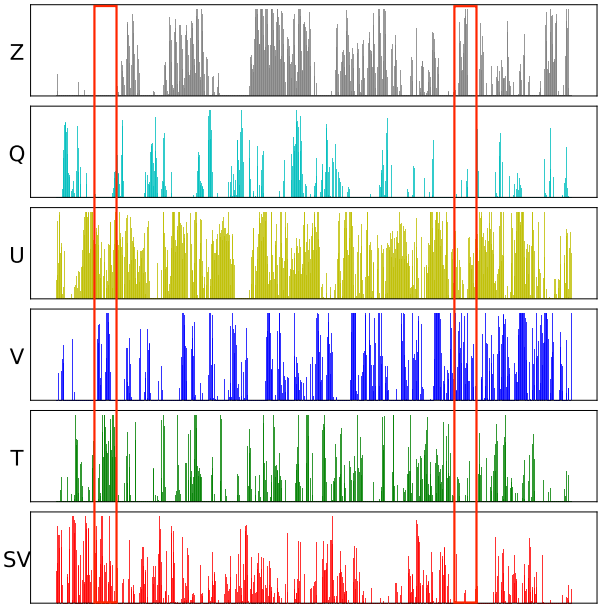


Figure 1. Frequency for selected Channels with different variables. 'SV' denotes the surface variables.

Top-K	**Z500**(m^2/s^2) ↓			**Q500**($\times e^{-3}, g/kg$) ↓			**U500**(m/s) ↓			**V500**(m/s) ↓			**T500**(K) ↓		
	6h	72h	120h	6h	72h	120h	6h	72h	120h	6h	72h	120h	6h	72h	120h
channel	26.95	198.3	405.4	0.20	0.57	0.75	0.98	3.60	5.81	0.99	3.69	6.06	0.30	1.18	2.04
channel/2	20.59	139.02	302.13	0.18	0.49	0.62	0.91	3.02	4.76	0.91	3.08	4.97	0.27	0.92	1.59
channel/4	31.30	227.1	432.8	0.19	0.61	0.79	0.98	3.95	6.13	0.97	4.08	6.45	0.34	1.40	2.24

Table 4. The ablation results of selected channels' number in the TopK selection layer with the VA-MoE method for upper-air variables prediction task. 'channel', 'channel/2', and 'channel/4' denote selecting all, half, and one-fourth of the original channels, respectively. **All experiments are in 1.5°with** 128×256 **resolutions.**

the 5 variables. Similarly, for the upper-air variables, Z500, Q500, U500, V500, and T500, the absolute errors for the same prediction periods are within 2.0%, 4.5%, and 5.4% for the 5 variables. These visualized results underscore the exceptional performance of our model in accurately predicting both surface and upper-air variables.

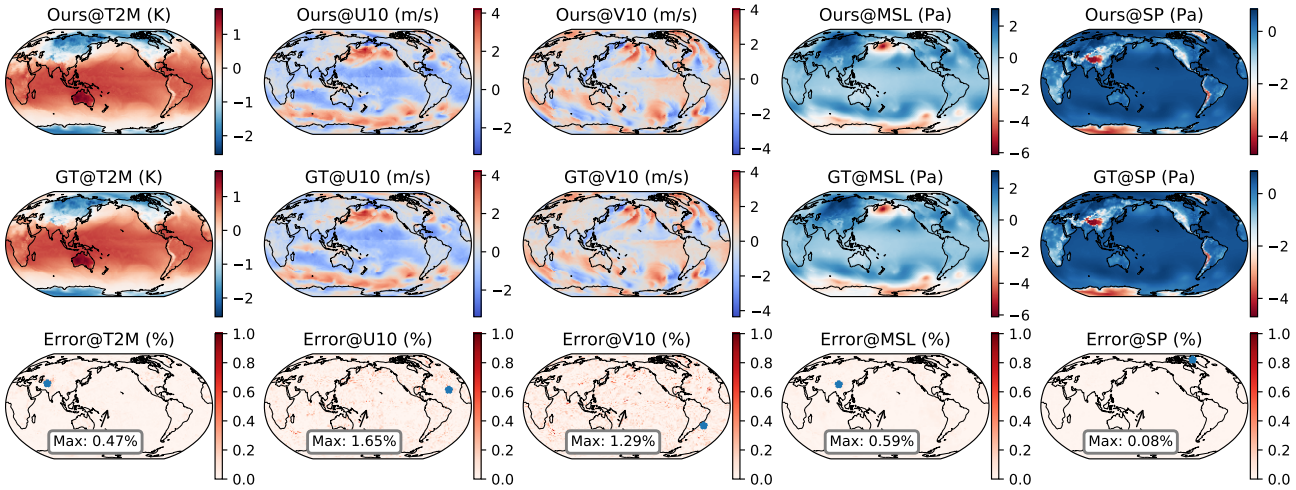
References

[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. 1

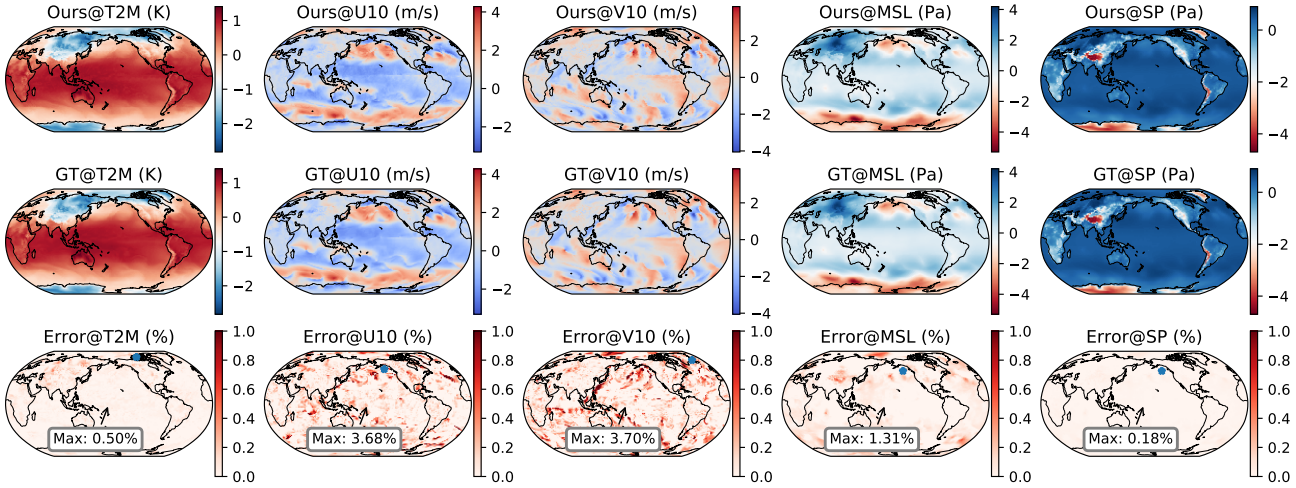
[2] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv:2304.02948*, 2023. 1

[3] 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. 1

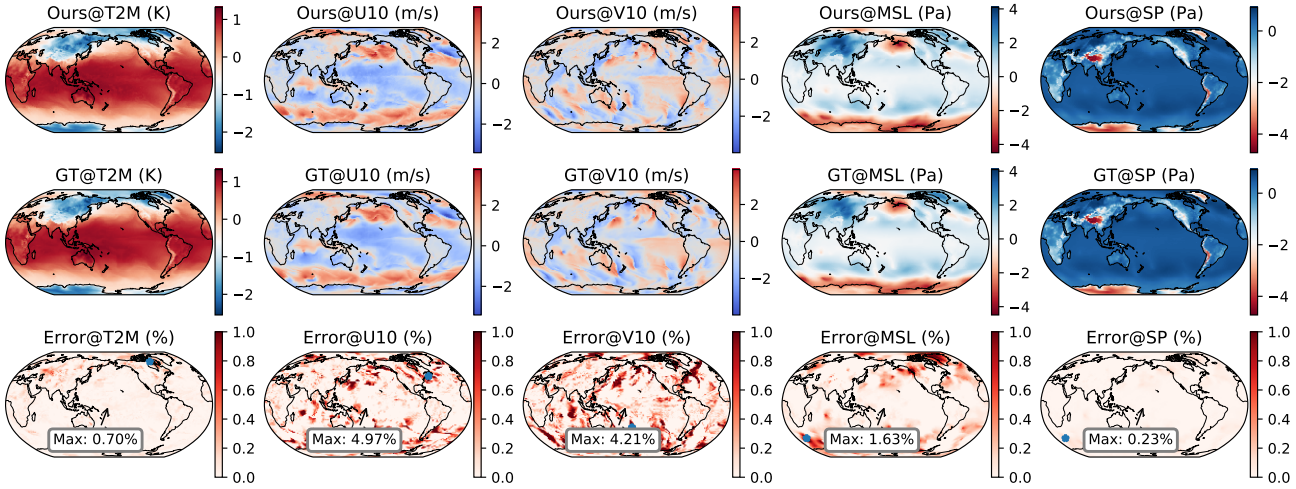
[4] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. 1



(a) 6h Prediction for Surface Variables

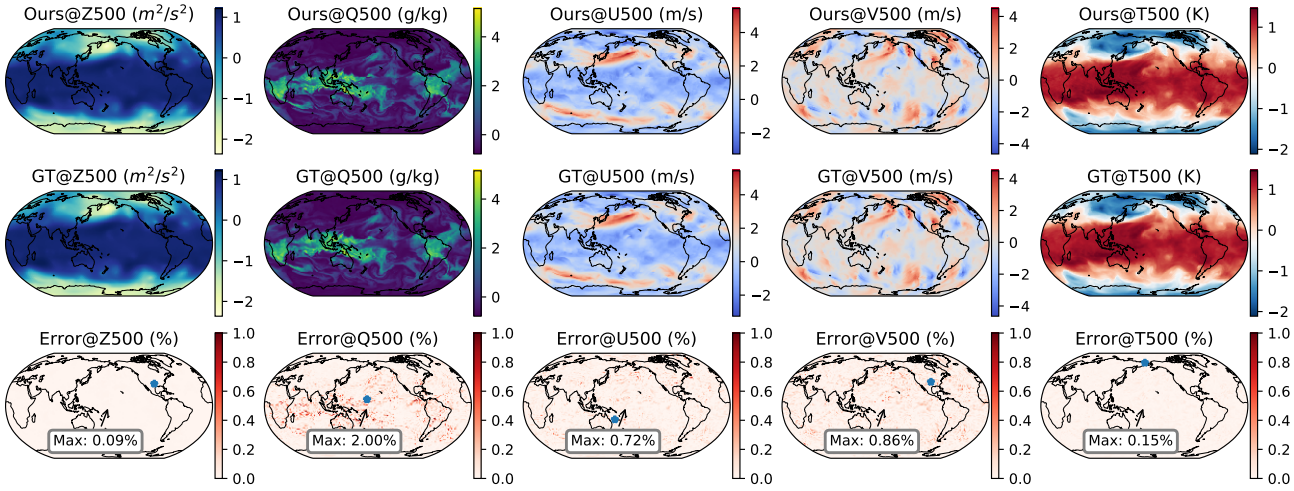


(b) 72h Prediction for Surface Variables

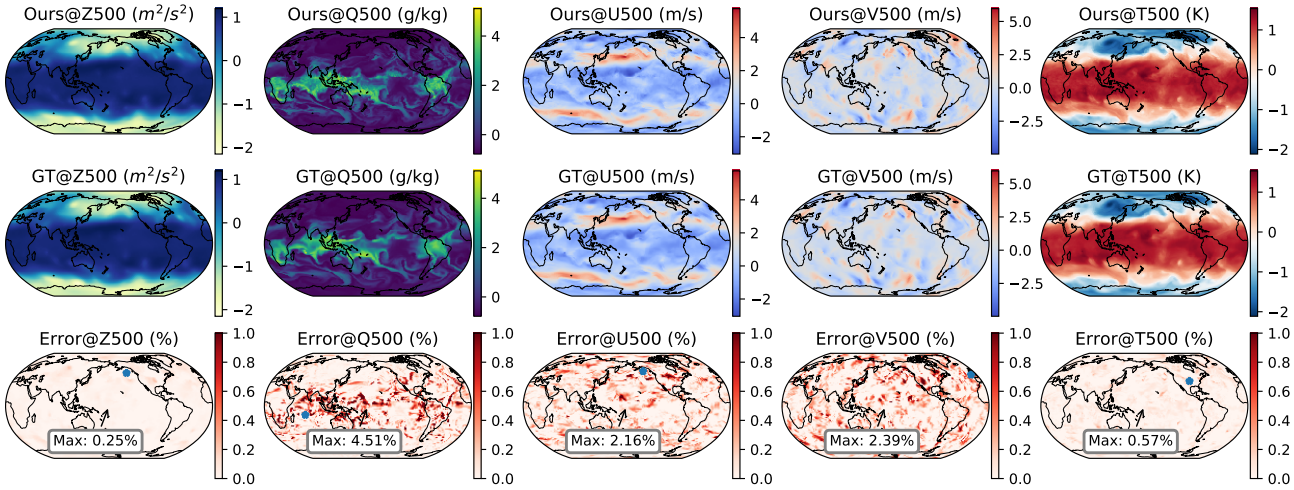


(c) 120h Prediction for Surface Variables

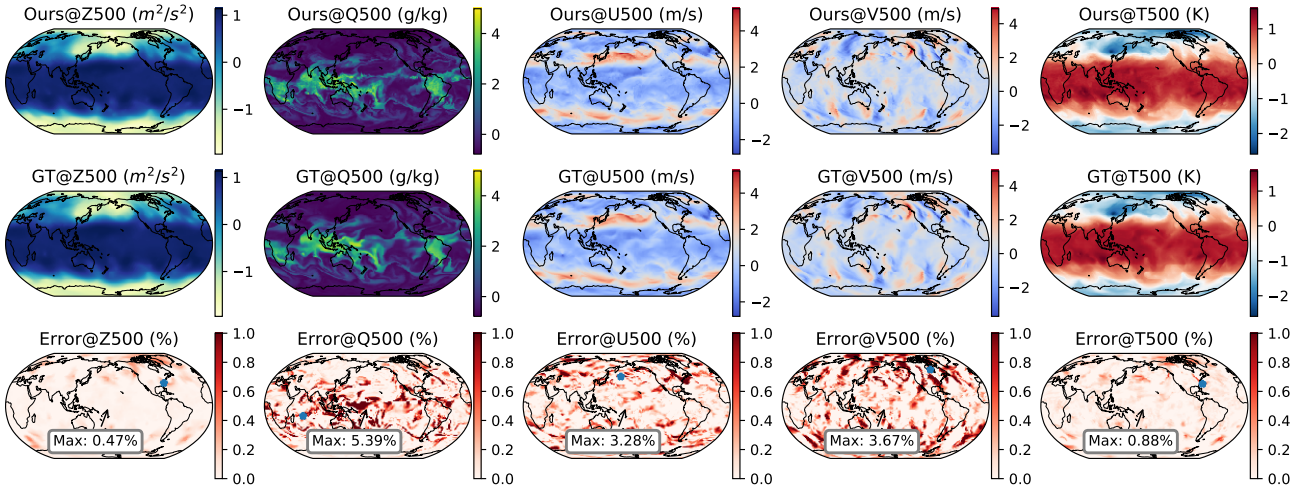
Figure 2. Visualization of the 6h, 72h, and 120h prediction results of 5 surface variables, including T2M, U10, V10, MSL, and SP. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.



(a) 6h Prediction for Upper-Air Variables

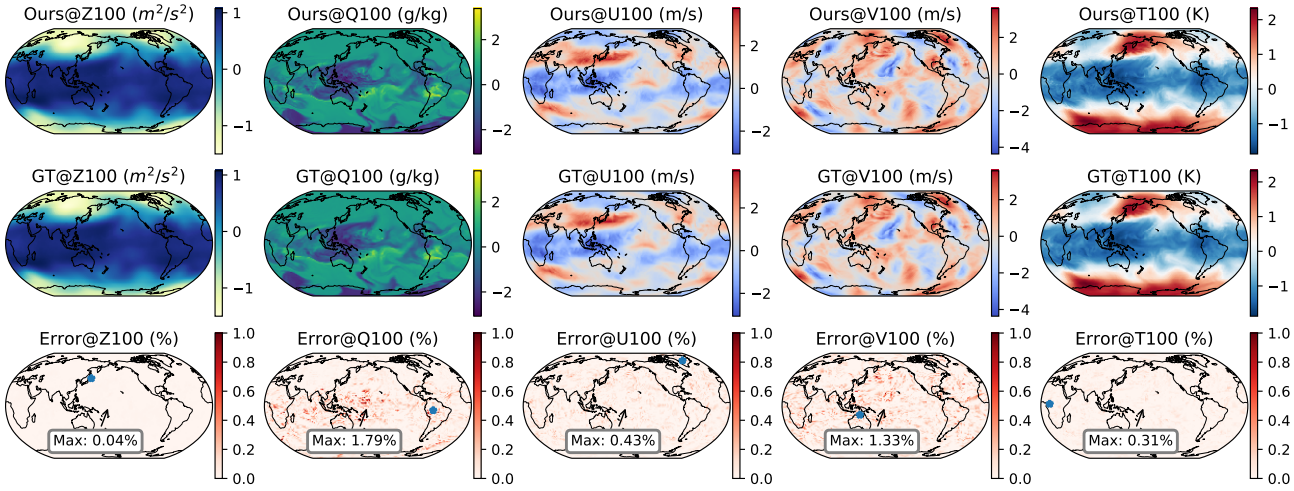


(b) 72h Prediction for Upper-Air Variables

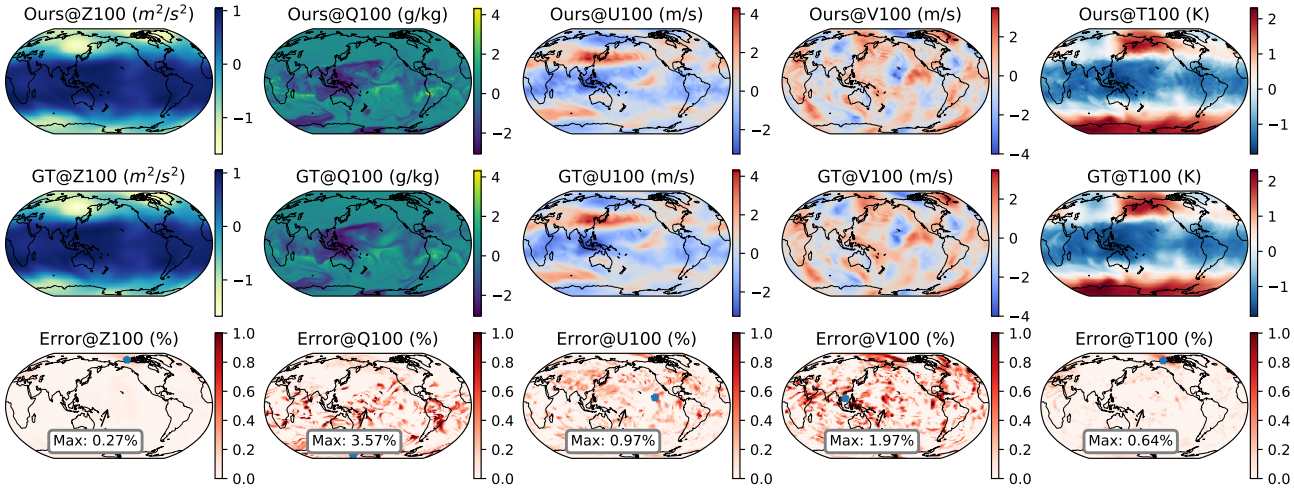


(c) 120h Prediction for Upper-Air Variables

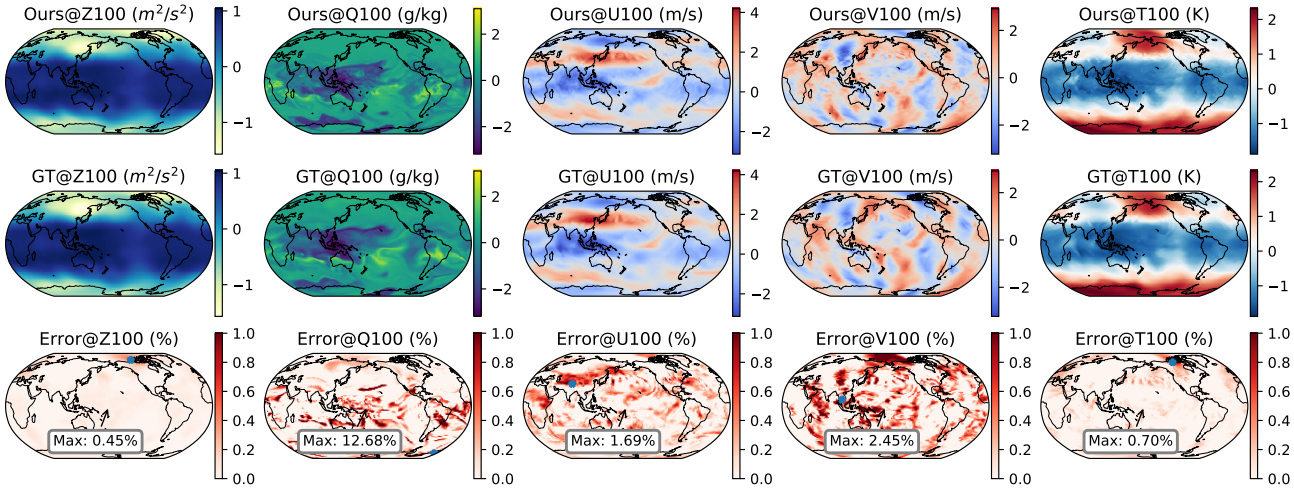
Figure 3. Visualization of the 6h, 72h, and 120h prediction results of 5 upper-air variables, including Z500, Q500, U500, V500, and T500. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.



(a) 6h Prediction for Upper-Air Variables

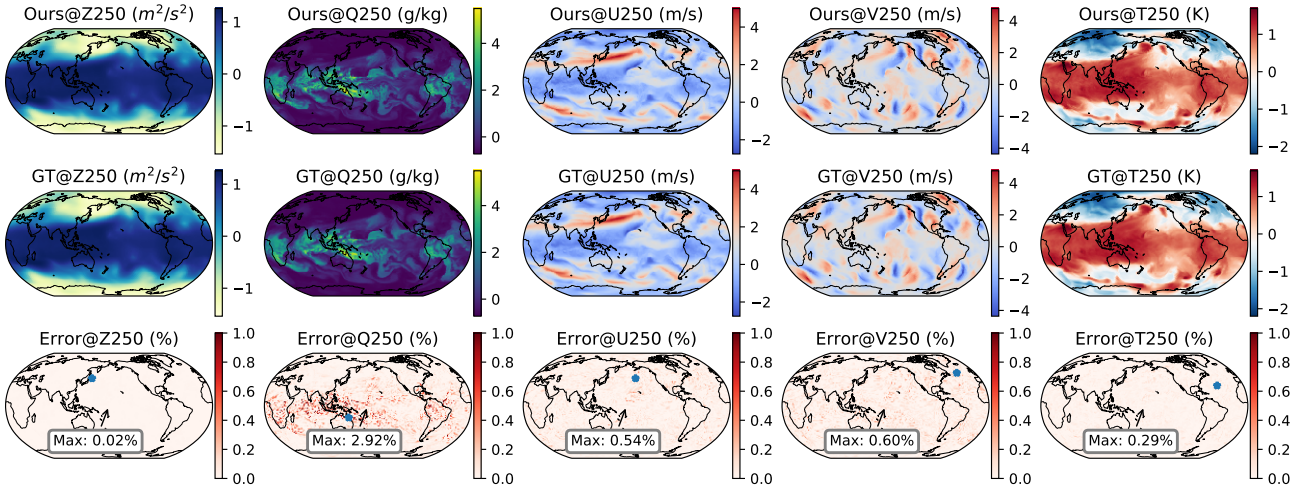


(b) 72h Prediction for Upper-Air Variables

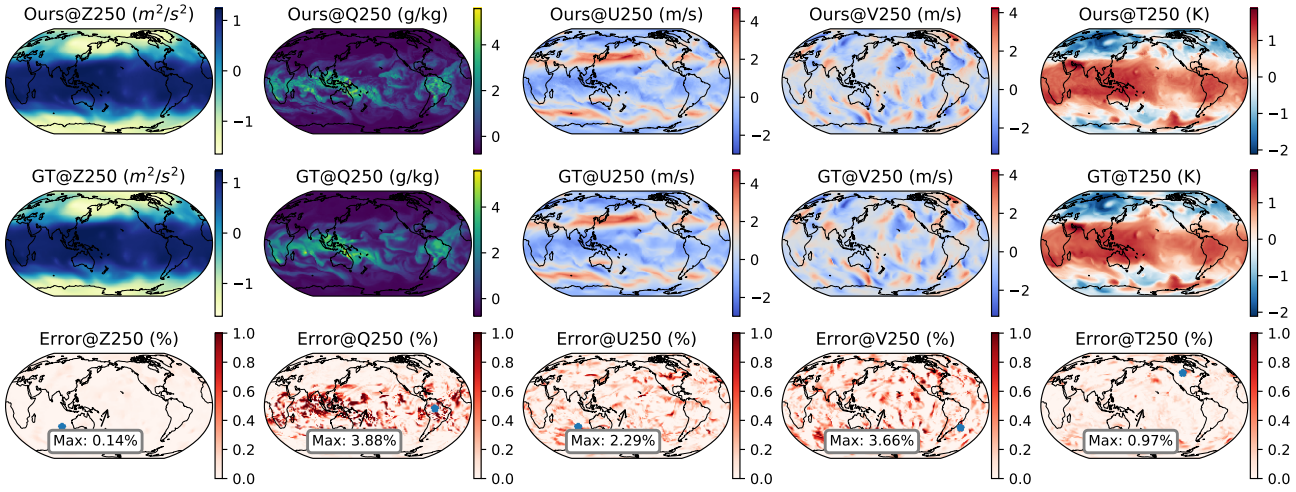


(c) 120h Prediction for Upper-Air Variables

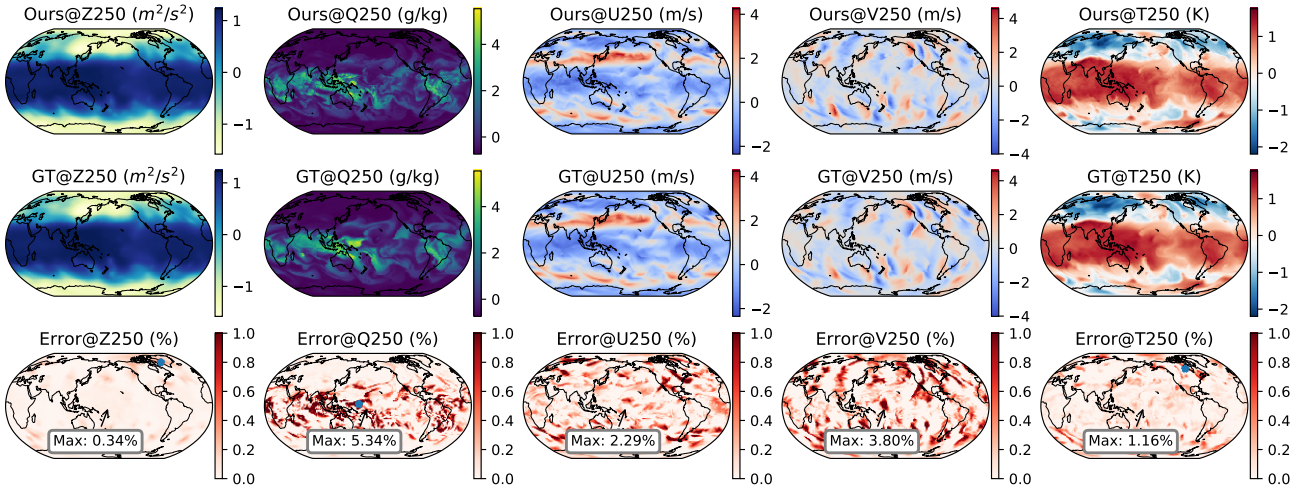
Figure 4. Visualization of the 6h, 72h, and 120h prediction results of 5 upper-air variables, including Z100, Q100, U100, V100, and T100. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.



(a) 6h Prediction for Upper-Air Variables

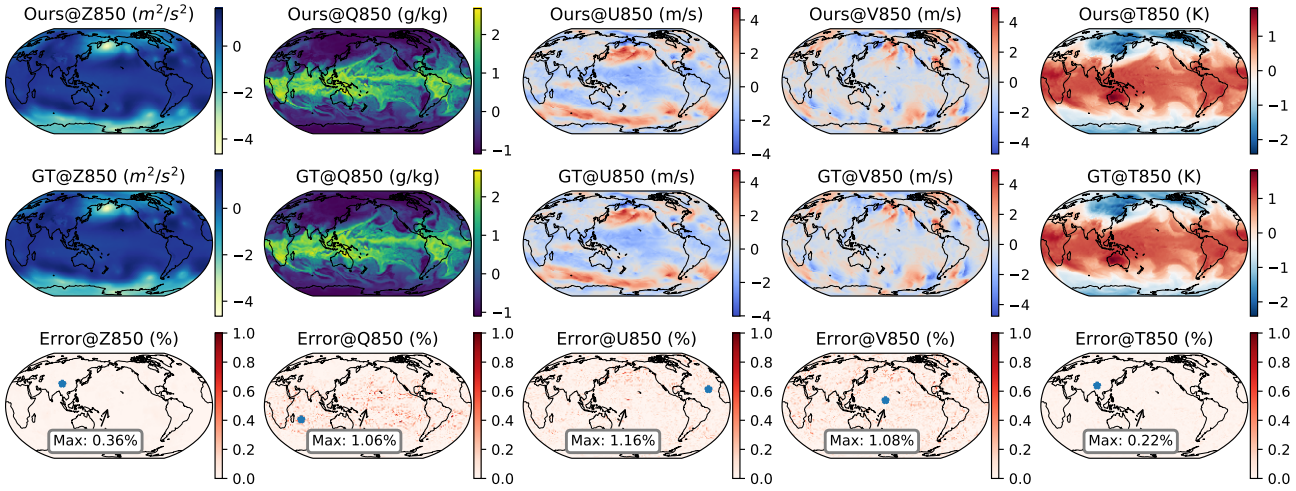


(b) 72h Prediction for Upper-Air Variables

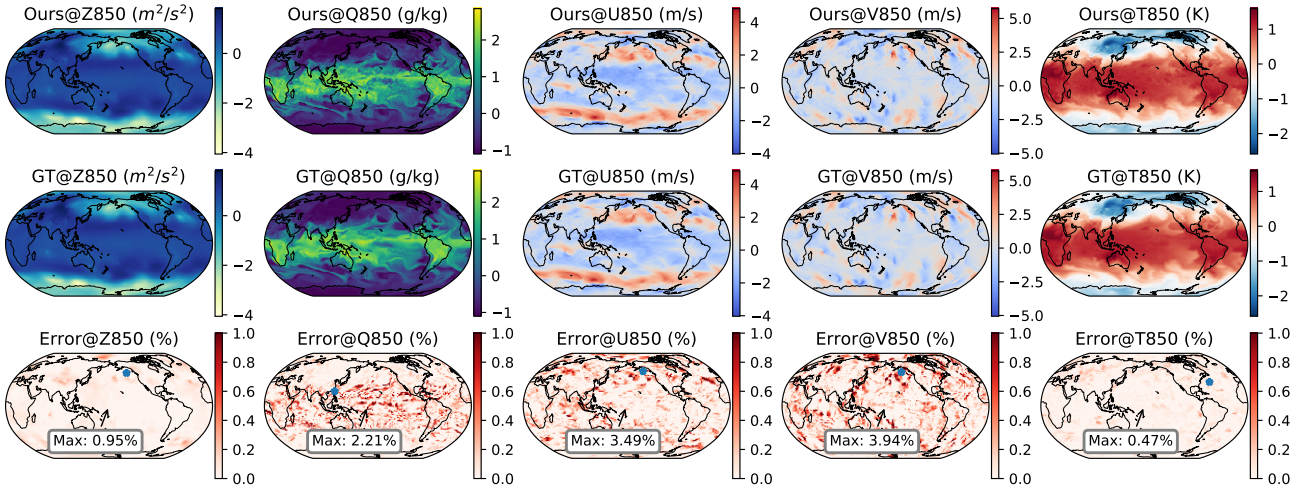


(c) 120h Prediction for Upper-Air Variables

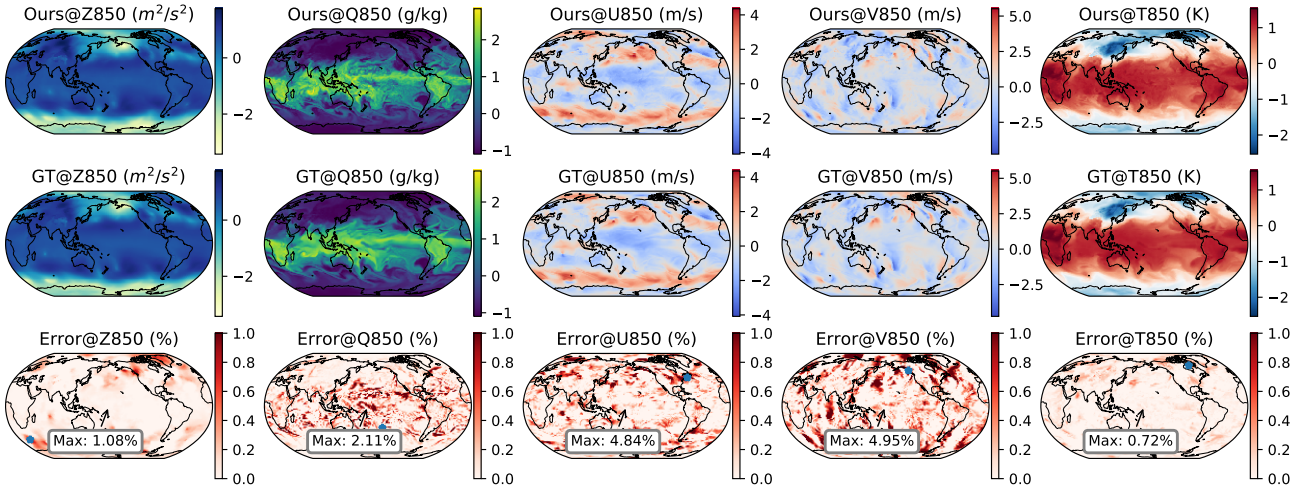
Figure 5. Visualization of the 6h, 72h, and 120h prediction results of 5 upper-air variables, including Z250, Q250, U250, V250, and T250. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.



(a) 6h Prediction for Upper-Air Variables

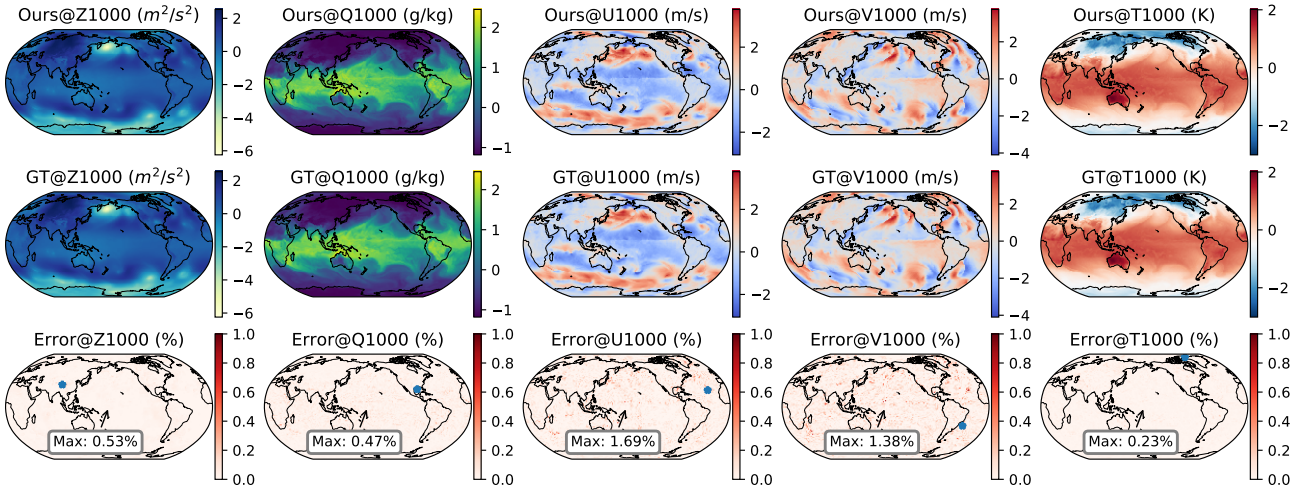


(b) 72h Prediction for Upper-Air Variables

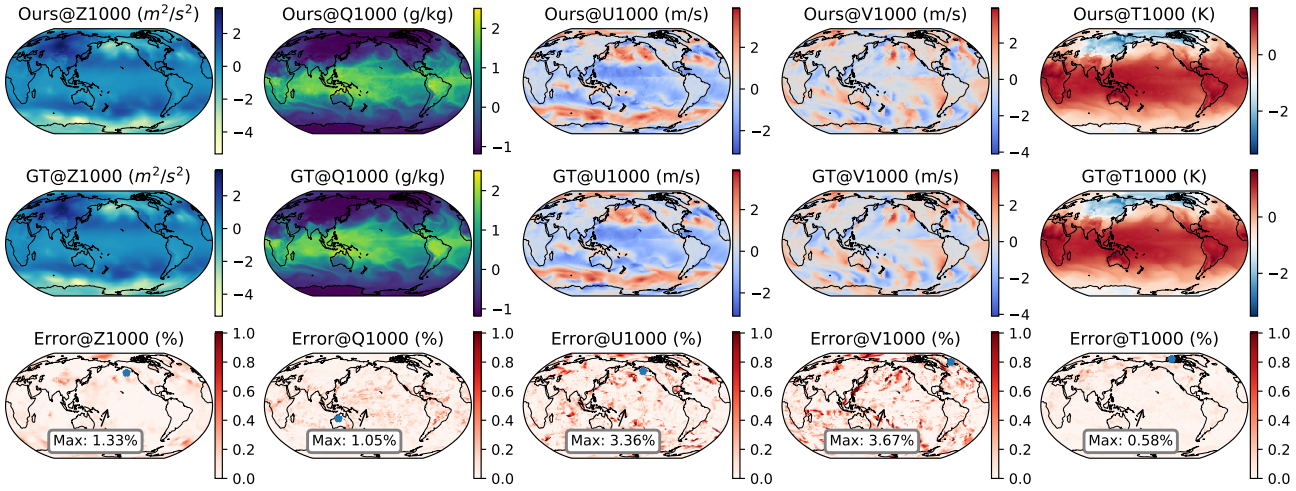


(c) 120h Prediction for Upper-Air Variables

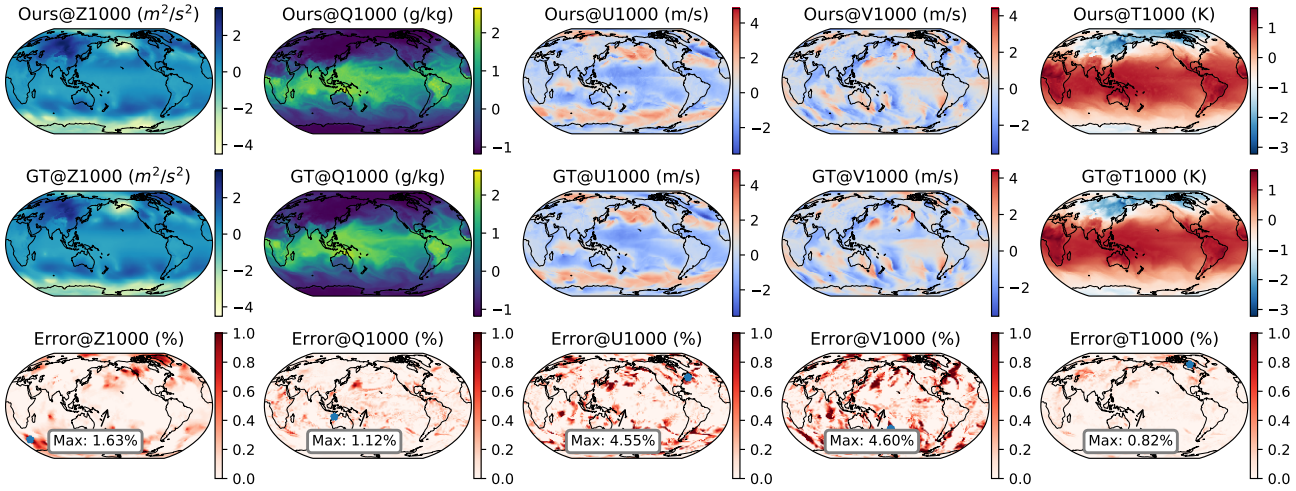
Figure 6. Visualization of the 6h, 72h, and 120h prediction results of 5 upper-air variables, including Z850, Q850, U850, V850, and T850. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.



(a) 6h Prediction for Upper-Air Variables



(b) 72h Prediction for Upper-Air Variables



(c) 120h Prediction for Upper-Air Variables

Figure 7. Visualization of the 6h, 72h, and 120h prediction results of 5 upper-air variables, including Z1000, Q1000, U1000, V1000, and T1000. 'Ours' denotes VA-MoE, 'GT' denotes Ground Truth, and 'Error' denotes the absolute error between the predicted results and ground truth.