VA-MoE: Variables-Adaptive Mixture of Experts for Incremental Weather Forecasting

Hao Chen[1†] Tao Han[1†] Song Guo[1*] Jie Zhang[1] Yonghan Dong[2] Yunlong Yu[3] Lei Bai[4]
[1]Hong Kong University of Science and Technology (HKUST) [2]Huawei Technologies Ltd.
[3]Zhejiang University [4]Shanghai AI Laboratory

2025/9/9

Contents

- **Motivation**
- **Main Contributions**
- **Methodology**
- **Experiments & Results**
- **Conclusion**

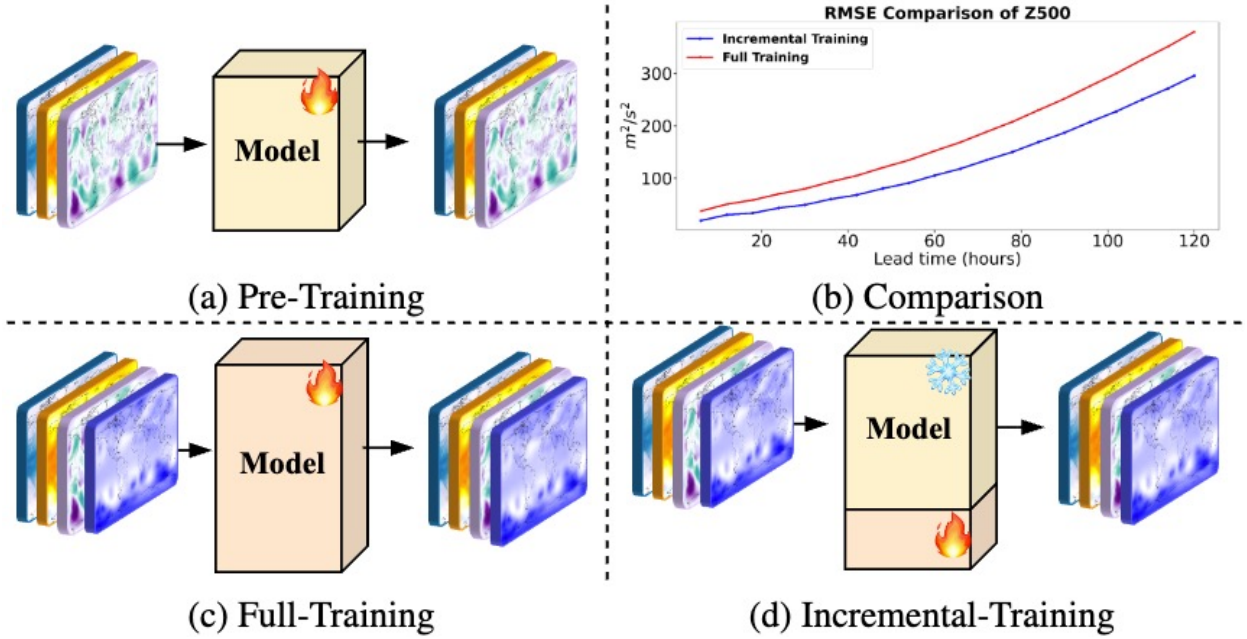
Motivation

- **Drawbacks of previous works in Earth System:**
  1. Traditional weather prediction models often struggle with **exorbitant computational expenditure** and the need to **continuously update forecasts** as new observations arrive.
  2. For instance, the upper-air variables (e.g., temperature profiles) are sparse and sampled via radiosondes/satellites, while the surface variable (e.g., precipitation, wind) are dense but updated in near-real-time.
  3. This asynchrony poses significant challenges: when introducing new variables (e.g., satellite-derived aerosol data), **existing models must be entirely retrained from scratch**, incurring prohibitive computational costs.

Motivation

- **Previous works & Ours:**
  1. **Previous Works:** Entirely retraining the whole model from scratch when new variable comes .
  2. **Ours:** Incrementally training limited number of parameters from the pretrained model.

- **Two strategies on Forecasting:**
  a) Pre-Training;
  b) RMSE Comparison of Z500;
  c) Full Training;
  d) Incremental Training;



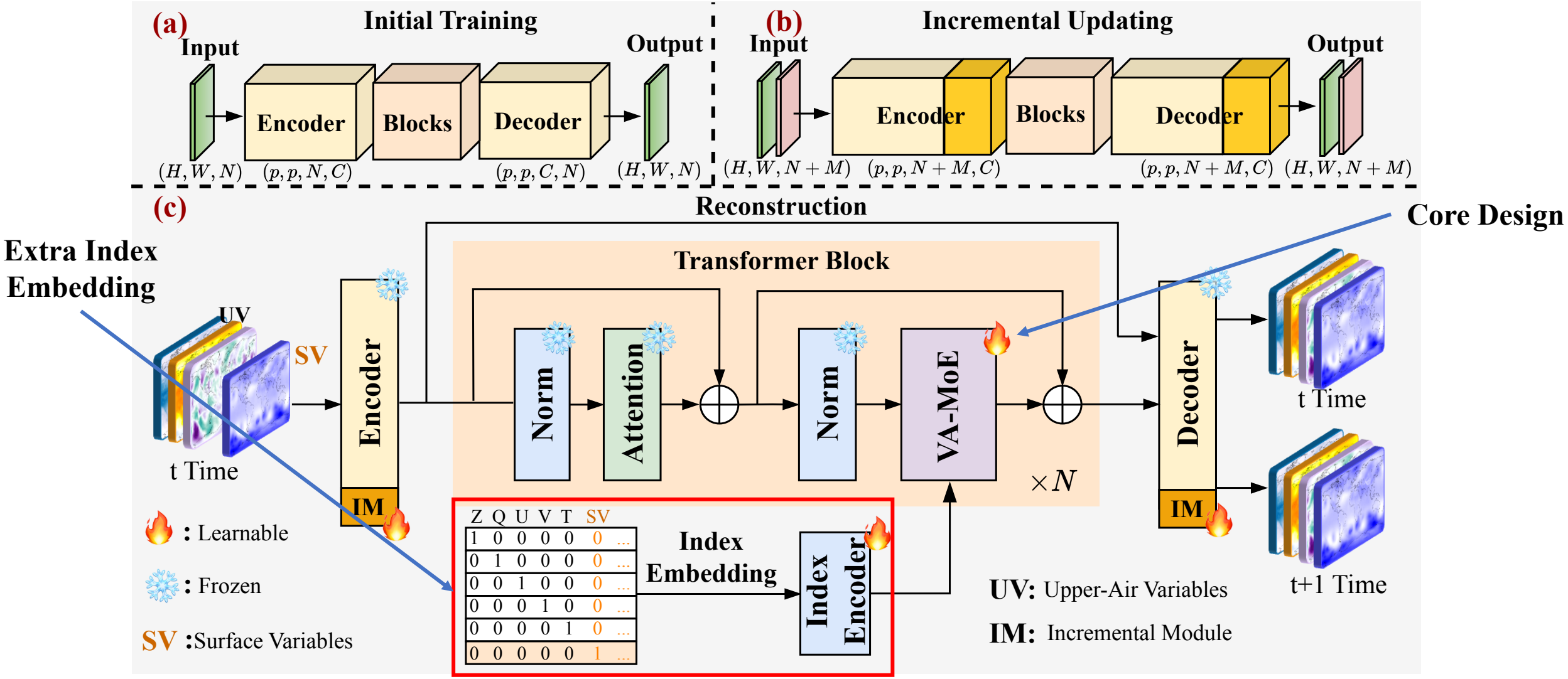
Main Contribution

1. This work initiates the research on **incremental learning paradigms for weather forecasting**. We propose the quantitative benchmark to evaluate the performance.

2. We present **Variables-Adaptive Mixture of Experts (VA-MoE),** the first framework tailored for incremental atmospheric modeling. VA-MoE achieves **expert specialization** through contextual variable activation driven by **variable index embeddings**, enabling dynamic assignment of experts to variables during both training and inference.

3. Extensive experiments on the ERA5 dataset demonstrate that VA-MoE achieves **comparable performance** for surface variables, while delivering **superior accuracy** in upper-air variables.

Methodology – Main Framework



Methodology – Incremental Setting

1. The weather variables X^t are divided into two sets: (i) the initial training variables at t time $\mathrm{X}_{\mathrm{h}}^{\mathrm{t}} \in \mathbb{R}^{H \times W \times N}$, and, (ii) the incremental variables at t time $\mathrm{X}_{\mathrm{sv}}^{\mathrm{t}} \in \mathbb{R}^{H \times W \times M}$.

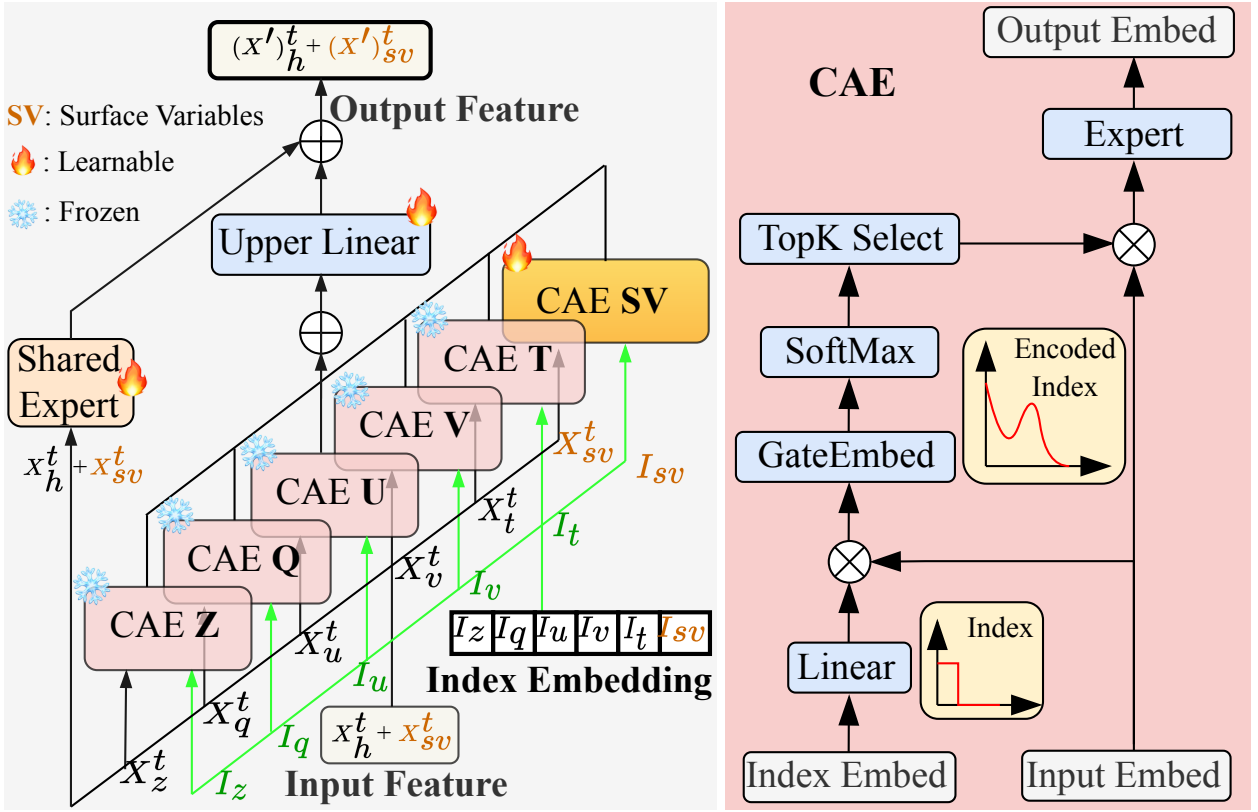
2. When new variables are incrementally introduced to the model, **specialized experts are dynamically integrated into the transformer blocks to process these variables**. The Encoder, Decoder, and the previous Experts are frozen. Only the newly-added experts and Index Encoder are trained.

Methodology – VAMoE Structure

- VA-MoE introduces index embedding $\mathbf{I_Z}$ within transformer blocks, which **dynamically guides experts in learning hierarchical relationships** between atmospheric variables.
- As new variables $\mathrm{X_h^t}$ are incrementally integrated, corresponding experts are added to the transformer architecture and optimized via **index-based affinity assignments**.

$$\mathrm{I_Z^{topk}}, \mathrm{W_Z^{topk}} = \mathrm{TOP_k}(\mathrm{Softmax}\,(\mathrm{MLP_Z}(\mathrm{X_h^t} \odot \mathrm{I_Z})))$$

$$\mathrm{X_h^{t,\,selected}} = \mathrm{W_Z^{topk}} * \mathrm{Select_Z}(\mathrm{X_h^t}, \mathrm{I_Z^{topk}})$$



llustration of the VA-MoE. **In the left subgraph**, the input features comprise both upper-air and surface variables. **The right subgraph** details CAE module.

Methodology – Loss Function

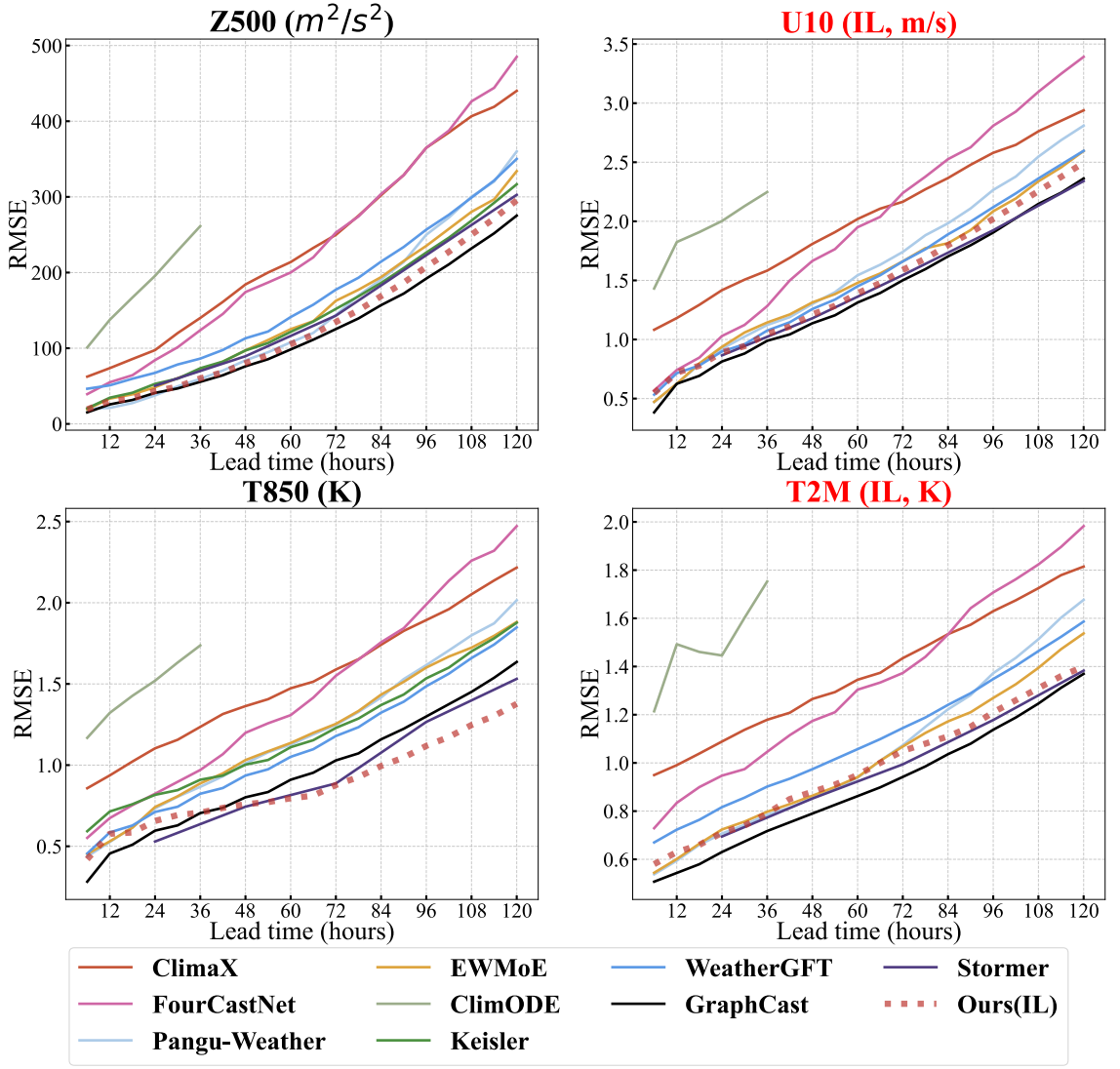
- A variable-adaptive loss function that aligns variable rates with their inherent spatiotemporal characteristics.

- **Prediction Loss:** $\mathrm{Obj}_{\mathrm{pred}} = \left(\hat{\mathrm{X}}^{\mathrm{t+1}} - \mathrm{X}^{\mathrm{t+1}}\right) \odot \left(\hat{\mathrm{X}}^{\mathrm{t+1}} - \mathrm{X}^{\mathrm{t+1}}\right) / \mathrm{e}^{\mathrm{w}} + \mathrm{w}$
- **Reconstruction Loss:** $\mathrm{Obj}_{\mathrm{recons}} = \left(\hat{\mathrm{X}}^{\mathrm{t}} - \mathrm{X}^{\mathrm{t}}\right) \odot \left(\hat{\mathrm{X}}^{\mathrm{t}} - \mathrm{X}^{\mathrm{t}}\right)$
- **Joint Loss:** $\mathrm{Obj}_{\mathrm{total}} = \mathrm{Obj}_{\mathrm{pred}} + \lambda * \mathrm{Obj}_{\mathrm{recons}}$

Experiments & Results

- **Dataset.** In this work, we conduct experiments on **ERA5**. We train the model on 40-year dataset in the initial stage and 20-year dataset in the incremental stage.

- **The right image is the comparative analysis of RMSE across 10 models for 4 variables**, including Z500 and T850 in the initial stage, as well as T2M and U10 in the incremental stage.



Experiments & Results

	Dataset (years)	Iteration ($\times 10^4$)	T2M (K) ↓			U10 (m/s) ↓			V10 (m/s) ↓			MSL (Pa) ↓			SP (Pa) ↓		
			6h	72h	120h	6h	72h	120h	6h	72h	120h	6h	72h	120h	6h	72h	120h
Plain Training																	
ViT* [13]	1979-2020	40	0.72	1.35	1.86	0.66	1.98	3.01	0.68	2.02	3.11	40.2	208.5	393.9	63.3	222.1	397.0
IFS [32]	1979-2020	40	1.09	1.38	1.74	0.96	1.87	2.78	0.99	1.93	2.87	-	-	-	-	-	-
Pangu-Weather [3]	1979-2020	40	0.82	1.09	1.53	0.77	1.63	2.54	0.79	1.68	2.65	-	-	-	-	-	-
FourCastNet [22]	1979-2020	40	0.82	1.02	1.77	0.82	2.08	3.34	0.84	2.11	3.41	-	-	-	-	-	-
ClimaX [27]	1979-2020	40	1.11	1.47	1.83	1.04	2.02	2.79	-	-	-	-	-	-	-	-	-
Graphcast [23]	1979-2020	40	**0.51**	**0.94**	**1.37**	**0.38**	1.51	2.37	-	-	-	23.4	135.2	278.2	-	-	-
Fengwu [9]	1979-2020	40	0.58	1.03	1.41	0.42	1.53	2.32	-	-	-	**23.2**	137.1	276.9	-	-	-
FuXi [10]	1979-2020	40	0.55	0.99	1.41	0.42	1.50	2.36	**0.43**	1.54	2.44	27.2	136.7	282.9	-	-	-
VA-MoE	1979-2020	40	0.57	1.03	1.42	0.43	**1.41**	**2.25**	0.44	**1.46**	**2.34**	27.5	**131.1**	**275.9**	**57.1**	**168.9**	**302.4**
Incremental Training from **65 Upper-Air Variables (79-20)** to **5 Surface Variables (79-20)**																	
VA-MoE (IL)	1979-2020	20	0.58	1.05	1.45	0.48	1.47	2.33	0.47	1.54	2.41	27.9	137.3	281.6	59.3	173.4	312.4
Incremental Training from **65 Upper-Air Variables (79-20)** to **5 Surface Variables (00-20)**																	
VA-MoE (IL)	2000-2020	10	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

Prediction performances on Incremental Training with **5 surface variables**, i.e., T2M, U10, V10, MSL, and SP. * denotes running by ourselves. All experiments are in 0.25°with 721 × 1440 resolutions.

Experiments & Results

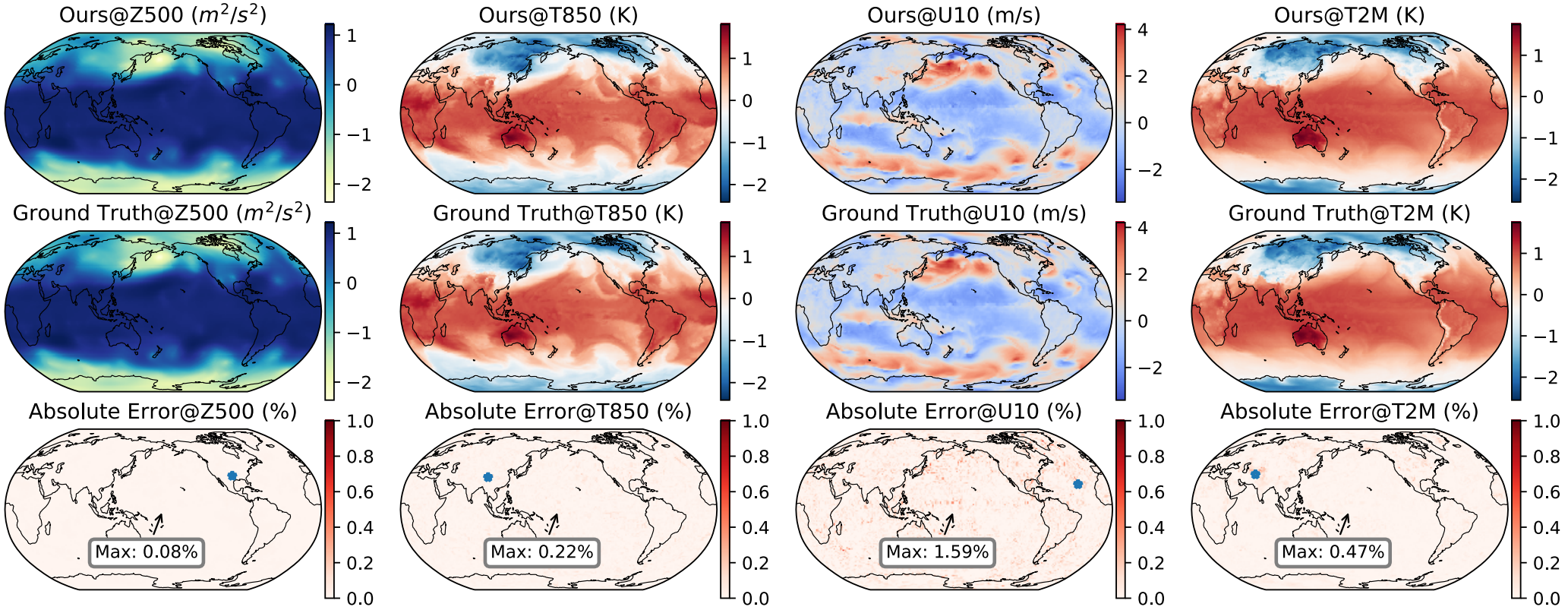
	Para. (M)	Z500(m^2/s^2) ↓ 6h	72h	120h	Q500($\times e^{-3}, g/kg$) ↓ 6h	72h	120h	U500(m/s) ↓ 6h	72h	120h	V500(m/s) ↓ 6h	72h	120h	T500(K) ↓ 6h	72h	120h
IFS [32]	-	28.31	154.08	333.96	0.31	0.61	0.75	1.43	3.23	5.12	1.40	3.58	5.64	0.36	0.98	1.70
Pangu-Weather [3]	-	24.88	167.90	391.26	0.25	0.55	0.69	0.96	3.13	4.73	0.91	3.52	5.15	0.27	0.94	1.56
Graphcast [23]	-	15.23	125.42	275.35	-	-	-	0.77	2.86	4.49	0.74	2.92	4.67	0.23	0.87	1.48
VA-MoE	665	19.28	134.63	295.52	0.17	0.49	0.62	0.84	2.99	4.71	0.84	3.04	4.89	0.25	0.76	1.36
Incremental Training from **65 Upper-Air Variables (79-20)** to **5 Surface Variables (00-20)**																
VA-MoE (IL)	137	18.23	133.14	292.63	0.17	0.49	0.61	0.84	3.01	4.74	0.84	3.51	4.93	0.25	0.76	1.37

Prediction performances on **Initial Training with 5 upper-air variables**.

	Para. (M)	Z500 (m^2/s^2) ↓ 6h	72h	120h	Q500 ($\times e^{-3}, g/kg$) ↓ 6h	72h	120h	U500 (m/s) ↓ 6h	72h	120h	V500 (m/s) ↓ 6h	72h	120h	T500 (K) ↓ 6h	72h	120h
ViT*[13]	307	33.38	209.4	517.81	0.22	0.61	1.06	1.24	3.66	6.52	1.22	3.76	7.41	0.42	1.18	2.40
ViT+MoE (light)*[30]	609	37.92	207.11	405.73	0.22	0.60	0.78	1.30	3.84	5.87	1.27	3.89	6.11	0.46	1.23	2.02
ViT+MoE*[30]	1113	28.31	169.61	356.02	0.23	0.56	0.72	1.21	3.46	5.44	1.23	3.54	5.69	0.35	1.07	1.83
VA-MoE	665	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
VA-MoE (IL)	137	20.29	138.52	301.41	0.18	0.50	0.63	0.91	3.04	4.79	0.91	3.10	5.03	0.27	0.93	1.60

Architectural impact on **5 upper-air variables** under 500 hPa. All experiments are with 128 × 256 resolutions.

Experiments & Results



6-hour visualization of upper-air and surface variables.

Conclusion

- **Our work:** In this work, we proposed incremental weather forecasting, a novel task that addresses the challenge of dynamically expanding weather models to incorporate new variables without retraining from scratch.

- **Limitation:** This work incrementally transforms the upper-air model to surface variables. It might be more meaningful to test the model's performance on larger datasets and in a wider range of scenarios, for instance, extending the atmospheric model to the ocean and soil.