DAWP: A framework for global observation forecasting via Data Assimilation and Weather Prediction in satellite observation space

Junchao Gong *
Shanghai Jiao Tong University
gjchimself@sjtu.edu.cn

Jingyi Xu *
Fudan University
jyxu22@m.fudan.edu.cn

Ben Fei †
The Chinese University of Hong Kong
Shanghai AI Laboratory
benfei@cuhk.edu.hk

Fenghua Ling
Shanghai AI Laboratory
lingfenghua@pjlab.org.cn

Wenlong Zhang
Shanghai AI Laboratory
zhangwenlong@pjlab.org.cn

Kun Chen
Shanghai AI Laboratory
chenkun@pjlab.org.cn

Wanghan Xu
Shanghai AI Laboratory
xuwanghan@pjlab.org.cn

Weidong Yang
Fudan University
wdyang@fudan.edu.cn

Xiaokang Yang
Shanghai Jiao Tong University
xkyang@sjtu.edu.cn

LEI BAI †
Shanghai AI Laboratory
bailei@pjlab.org.cn

Abstract

Weather prediction is a critical task for human society, where impressive progress has been made by training artificial intelligence weather prediction (AIWP) methods with reanalysis data. However, reliance on reanalysis data limits the AIWPs with shortcomings, including data assimilation biases and temporal discrepancies. To liberate AIWPs from the reanalysis data, observation forecasting emerges as a transformative paradigm for weather prediction. One of the key challenges in observation forecasting is learning spatiotemporal dynamics across disparate measurement systems with irregular high-resolution observation data, which constrains the design and prediction of AIWPs. To this end, we propose our DAWP as an innovative framework to enable AIWPs to operate in a complete observation space by initialization with an artificial intelligence data assimilation (AIDA) module. Specifically, our AIDA module applies a mask multi-modality autoencoder (MMAE) for assimilating irregular satellite observation tokens encoded by mask ViT-VAEs. For AIWP, we introduce a spatiotemporal decoupling transformer with cross-regional boundary conditioning (CBC), learning the dynamics in observation space, to enable sub-image-based global observation forecasting. Comprehensive experiments demonstrate that AIDA initialization significantly improves the rollout and efficiency of AIWP. Additionally, we show that DAWP holds promising potential to be applied in global precipitation forecasting. Code will be available at this github repo.

*Equal Contribution

†Corresponding Authors: Ben Fei (benfei@cuhk.edu.hk) and Lei Bai (bailei@pjlab.org.cn)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Weather prediction is a critical task that significantly impacts various socioeconomic aspects, including transportation, agriculture, and public safety. Traditional numerical weather prediction (NWP) systems rely on intricate human-designed workflows [1, 2], such as numerical assimilation systems and physical solvers, to generate global precipitation predictions.

Recently, transformative progress has been made by artificial intelligence weather prediction (AIWP) models. These AIWP models now achieve forecast skill scores comparable to or even surpassing those of leading physics-based NWP systems [3, 4, 5, 6]. To learn the atmospheric dynamics, reanalysis products are widely used [7, 8, 9, 10].

However, reanalysis data, generated by numerical data assimilation, introduce intrinsic limitations in AIWP models built upon them. (I) **Data Assimilation Biases:** (Re)analysis products are synthesized by numerical data assimilation (DA), where direct observations are blended with a physics-based forecast. During DA, information loss of direct observation occurs due to the limited utilization of raw observational data and the preprocessing that resamples observations to regular grids of reanalysis data format with finite resolution [11, 12, 13, 14]. Additionally, the incomplete physical process modeling and uncertainty parameterizations in physics-based forecast systems also introduce biases, which could hinder learning the actual dynamics of the atmosphere [15]. (II) **Temporal Discrepancies:** The temporal lag between direct observation acquisition (nearly real-time) and analysis data generation (up to six hours) severely degrades the quick response ability of AIWP models [2, 16]. These limitations may be further exacerbated by the discrepancy between real-world observation space and physical forecasting space which is required by NWP systems to implement dynamic equations of atmospheric [17]. As AIWP models do not require physical solvers to predict the evolution of the atmosphere, there is potential for them to directly predict atmosphere states in real-world observation space.

Artificial Intelligence Direct Observation Prediction (AI-DOP) is emerging as a transformative, data-driven approach with the potential to overcome the limitations in reanalysis-driven AIWP methods. The key challenge of AI-DOP is learning the spatiotemporal dynamics not only within a single observation source but also across disparate measurement systems, given the irregular and high-resolution observation data [17]. The spatiotemporal modeling approaches are restricted to learn the relationships between different observations with irregular data. [17] applies a transformer with mask tokens to reformat multiple observations into regular ones. Further, [18] uses a graph encoder to flexibly encode different measurements into a uniform latent representation for latent forecasting with a naive transformer backbone. In addition, as AI-DOP requires being generalized to any location grids given observations with variable missing values [17], the dense forecasts and sparse observation inputs result in **input-output distribution shift** in rollout as shown in (b) and (c) of Figure 1. Motivated by these questions, we argue that training AIWP models in **a uniform observation space** where input and output are both regular grid data.

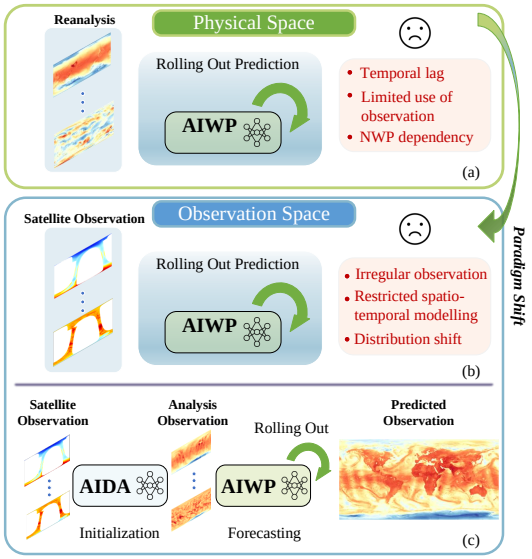

Figure 1: A paradigm shift from physical space to observation space. Our DAWP is illustrated in (c).

We propose our DAWP, an AI-DOP system composed of an observation space data assimilation (AIDA) module and an AIWP module, to learn the spatiotemporal dynamics between various satellite observations with irregular and high-resolution characteristics. To begin with, a transformer VAE encoder/decoder is designed with observation masks to regionally encode high-resolution irregular direct earth observations for efficient I/O and computation. Then, we implement a sub-image-based AIDA module for observation space data assimilation by a multi-modal masked autoencoder with en-

coded observation tokens. By learning the spatiotemporal correlations between various observations in the AIDA initialization, irregular observations are transformed into a uniform completed observation space. In this imputed space, we train our sub-image-based AIWP module with Cross-regional Boundary Conditioning (CBC), which could forecast observations with a global state cache providing atmospheric states of neighbours. Finally, a combination of mapping operators is used to obtain global observation predictions or precipitation variables. We conduct comprehensive experiments to demonstrate the effectiveness of our DAWP framework for global direct observation predictions. We summarize the contributions of this paper as follows:

- **Innovative framework** integrating AIDA and AIWP methods for direct observation predictions: We propose a brand-new framework that leverages an observation space AIDA module, transforming irregular observations into a uniform observation space, with AIWP modules to achieve skillful direct observation predictions, bypassing the limitations of reanalysis data.
- **High-resolution global forecasting** for observation and precipitation: We introduce a mask ViT-VAE and a spatiotemporal transformer with cross-regional boundary information for encoding high-resolution irregular observations, and implement global forecasting efficiently on sub-images of global observations.
- **Comprehensive experiments**: We organize a composite satellite observation dataset with a size of over 35TB, which has a spatiotemporal resolution of $12\times1152\times2304$. Comprehensive experiments and reanalyses are presented, demonstrating the effectiveness of our DAWP framework and the potential of direct observation predictions.

2 Related work

Weather prediction with deep learning. Recent studies have demonstrated that machine learning systems can produce accurate medium-range forecasts, comparable to physics-based models, for key weather parameters [3, 4, 6, 5, 19, 20, 21, 22]. FourcastNet was the first to propose using deep neural networks to learn global atmospheric dynamics [23]. By scaling up the training stage, Pangu-Weather [3] and GraphCast [4] simultaneously achieved accuracy levels comparable to those obtained by the operational IFS systems at ECMWF. Other works extend the AIWP from aspects including forecasting skill [5, 6], resolution [19], probabilistic modelling [24], and physics informed [25]. Although impressive progress has been made, previous AIWP methods still rely on reanalysis data, which introduces inherent limitations, including temporal lag, limited observation use, and dependency on NWP systems.

Direct observation prediction. Direct observation prediction holds transformative potential for overcoming the dependency on reanalysis, enabling the forecasting of weather using direct observations. Although the concept of direct observation prediction has been widely applied in fields such as precipitation nowcasting, where radar echoes are utilized for short-term forecasting [26, 27, 28, 29, 30], its application in global weather prediction remains limited. In contrast to gridded radar observations, direct observations of the global atmosphere are irregular and non-gridded, making global weather DOP challenging. Transformer-DOP proposes using a transformer with mask tokens to handle the irregular global Earth observations [17]. Simultaneously, Graph-DOP employs graph neural networks to flexibly encode direct observations [18]. EarthNet proposes pretraining the backbone with an observation assimilation task and then finetuning it with a prediction task [31]. Although they successfully apply irregular global observation for prediction, these designs suffer from rollout distribution and ineffective spatiotemporal learning as the input space is discrete while the output space is dense. We propose applying AIDA to transform the input space into a dense one, thereby solving the misalignment between input and output.

3 Method

Our DAWP is designed for global observation forecasting in a uniform satellite observation space. The key components of DAWP are an observation space data assimilation module and a cross-regional boundary conditioning weather prediction module. Additionally, we introduce the mask ViT-VAE to encode observations and produce precipitation variables. Our DAWP is illustrated in Figure 2.

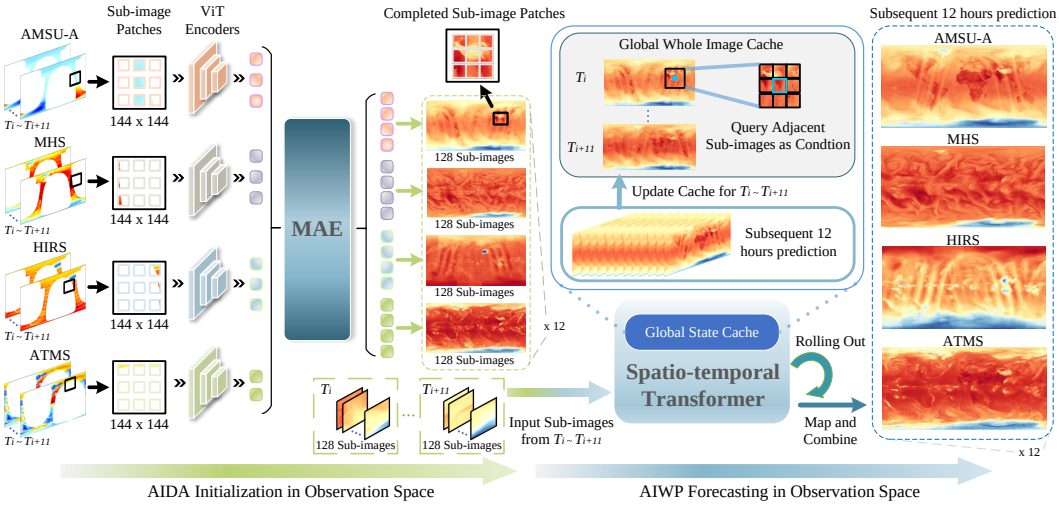

Figure 2: The framework of our DAWP. There are two stages in our DAWP: (1) Initialization and (2) Forecasting.

3.1 Initialization: observation space assimilation by multi-modal masked autoencoder

The missing and sparse observations, attributed to the inherent characteristics of orbital motion, present an irregular input observation space, as shown in the left column of Figure 2. Directly taking these satellite observations as inputs not only restricts the network design for spatiotemporal modeling but also leads to a distribution shift when implementing rollout forecasting, as the output space is required to be a regular one. To meet the gap, we propose using observation space assimilation with a Multi-modal Masked Autoencoder [32, 33] (MMAE) as the initialization stage for direct observation predictions. The MMAE fills in missing areas by leveraging contextual information from different sensors and spatiotemporally nearby observations.

The core of our assimilation module is a naive MAE that imputes masked multi-modal satellite tokens following [34, 32, 31]. Since imputation mainly relies on space-time nearby observations from multiple satellite observations, our MMAE sub-regionally processes data from multiple sources within a fixed time window of 12 and a sub-image of 144×144. Observations in the time window are tokenized frame by frame by pretrained satellite-specific mask ViT encoders, where missing patches are ignored. Remaining spatiotemporal tokens from each satellite are concatenated and passed through MAE for complete missing information in the observation space. For MAE training, we randomly mask a given number of tokens from the whole concatenated remaining tokens and reconstruct the left observed but masked tokens. As the number of observed tokens from different time windows can vary, we flexibly pad [EOS] tokens to maintain a uniform sequence length for efficient attention computation. In the inference stage, masks for MAE are released to utilize as many available observations as possible.

3.2 Forecasting: cross-regional boundary conditioning direct observation prediction

After AIDA, we implement an efficient weather prediction module for direct observation prediction through cross-regional boundary conditioning in the imputed observation space, initialized by our assimilation module. For efficiently integrating with the assimilation module pretrained by sub-images, our weather prediction module is also applied to sub-images generated by the assimilation module. Since sub-images only contain local atmospheric states, predictions on sub-images require cross-regional information interaction for continuous spatiotemporal modeling. We introduce a global state cache to store observation states for cross-regional boundary conditioning during forecasting.

With the global state cache, our weather prediction module achieves efficient cross-regional boundary conditioning observation forecasting by applying spatiotemporal decoupling attention structure. In the forecasting stage, the assimilated observations simplify the prediction task into a standardized spatiotemporal forecasting problem, which is widely explored [35, 36, 37]. We follow the concept of spatiotemporal decoupling in spatiotemporal forecasting [35, 30]. Specifically, our weather prediction

Table 1: Dataset overview.

Modality/Sensor	Satellite	Channels	Level	Period
Advanced Microwave Sounding Unit-A (AMSU-A) [40]	NOAA18 & 19	Microwave radiance 15 bands	1B	2007-2023
Advanced Technology Microwave Sounder (ATMS) [41, 42]	NPP & NOAA20	Brightness temperature 9 bands	1C	2012-2023
High Resolution Infrared Radiation Sounder (HIRS) [40]	NOAA18 & 19	Infrared radiance 20 bands	1B	2007-2023
Microwave Humidity Sounder (MHS) [40]	NOAA18 & 19	Microwave radiance 5 bands	1B	2007-2023
ATMS-Precipitation [43, 44]	NPP & NOAA20	Precipitation product 2 channels	2A	2012-2023

module is composed of N Temporal-Spatial (TS) attention blocks [37, 36]. Attention blocks have the advantages of simplicity and scalability, while the TS spatiotemporal decoupling reduces the sequence length of spatiotemporal tokens for efficient computation. Utilizing the efficiency of TS decoupling, we simply pad tokens from neighbouring areas to the tokens of the prediction region as SD3 [38] and Flux [39] do. In this way, cross-regional boundary information of the border area is passed to the center forecasting region. To forecast multistep global observations, we maintain a global state cache during the inference phase, ensuring consistent updates for subsequent steps. Specifically, when forecasting observations of center sub-images, current adjacent sub-images are queried as conditions from the state cache according to the relative coordinates. After one step of global prediction has been completed, the global state cache is updated with the subsequent 12 hours prediction to support rollout forecasting. More details can be found in the Appendix A.

3.3 Encoding and precipitation mapping via mask ViT-VAE

In our DAWP, we introduce a mask ViT-VAE both for encoding and mapping multiple-channel satellite observations with missing values. The mask ViT-VAE consists of a vision transformer (ViT) encoder/decoder with masks enabling the model to ignore patches without sufficient observations. The ViT encoder/decoder provides better compression capability, as detailed in Appendix D, compared to SD-VAE for satellite observations, which typically have more channels than natural images. Moreover, it explicitly maintains spatial consistency between tokens and pixels by position embeddings and mitigates the influence of missing values through mask attention. With our mask Vit-VAE, encoding/decoding is pretrained as a reconstruction task, while the precipitation mapping is trained with ATMS inputs and ATMS-precipitation outputs.

4 Experiments

In the experiment part, a comprehensive analysis of our DAWP is presented. First, we introduce the composition of our data and training details. Based on the observation data, a comparison of our DAWP with other AI-DOP methods is implemented. Further, we evaluate the precipitation forecasting skill of these AI-DOP methods by applying a precipitation mapping network. In addition to evaluating DAWP's capabilities of forecasting capabilities, we also conducted ablation studies to validate the effectiveness of our modular designs. Finally, the importance of each satellite for DOP is tested by ablating the input modalities of the assimilation module.

4.1 Experimental Setups

Data. The top four datasets listed in Table 1 are used for training our DAWP, while ATMS-precipitation is used to train the precipitation mapping with ATMS. We generate hourly, 0.16° resolution composites by interpolating and reprojecting the raw data as detailed in Appendix E. Additional information for dataset split and introduction is presented in Appendix F.

Training details. The training details are presented in Appendix G. Training our mask ViT-VAE, MMAE AIDA, and TS decoupling AIWP takes about 1 day, 5 days, and 4 days, respectively. Besides, our DAWP is compared with the persistence model, using our AIDA module for completion of missing values, as mentioned in [35]. We replicate the EarthNet [31] and Transformer-DOP [26] models on our composite dataset, as their codes and data are closed source.

Table 2: MAE error of forecasting during 3 lead time periods (0-12h, 12-24h, and 24-36h) for different channels of the satellite data. We use the unit of 1e-5 for AMSU-A, 1e-4 for MHS, and 1e-0 for both ATMS and HIRS.

Methods	Lead time	AMSU-A		ATMS		HIRS		MHS	
		ch0	ch1	ch0	ch1	ch9	ch10	ch0	ch1
Persistence [35]	0-12h	5.86	9.15	14.37	11.69	12.43	2.21	7.01	14.74
EarthNet [31]		2.93	4.89	6.96	6.14	9.15	1.39	3.96	9.65
Transformer-DOP [17]		2.67	4.48	6.40	5.61	9.22	1.42	3.91	9.48
Ours		1.92	3.39	3.36	3.27	7.70	1.12	3.07	7.91
Persistence [35]	12-24h	4.35	6.94	10.40	8.86	13.58	2.30	6.07	15.09
EarthNet [31]		4.12	6.65	11.25	9.00	11.14	1.98	5.46	13.11
Transformer-DOP [17]		3.84	6.14	10.04	8.04	11.08	1.95	5.19	12.65
Ours		3.11	5.12	7.35	6.35	9.57	1.54	4.51	10.54
Persistence [35]	24-36h	6.39	9.84	15.37	12.52	14.61	2.61	8.00	17.86
EarthNet [31]		5.17	8.14	12.52	10.08	12.37	2.36	6.41	15.08
Transformer-DOP [17]		4.91	7.54	11.35	9.07	12.39	2.27	6.22	14.70
Ours		3.66	5.80	7.84	6.81	10.71	1.79	5.15	12.22

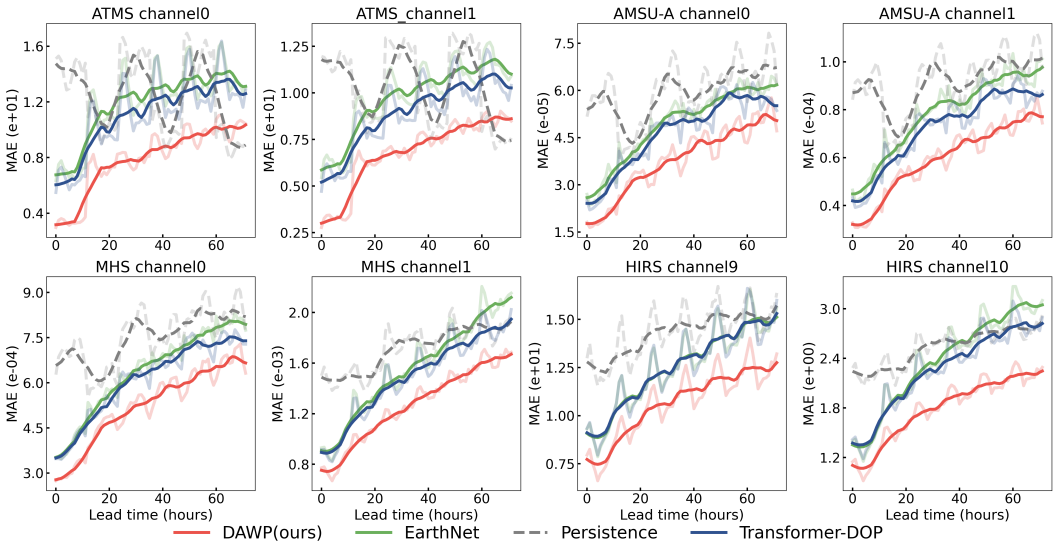

Figure 3: Curves of MAE for the prediction of different modalities. The max leadtime is 72h with a 1h temporal resolution.

4.2 Direct observation prediction

We compare our DAWP with other AI-DOP methods to evaluate the ability to predict direct observations. The metric for evaluation is the mean absolute error (MAE) between the predictions and observed values. The baselines are also trained with sub-images as our DAWP. Among them, persistence is a naive method that uses the last observation as the prediction of the next one.

The time-averaged MAEs within fixed time windows of 0-12h, 12-24h, and 24-36h are presented in Table 2. We select two channels with meaningful patterns of each modality observation to calculate the MAE. The results show that our DAWP significantly outperforms other methods. Specifically, our DAPW's MAE on AMSU-A among 24-36h is 3.66 and 5.80, which not only outperforms those of EarthNet and Transformer-DOP, but also surpasses EarthNet's (3.84 and 6.14) and Transformer-DOP's (4.12 and 6.65) MAEs on AMSU-A among 12-24h. It is the same when comparing our DAWP with baselines on other modalities. These results indicate our DAWP has a 12-hour lead time advantage in direct observation prediction. To provide a more detailed analysis of the observation prediction, we present the MAE figure in a time range of 0-72h in Figure 3. It can be observed that the 1h temporal resolution of the figure depicts the periodicity of the prediction errors. This periodicity is attributed to the strip-scanning characteristics of polar-orbiting satellites, which generate periodic

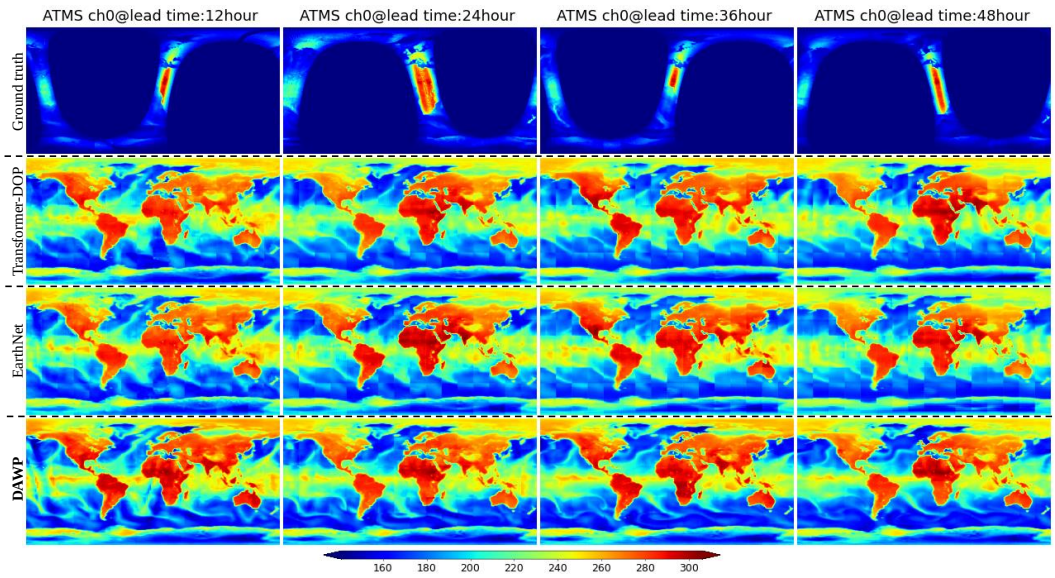

Figure 4: A visualization of rollout predictions for global satellite observation forecasting.

Table 3: Forecasting skills in 12 hours on precipitation-related variables Total Column Water Vapor (TCWV) and Surface Precipitation (SP). CSI and FAR scores are calculated on different thresholds.

Method	TCWV (mm)						SP (mm/h)					
	CSI-10	CSI-20	CSI-30	FAR-10	FAR-20	FAR-30	CSI-0.5	CSI-1.0	CSI-2.0	FAR-0.5	FAR-1.0	FAR-2.0
Persistence [35]	0.853	0.702	0.636	0.121	0.266	0.332	0.110	0.073	0.031	0.844	0.861	0.655
EarthNet [31]	0.909	0.822	0.786	0.047	0.130	0.172	0.127	0.050	0.008	0.666	0.692	**0.180**
Transformer-DOP [17]	0.905	0.822	0.789	0.053	0.130	0.169	0.136	0.057	0.010	0.655	0.685	0.214
Ours	**0.917**	**0.841**	**0.807**	**0.034**	**0.088**	**0.120**	**0.197**	**0.102**	**0.035**	**0.529**	**0.577**	0.283

observations. Although there is periodicity in the MAE, our DAWP consistently outperforms other methods, establishing a lead time advantage from the beginning. Specifically, in subfigure (a), at the beginning of the prediction, our DAWP's MAE is about 3.5 while the MAE of EarthNet is about 6.8, which is almost 2 times larger than ours. In addition, EarthNet is surpassed by the naive persistent baseline at the lead time of 16h, but our DAWP doesn't meet the persistent baseline until about 64h.

In Figure 4, we visualize the prediction results of these methods at different lead times. At the lead time of 12 hours (the first prediction step), our DAWP exhibits a slightly better prediction than others. When the lead time increases, a significant distortion appears in the predictions of EarthNet and Transformer-DOP, while our DAWP maintains a relatively stable structure. It has the same trend as the MAE curve in Figure 3, indicating that our DAWP is more robust in rollout predictions.

The results of direct observation prediction indicate that our DAWP outperforms other methods in the 0-72h time range and selected channels, demonstrating the effectiveness of our method in both initial prediction and rollout prediction. More evaluations on other channels are shown in the Appendix I.

4.3 Global precipitation forecasts

In this section, we evaluate the potential of AI-DOP methods for global precipitation forecasting. A precipitation mapping network is trained to transform observation predictions into precipitation products. It is worth noting that in this way, our DAWP does not require satellite precipitation products as input, making the forecasting a quick response to observations.

The precipitation skill is evaluated by the Critical Success Index (CSI) [45] and False Alarm Ratio (FAR) [46] metrics on observed points in a time range of 0-12h. CSI measures the ability of the model to correctly identify precipitation events, while FAR assesses the reliability of the model's predictions. The combined application of CSI and FAR enables a comprehensive precipitation forecasting skill assessment. We present a detailed definition of CSI and FAR in Appendix H.

Table 3 presents the quantitative results of the precipitation forecasting skill of AI-DOP methods on total column water vapor index (TCWV) and surface precipitation (SP). The thresholds for TCWV and SP are set to [10mm, 20mm, 30mm] and [0.5mm/h, 1.0mm/h, 2.0mm/h], respectively. On variable TCWV, our DAWP achieves a slight advantage in CSI over other methods. Specifically, our DAWP's achieves 0.807 on CSI-30, while Earthnet and transformer-dop achieve 0.786 and 0.789, respectively. This advantage is maintained across all thresholds. In terms of FAR, our DAWP is significantly lower than other methods. It indicates that our DAWP predicts more accurate TCWV with lower false alarm rates than baseline methods, demonstrating the reliability of our DAWP. We also analyze the forecasting skill on SP variable. The CSI of our DAWP is significantly higher than that of other methods on CSI-0.5 and CSI-1.0, achieving 0.197 and 0.102. Compared with transformer-dop, our DAWP increases CSI-0.5 and CSI-1.0 by 44.8% and 78.9%. Another observation about CSI is a decreasing trend with increasing thresholds. Especially, when the threshold increases to 2.0, the CSI of EarthNet and transformer-dop are even lower than that of the naive persistent model. Only our DAWP maintains a slightly higher forecast skill on CSI 2.0. The decrease of forecasting skill is caused by the temporal decay of prediction intensity [26, 29], which could further hamper the mapping network's performance. When evaluating FAR on threshold 2.0mm/h, our DAWP uncommonly surpasses Earthnet and transformer-dop slightly. It could be explained by considering the CSI-2.0 results, as it is difficult for EarthNet and transformer-dop to produce predictions greater than 2.0mm/h, resulting in few false alarms. Besides, compared to the persistent baseline with a comparable CSI-2.0 skill, our DAWP has a FAR score which 57.8% lower than persistent's, showing stronger reliability. The evaluation of CSI and FAR scores demonstrates the potential of our DAWP for global precipitation forecasting, simultaneously increasing the accuracy and reliability for precipitation forecasting skill of AI-DOP methods.

4.4 Ablation study

Effect of AIDA initialization. An ablation study is conducted to validate the effectiveness of our AIDA module on our DAWP and other AI-DOP methods. The results demonstrate that AIDA enables efficient spatiotemporal modeling and enhances the rollout prediction ability of AI-DOP methods by imputing the observation space.

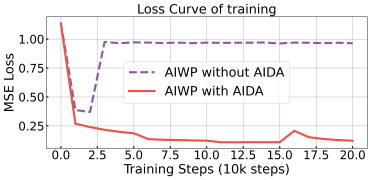

Figure 5: Training loss curve.

The training is unstable after cancelling the AIDA stage in our DAWP. We show the curve of training loss in Figure 5. It can be observed that the training loss of our DAWP with AIDA is decreasing steadily, while the loss curve without the initialization of AIDA dramatically increases at about 20k training steps and maintains a MSE loss of 1.0. In contrast, the loss of our DAWP with AIDA could converge to less than 0.1 after 200k training steps. This phenomenon indicates that the classical spatiotemporal learning methods are difficult to learn the dynamics in an observation space with variable missing values. By imputing the observation space into a completed spatiotemporal space, AIDA benefits efficient spatiotemporal learning modeling.

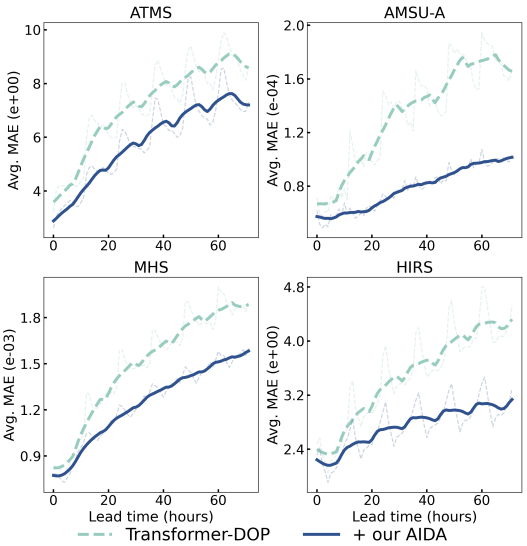

Figure 6: The MAE curves of Transformer-DOP with(w/o) our AIDA initialization.

Applying AIDA initialization enhances the rollout prediction ability of AI-DOP methods. The quantitative results of transformer-dop are depicted in Figure 6. We exhibit the averaged MAE of 0-72h prediction on each sensor with-/without AIDA initialization. Transformer-DOP with AIDA initialization significantly outperforms the original Transformer-DOP when the lead time increases. It reveals the potential of AIDA initialization for boosting the AI-DOP model for multi-step rollout prediction.

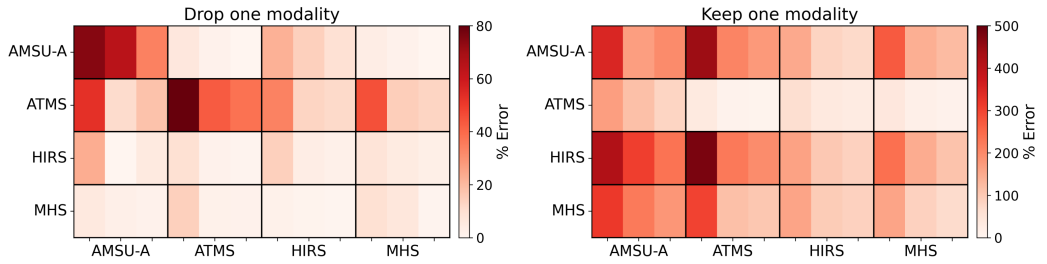

Figure 8: Matrix of relative errors under the setting of dropping one modality and keeping one modality. The three columns in a black rectangle represent the relative MAE error ratios of 0-12h, 12-24h, and 24-36h lead times, respectively.

Gains of cross-regional boundary conditioning. We compare the observation predictions with and without cross-regional conditions by training a DAWP without neighbour regions as inputs. This ablation study is conducted to verify the effect on the accuracy and spatial continuity of the prediction.

The convergence loss of our DAWP with and without cross-regional condition is presented in Table 4, showing conditioning on neighbour regions improves the prediction accuracy. In this table, we evaluate the convergence loss of center areas and neighbor areas. For DAWP without Cross-regional Boundary Conditioning (CBC), we take the center area as the neighbour area. The average convergence loss of the center area is 0.106, which is significantly than that of the neighbour area. Besides, it also significantly outperforms the DAWP without CBC on the centre area by about 15.6%. This result validates the necessity of cross-regional conditions for weather prediction, as the boundary information of atmospheric physical motion is crucial for accurate weather forecasting.

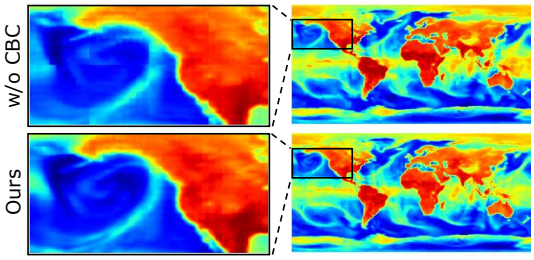

Figure 7: A visualization of forecasting results with(w/o) Cross-regional Boundary Conditioning.

Another benefit of cross-regional conditions is the improvement of prediction continuity, as shown in Figure 7. The first and second rows show the predictions of ATMS channel 0 at a lead time of 6 hours without and with CBC, respectively. It can be observed in the black box that the continuity of the adjacent regions is improved, which is helpful for keeping the atmospheric structure.

Modality sensitivity analysis. We introduce the experiment of using different modality combinations for multistep observation prediction to gain insight into the importance of each modality's data. As shown in Figure 8, a drop one and a keep one combination are explored. The setting of drop one means removing one modality's observation before AIDA initialization and conducting multi-step prediction, while in the keep one setting, we only keep one modality's observation before AIDA initialization for forecasting. We evaluate the influence of each modality by calculating the relative MAE error ratios between the MAE of DAWP with completed modality inputs.

Table 4: Converged loss of our DAWP with and without CBC module. The average loss is calculated by averaging the loss of all four modalities.

Module	Area	Modalitiy				Avg.
		AMSU-A	ATMS	HIRS	MHS	
w/o CBC	border	0.063	0.101	0.236	0.111	0.128
	center	0.063	0.101	0.236	0.111	0.128
with CBC	border	0.069	0.098	0.324	0.101	0.148
	center	**0.054**	**0.074**	**0.214**	**0.084**	**0.106**

The result of drop one is shown in the left of Figure 8. Row names of the figure indicate the modality that is dropped, while columns sequentially included in a modality name represent the relative MAE error ratio of this modality in time windows 0-12h, 12-24h, and 24-36h. The overall trend of the drop one setting demonstrates that dropping any modality's data leads to an increase of MAE error. Besides, it's observed that for the rollout predictions, the MAE error ratio is gradually decreasing, indicating the insufficient usage of observations for multistep predictions. When focusing on single modalities, we find that HIRS and MHS can still achieve a relatively low MAE error ratio when their own data is dropped, indicating that they have redundant information for the prediction.

The error ratio matrix of keep one setting is also shown in Figure 8. When only keeping ATMS observation, other modalities' MAE error ratios of predictions are the lowest, which are even lower than keeping the satellites themselves. This indicates that ATMS has a substantial amount of information for prediction. In contrast, for predicting one modality itself, AMSU-A exhibits the highest MAE error ratio, showing that it has the least spatiotemporal dynamics information.

5 Conclusion

In this paper, we propose DAWP, a novel framework using AIWP for observation prediction with an AIDA module as initialization. Comprehensive experiments are conducted to validate the efficiency and potential of our DAWP framework for observation forecasting and downstream applications such as precipitation forecasting. **Broader Impacts&Future Work**: First, our framework readily integrates variable observations, demonstrating its potential as an implicit Earth system modeling framework. Second, our framework has broad application prospects. It can seamlessly adapt to diverse downstream tasks-such as surface parameter estimation, wildfire monitoring, and sea ice mapping-whenever observations or retrieval operators are available, similar to precipitation forecasting. Third, our framework holds a promising potential for directly predicting physical variables by integrating observations of weather variables such as station data. **Limitations**: Although our DAWP framework improves the observation forecasting, the sources of observation are still homogeneous in satellite observations. More observation sources will be integrated with DAWP in the future.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory and the JC STEM Lab of AI for Science and Engineering, funded by The Hong Kong Jockey Club Charities Trust, the Research Grants Council of Hong Kong (Project No. CUHK14213224). This work was done during Junchao Gong's internship at Shanghai Artificial Intelligence Laboratory.

References

[1] Peter D Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009, 2018.

[2] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, pages 1–3, 2025.

[3] 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.

[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.

[5] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

[6] 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.

[7] 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.

[8] Michele M Rienecker, Max J Suarez, Ronald Gelaro, Ricardo Todling, Julio Bacmeister, Emily Liu, Michael G Bosilovich, Siegfried D Schubert, Lawrence Takacs, Gi-Kong Kim, et al. Merra: NasaâĂŹs modern-era retrospective analysis for research and applications. *Journal of climate*, 24(14):3624–3648, 2011.

[9] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[10] Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, et al. The ncep/ncar 40-year reanalysis project. In *Renewable energy*, pages Vol1_146–Vol1_194. Routledge, 2018.

[11] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[12] Arianna Valmassoi, Jan D Keller, Daryl T Kleist, Stephen English, Bodo Ahrens, Ivan Bašták Ďurán, Elisabeth Bauernschubert, Michael G Bosilovich, Masatomo Fujiwara, Hans Hersbach, et al. Current challenges and future directions in data assimilation and reanalysis. *Bulletin of the American Meteorological Society*, 104(4):E756–E767, 2023.

[13] Guannan Hu, Sarah L Dance, Ross N Bannister, Hristo G Chipilski, Oliver Guillet, Bruce Macpherson, Martin Weissmann, and Nusrat Yussouf. Progress, challenges, and future steps in data assimilation for convection-permitting numerical weather prediction: Report on the virtual meeting held on 10 and 12 november 2021. *Atmospheric Science Letters*, 24(1):e1130, 2023.

[14] Jing-An Sun, Hang Fan, Junchao Gong, Ben Fei, Kun Chen, Fenghua Ling, Wenlong Zhang, Wanghan Xu, Li Yan, Pierre Gentine, and Lei Bai. Align-da: Align score-based atmospheric data assimilation with multiple preferences, 2025.

[15] Patrick Laloyaux, Thorsten Kurth, Peter Dominik Dueben, and David Hall. Deep learning to estimate model biases in an operational nwp assimilation system. *Journal of Advances in Modeling Earth Systems*, 14(6):e2022MS003016, 2022.

[16] Haiyu Dong. Omg-hd: A high-resolution ai weather model for end-to-end forecasts from observations. In *AI4X 2025 International Conference*, 2025.

[17] Anthony McNally, Christian Lessig, Peter Lean, Eulalie Boucher, Mihai Alexe, Ewan Pinnington, Matthew Chantry, Simon Lang, Chris Burrows, Marcin Chrust, et al. Data driven weather forecasts trained and initialised directly from observations. *arXiv preprint arXiv:2407.15586*, 2024.

[18] Mihai Alexe, Eulalie Boucher, Peter Lean, Ewan Pinnington, Patrick Laloyaux, Anthony McNally, Simon Lang, Matthew Chantry, Chris Burrows, Marcin Chrust, et al. Graphdop: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. *arXiv preprint arXiv:2412.15687*, 2024.

[19] Tao Han, Song Guo, Fenghua Ling, Kang Chen, Junchao Gong, Jingjia Luo, Junxia Gu, Kan Dai, Wanli Ouyang, and Lei Bai. Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting. *arXiv preprint arXiv:2402.00059*, 2024.

[20] Sijie Zhao, Feng Liu, Xueliang Zhang, Hao Chen, Tao Han, Junchao Gong, Ran Tao, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Transforming weather data from pixel to latent space. *arXiv preprint arXiv:2503.06623*, 2025.

[21] Xiangyu Zhao, Zhiwang Zhou, Wenlong Zhang, Yihao Liu, Xiangyu Chen, Junchao Gong, Hao Chen, Ben Fei, Shiqi Chen, Wanli Ouyang, et al. Weathergfm: Learning a weather generalist foundation model via in-context learning. *arXiv preprint arXiv:2411.05420*, 2024.

[22] Wanghan Xu, Fenghua Ling, Tao Han, Hao Chen, Wanli Ouyang, and LEI BAI. Generalizing weather forecast to fine-grained temporal scales via physics-ai hybrid modeling. *Advances in Neural Information Processing Systems*, 37:23325–23351, 2024.

[23] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[24] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[25] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

[26] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[27] Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.

[28] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

[29] Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. Postcast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. *arXiv preprint arXiv:2410.05805*, 2024.

[30] Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

[31] Thomas J Vandal, Kate Duffy, Daniel McDuff, Yoni Nachmany, and Chris Hartshorn. Global atmospheric data assimilation with multi-modal masked autoencoders. *arXiv preprint arXiv:2407.11696*, 2024.

[32] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

[33] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[35] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

[36] Yujin Tang, Lu Qi, Fei Xie, Xiangtai Li, Chao Ma, and Ming-Hsuan Yang. Predformer: Transformers are effective spatial-temporal predictive learners. *arXiv preprint arXiv:2410.04733*, 2024.

[37] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[38] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[39] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[40] EUMETSAT. User services client. `https://archive.eumetsat.int/usc/UserServicesClient.html`.

[41] GES DISC. Gpm atms on suomi-npp common calibrated brightness temperature. `https://disc.gsfc.nasa.gov/datasets/GPM_1CNPPATMS_07/summary`.

[42] GES DISC. Gpm atms on noaa-20 common calibrated brightness temperatures. `https://disc.gsfc.nasa.gov/datasets/GPM_1CNOAA20ATMS_07/summary`.

[43] GES DISC. Gpm atms on suomi-npp radiometer precipitation profiling. `https://disc.gsfc.nasa.gov/datasets/GPM_2AGPROFNPPATMS_CLIM_07/summary`.

[44] GES DISC. Gpm atms on noaa-20 climate-based radiometer precipitation profiling. `https://disc.gsfc.nasa.gov/datasets/GPM_2AGPROFNOAA20ATMS_CLIM_07/summary`.

[45] Joseph T Schaefer. The critical success index as an indicator of warning skill. *Weather and forecasting*, 5(4):570–575, 1990.

[46] Lindsey R Barnes, David M Schultz, Eve C Gruntfest, Mary H Hayden, and Charles C Benight. Corrigendum: False alarm rate or false alarm ratio? *Weather and Forecasting*, 24(5):1452–1454, 2009.

[47] Yi Xiao, Lei Bai, Wei Xue, Hao Chen, Kun Chen, Tao Han, Wanli Ouyang, et al. Towards a self-contained data-driven global weather forecasting framework. In *Forty-first International Conference on Machine Learning*, 2024.

[48] Kun Chen, Peng Ye, Hao Chen, Tao Han, Wanli Ouyang, Tao Chen, LEI BAI, et al. Fnp: Fourier neural processes for arbitrary-resolution data assimilation. *Advances in Neural Information Processing Systems*, 37:137847–137872, 2024.

[49] Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D Dueben, and Torsten Hoefler. Diffda: a diffusion model for weather-scale data assimilation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19798–19815, 2024.

[50] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[51] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.

[52] Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[53] 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.

[54] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782, 2023.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[56] Tao Han, Zhenghao Chen, Song Guo, Wanghan Xu, and Lei Bai. Cra5: Extreme compression of era5 for portable global climate and weather research via an efficient variational transformer. *arXiv preprint arXiv:2405.03376*, 2024.

[57] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[58] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1âĂŞ2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clarify our contributions in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] ,

   Justification: This is a paper for applications.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all experimental settings for reproductivity.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We train our model on ATMS, AMSU-A, MHS, and HIRS, which could be obtained publicly on their websites. The detailed settings for data resampling are provided in supplemental material. We will release our code on github.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so âĂIJNoâĂİ is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The complete experimental settings, including the dataset, hyperparameters, type of optimizer, etc, are provided in the Experiments section and supplemental materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars, like other related work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The sufficient information on the computer resources is provided in the Experimets section

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We perform this work following the NeurIPS Code of Ethics exactly.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts of the work performed in the conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: This work does not involve such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original papers or websites that produced the code or dataset are properly cited and we use an open-source dataset for our experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code in the paper is well documented and the documentation is provided alongside the code.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: The paper has no crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Global state cache

In this section, we clarify the design of the global state cache. It is composed of a current cache and a previous cache. The current one is used to store the predicted sub-images of $T_{12i:12(i+1)}$, and the previous one, storing sub-images of $T_{12(i-1):12i}$ is used to provide neighbour information for the prediction.

```
class GlobalStateCache:
    def __init__(self, domains, domain_Chs, time_window,):
        self.domains = domains
        self.time_window = time_window
        self.domain_Chs = domain_Chs

        self.cur_cache = self.init_cache()
        self.prev_cache = self.init_cache()

    def init_cache(self):
        domain_cache = {}
        for domain in self.domains:
            domain_Ch = self.domain_Chs[domain]
            # original image size is 1152x2304.
            # 8x16 subimages with the size of 144x144.
            # 9x9 is the number of tokens in a subimage.
            domain_cache[domain] = /
            torch.zeros((8, 16, self.time_window, domain_Ch, 9, 9))
        return domain_cache
```

The main operators of the global state cache are query neighbors and update cache. Query neighbors is used to get the adjacent sub-images as conditions for the prediction of central areas. Update cache is used to update the previous and current cache with the subsequent 12 hours predictions.

```
    def query_neighbours(self, rel_coordinate):
        neighbour_coords = self._get_8_neighbour_coord(rel_coordinate)
        neighbours = {}
        for domain in self.domains:
            neighbours[domain] = []
            for coord in neighbour_coords:
                neighbours[domain].append(
                    self.prev_cache[domain][coord[0], coord[1]]
                    )
        return neighbours


    def update_cache(self, rel_coordinate, pred_subimg):
        for domain in self.domains:
            coords = rel_coordinate[0], rel_coordinate[1]
            self.cur_cache[domain][coords] = pred_subimg[domain]
        if have_pred_whole_img:
            self.prev_cache = self.cur_cache
        return None
```

To get the neighbour coordinates from a Plane Rectangular Coordinate System, we utilized a _get_8_neighbour_coord function that incorporates Earth's spherical geometry, specifically handling the left-right and top-bottom boundaries of the image.

```
    def _get_8_neighbour_coord(self, rel_coordinate, h=8, w=16):
        """
        h, w is the height and width of the image
        r, c is the coordinate of the pixel
        case 1: if the pixel is in the center of the image, return all
                                                8 neighbours
        case 2: if the pixel in w border, treat the image as h border
                                                is connected
```

```
        case 3: if the pixel in h border, symmetrically get the
                                            neighour
        case 4: if the pixel in corner, use the rule of both w border
                                            and h border
        ret:
        8 neighbours ordered as [up_left, up, up_right, left, right,
                                            down_left, down, down_right
                                            ]
        """
        r, c = rel_coordinate[0], rel_coordinate[1]
        assert r >= 0 and r < h
        assert c >= 0 and c < w
        w_border_flag = (c == 0 or c == w - 1)
        h_top_border_flag = (r == 0)
        h_bottom_border_flag = (r == (h - 1))

        neighbours = []
        if not (h_top_border_flag or h_bottom_border_flag):
            up_left = ((r - 1) % h, (c - 1) % w)
            up = ((r - 1) % h, c)
            up_right = ((r - 1) % h, (c + 1) % w)
            left = (r, (c - 1) % w)
            right = (r, (c + 1) % w)
            down_left = ((r + 1) % h, (c - 1) % w)
            down = ((r + 1) % h, c)
            down_right = ((r + 1) % h, (c + 1) % w)
        elif h_top_border_flag:
            up_left = (r, (c + 1 + w//2) % w)
            up = (r, (c + w // 2)% w)
            up_right = (r, (c - 1 + w//2) % w)
            left = (r, (c - 1) % w)
            right = (r, (c + 1) % w)
            down_left = ((r + 1) % h, (c - 1) % w)
            down = ((r + 1) % h, c)
            down_right = ((r + 1) % h, (c + 1) % w)
        elif h_bottom_border_flag:
            up_left = ((r - 1) % h, (c - 1) % w)
            up = ((r - 1) % h, (c)% w)
            up_right = ((r - 1) % h, (c + 1) % w)
            left = (r, (c - 1) % w)
            right = (r, (c + 1) % w)
            down_left = (r, (c + 1 + w//2) % w)
            down = (r, (c + w//2)% w)
            down_right = (r, (c - 1 + w//2) % w)
        else:
            raise NotImplementedError


        neighbours.append(up_left)
        neighbours.append(up)
        neighbours.append(up_right)
        neighbours.append(left)
        neighbours.append(right)
        neighbours.append(down_left)
        neighbours.append(down)
        neighbours.append(down_right)
        return neighbours
```

B Artificial intelligence data assimilation

The development of data assimilation has also been revolutionized by artificial intelligence. Xiao et al. [47] were the first to apply the popular traditional numerical data assimilation method, Four-Dimensional Variational, to the AIWP model FengWu. Furthermore, researchers have explored the

development of artificial intelligence assimilation methods, such as FNP [48] and DiffDA [49], which could be applied to both NWP and AIWP models. Although impressive progress has been made, these methods remain limited by reanalysis data and NWP models, which require transforming the observations into physical space. Unlike previous methods, EarthNet [31] proposes implementing observation space data assimilation with masked reconstruction. We are motivated to use an observation AIDA model for formulating a complete observation space.

C Comparisons with spatiotemporal learning methods

Table 5: More results on spatiotemporal methods. MAE error of forecasting during 3 lead time periods (0-12h, 12-24h, and 24-36h) for different channels of the satellite data. We use the unit of 1e-5 for AMSU-A, 1e-4 for MHS, and 1e-0 for both ATMS and HIRS.

Methods	Lead time	AMSU-A		ATMS		HIRS		MHS	
		ch0	ch1	ch0	ch1	ch9	ch10	ch0	ch1
Persistence [35]	0-12h	5.86	9.15	14.37	11.69	12.43	2.21	7.01	14.74
ConvLSTM [50]		127.82	208.90	73.21	78.44	77.40	11.05	62.27	158.57
PredRNN [51]		2.95	4.96	6.99	6.21	9.26	1.42	4.11	10.20
RainFormer [52]		3.95	6.52	9.08	8.05	10.41	1.63	4.98	11.69
EarthFormer [35]		18.97	33.94	37.33	38.75	22.49	3.62	18.00	55.04
SimVP [53]		3.61	6.02	7.29	6.61	9.84	1.53	4.68	11.35
TAU [54]		3.84	6.31	7.70	6.86	9.94	1.58	4.87	11.42
EarthNet [31]		2.93	4.89	6.96	6.14	9.15	1.39	3.96	9.65
Transformer-DOP [17]		2.67	4.48	6.40	5.61	9.22	1.42	3.91	9.48
Ours		1.92	3.39	3.36	3.27	7.70	1.12	3.07	7.91
Persistence [35]	12-24h	4.35	6.94	10.40	8.86	13.58	2.30	6.07	15.09
ConvLSTM [50]		128.13	211.14	81.22	82.60	71.64	9.98	59.26	145.44
PredRNN [51]		3.85	6.26	11.14	8.82	10.92	1.84	5.48	13.16
RainFormer [52]		11.46	15.72	19.43	17.76	18.75	3.28	11.74	32.10
EarthFormer [35]		18.98	33.96	37.33	38.76	22.50	3.62	18.01	55.03
SimVP [53]		4.83	7.74	11.48	9.10	11.49	2.02	6.43	14.20
TAU [54]		4.46	7.21	11.46	9.06	11.65	2.04	6.24	13.94
EarthNet [31]		4.12	6.65	11.25	9.00	11.14	1.98	5.46	13.11
Transformer-DOP [17]		3.84	6.14	10.04	8.04	11.08	1.95	5.19	12.65
Ours [35]		3.11	5.12	7.35	6.35	9.57	1.54	4.51	10.54
Persistence [35]	24-36h	6.39	9.84	15.37	12.52	14.61	2.61	8.00	17.86
ConvLSTM [50]		128.42	211.83	82.03	85.17	72.62	10.15	60.44	148.62
PredRNN [51]		4.58	7.23	12.18	9.69	12.03	2.08	6.24	15.04
RainFormer [52]		24.89	32.34	35.49	36.54	30.78	4.92	21.51	69.68
EarthFormer [35]		18.98	33.96	37.34	38.77	22.51	3.63	18.02	55.04
SimVP [53]		5.72	8.91	12.93	10.33	12.71	2.32	73.95	15.99
TAU [54]		5.61	8.76	13.23	10.53	12.98	2.34	7.15	15.85
EarthNet [31]		5.17	8.14	12.52	10.08	12.37	2.36	6.41	15.08
Transformer-DOP [17]		4.91	7.54	11.35	9.07	12.39	2.27	6.22	14.70
Ours		3.66	5.80	7.84	6.81	10.71	1.79	5.15	12.22

As shown in Table 5, we compare DAWP with more spatiotemporal methods including RNN-based ([50], [51]), CNN-based ([53], [54]), and transformer-based ([52], [35]). Our DAWP maintains a significant advantage over these methods, demonstrating the effectiveness of our AIDA module in improving the roll-out and efficiency of AIWP.

We present implementation details of EarthNet and Transformer-DOP here. There is no open-sourced code for EarthNet [31] or Transformer-DOP [17]. For EarthNet [31], it follows the implementation of MultiMAE [32] as detailed in EarthNet's Appendix C and D. Therefore, we reproduce it on our datasets following MultiMAE. As for Transformer-DOP, since the original paper presents only a sketch without details, we implemented it according to our best available understanding. Specifically, EarthNet is reproduced as a 12-layer encoder (hidden dimension 768) paired with an 8-layer decoder (hidden dimension 512), and Transformer-DOP is implemented as a transformer consisting of 18 layers (hidden dimension 1024). We employ sub-images because the full 12-hour global observation sequence would result in 124k-token sequence, which is computationally infeasible.

Table 6: Reconstruction error of various VAEs on different modalities. The column of **Improvement** represents relative average improvement over SD-VAE.

VAEs	Modality reconstruction error				Improvement
	AMSU-A (1e-3)	ATMS (1e-2)	HIRS (1e-3)	MHS (1e-2)	
SD-VAE [55]	1.07	1.26	5.84	2.28	-
Mask-SD-VAE	1.21(-13.1%)	1.31(-4.0%)	5.60(+4.1%)	2.35(-3.1%)	-4.1%
ViT-VAE [56]	0.92(+14.0%)	1.36(-7.9%)	4.28(+26.7%)	2.45(-7.4%)	+6.3%
Mask-ViT-VAE	0.78(+27.1%)	1.29(-2.3%)	4.11(+29.6%)	2.41(-5.7%)	+12.2%

D VAE comparison

We explore the ability of our mask ViT-VAE to compress satellite data with multiple channels and missing values by comparing it with other VAEs.

The results are shown in Table 6, where we compare the reconstruction loss of different VAEs. First, VAEs with ViT structure are more effective for reconstructing modality data with multiple channels, such as HIRS and AMSU-A, while on modalities with fewer channels, there is only a slight increase in reconstruction loss. Another observation is that the application of a mask consistently increases VAEs' reconstruction ability on HIRS. For ViT-VAE, it is beneficial to use the mask for the computation of attention between patch tokens, as it could directly weaken the influence of missing tokens.

E Satellite data preprocessing

Preprocessing: The original satellite observation data points are extremely sparse and irregular. To spatially align different observation sources and channels for model training, a remapping procedure is performed beforehand. The pseudocode for the preprocessing algorithm is given in algorithm 1.

Algorithm 1: Remapping Satellite Observation

Input: target resolution R, observation D_o, corresponding latitudes C_{lat} and longitudes C_{lon}
Output: remapped observation data points D_{grid} on desired global grid

```
Generate global grid C_R of desired resolution R that follows Equirectangular projection;
Assign latitudes and longitudes (C_lat,C_lon) to the nearest coordinates (C'_lat,C'_lon) on grid C_R;
D_grid ← NaN with the shape of C_R;
D_count ← Zeros with the shape of C_R;
for d_o,c'_lat,c'_lon in D_o,C'_lat,C'_lon do
    Locate corresponding data point d_grid of D_grid according to coordinates (c'_lat,c'_lon);
    Locate corresponding point counter d_count of D_count according to coordinates (c'_lat,c'_lon);
    if d_grid equals NaN then
        d_grid ← d_o;
    else
        d_grid ← (d_grid + d_o);
    end
end
Average each d_grid where d_count ≥ 1
```

Normalization: We normalize each modality for efficient convergence. The direct observation modalities are normalized by:

$$x^M_{norm} = \frac{x^M - Mean(x^M)}{Std(x^M)}, \tag{1}$$

where M represents the modality M. For ATMS-precipitation, we first implement the log-transformation as:

$$x_{prec} = \log\left(x/a + b\right), \tag{2}$$

to alleviate long-tail distribution, and then normalize these variables like other modalities. Specifically, we select $a = 1e-7$, $b = 1e2$ for SP channel, and choose $a = 1$, $b = 1$ for TCWV channel. The

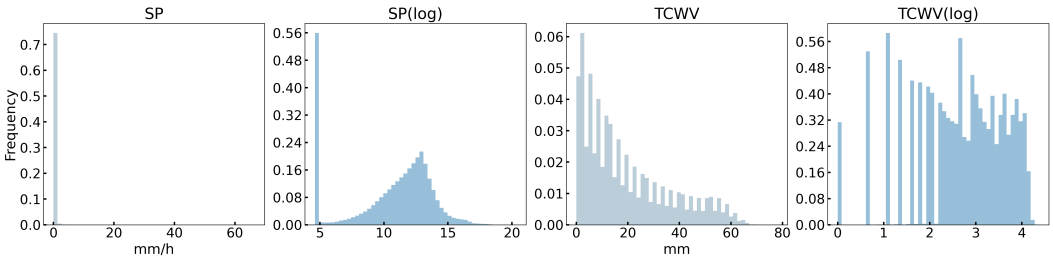

Figure 9: Distribution of ATMS precipitation productions. SP(log) indicates applying a log-transform on SP. It is the same for TCWV(log).

distribution shift is shown in Figure 9. The original distributions also motivate us to choose a threshold list of [0.5 mm/h, 1 mm/h, 2 mm/h] for SP.

F Dataset introduction

The four sensors below are organized into a satellite observation dataset for direct observation forecasting. This composite dataset has a training split with data from January of 2012 to June of 2022 for training and a testing split composed of data from May of 2023 to July of 2023. We present detailed introductions to these sensors.

AMSU-A: The Advanced Microwave Sounding Unit-A (AMSU-A) is a 15-channel microwave radiometer used for measuring global atmospheric temperature profiles and will provide information on atmospheric water in all of its phases (with the exception of small ice particles, which are transparent at microwave frequencies). AMSU-A will provide information even in cloudy conditions. AMSU-A measures Earth radiance at frequencies (in GHz) as listed under the instrument channel information. Level 1B data was collected from EUMETSAT at `https://archive.eumetsat.int/usc/UserServicesClient.html`.

ATMS: The Advanced Technology Microwave Sounder (ATMS) and the Cross-track Infrared Sounder (CrIS) work together to provide global high-resolution profiles of temperature and moisture. These advanced atmospheric sensors create cross-sections of storms and other weather conditions, helping with both short-term nowcasting and long-term forecasting. Level 1C data was collected from GES DISC at `https://disc.gsfc.nasa.gov/datasets?page=1`.

HIRS: The High Resolution Infrared Sounder (HIRS) operates at 20 channels (19 channels in the infrared and one in the visible). Its main purpose is to provide input for the vertical temperature and humidity profile retrievals. In addition, the HIRS pixel resolution serves as the standard grid resolution for all ATOVS level 2 products. Level 1B data was collected from EUMETSAT at `https://archive.eumetsat.int/usc/UserServicesClient.html`.

MHS: The Microwave Humidity Sounder (MHS) is a 5 channel instrument used to provide input to the retrieval of surface temperatures, emissivities, and atmospheric humidity. In combination with AMSU-A information it can also be used to process precipitation rates and related cloud properties, as well as to detect sea ice and snow coverage. Level 1B data was collected from EUMETSAT at `https://archive.eumetsat.int/usc/UserServicesClient.html`.

ATMS-Precipitation: The ATMS-Precipitation is one of the products of the Global Precipitation Measurement (GPM) mission. It is based on the L1C-level calibrated brightness temperature data of the ATMS sensor and extracts information such as precipitation rate and precipitation type through a physical inversion algorithm. Level 2A data was collected from GES DISC at `https://disc.gsfc.nasa.gov/datasets?page=1`.

G Training details

Our DAWP framework is trained in 3 stages on 4 A100 80G GPUs, including training mask ViT-VAEs for encoding and mapping, an MMAE for data assimilation in observation space, and a spatiotemporal transformer for direct observation prediction. Specifically, these modules are all trained within 144×144 sub-images. The encoder and decoder of the mask ViT-VAE use the same

Table 7: Hyperparameters for training the mask ViT-VAE of DAWP on the composite dataset.

Hyper-parameter	Value
Learning rate	0.0001
β_1	0.9
β_2	0.999
Weight decay	0.00001
Batch size	200
Training steps	200000
Warm up percentage	10%
Warmup learning rate	0.000001
Learning rate decay	Cosine
Min learning rate	0.000001
KL-loss weight	0.000001

Table 8: Hyperparameters for training the AIDA module of DAWP on the composite dataset.

Hyper-parameter	Value
Learning rate	0.0001
β_1	0.9
β_2	0.999
Weight decay	0.00001
Batch size	48
Training steps	200000
Warm up percentage	10%
Warmup learning rate	0.000001
Learning rate decay	Cosine
Min learning rate	0.000001

Table 9: Hyperparameter for training the AIWP module of DAWP on the composite dataset.

Hyper-parameter	Value
Learning rate	0.0001
β_1	0.9
β_2	0.999
Weight decay	0.00001
Batch size	8
Training steps	200000
Warm up percentage	10%
Warmup learning rate	0.000001
Learning rate decay	Cosine
Min learning rate	0.000001

transformer structure with a patch size of 16 and a hidden dimension of 768. It is trained with a reconstruction MAE loss and a KL loss weight of 0.000001 for robust representation. We freeze the pretrained mask ViT-VAE as the encoders for each modality in our MMAE. Each modality observation with a 12h time window in a 144×144 sub-image is tokenized into 972 spatiotemporal tokens. Thus, our MMAE totally received 3888 tokens. We randomly select 128 observed tokens of them (3.3%) to reconstruct the remaining observed tokens via MAE training. Given the 144× 144 sub-images assimilated by MMAE, our spatiotemporal transformer is trained. It is structured with 12 TS spatiotemporal decoupling blocks, whose hidden dimension is 768. The hyperparameters for optimizing these modules are similar. All of them use the AdamW optimizer with $\beta_0 = 0.9$, $\beta_1 = 0.999$, and a learning rate of 0.0001. The learning rate is scheduled by a cosine scheduler, warming up 10k steps, step by step.

In the table 13, we present the computation cost.

Table 10: The details of the mask ViT-VAE model on different satellite datasets. `Conv16` $\times$ `16` is the 2D convolutional layer with 16×16 kernel. The `FFN` consists of two `Linear` layers separated by a `GeLU` activation layer [57]. The operator SamplePosterior samples a latent representation from μ and σ as SD did [55].

Module	Layer	Resolution	Channels
Input	-	144×144	c
PatchEmbed	`Conv16` $\times$ `16`	9×9	$c \rightarrow 768$
	`Flatten`	$9 \times 9 \rightarrow 81$	768
	`PosEmbed`	81	768
Trnasformer Block $\times$ 10	`LayerNorm`	81	768
	`MaskAttention`	81	768
	`LayerNorm`	81	768
	`FFN`	81	768
Qauntify	`TransformerBlock`	81	768
	`TransformerBlock`	81	768
	`Concat`	81	$768 \rightarrow 1536$
	`Linear`	81	$1536 \rightarrow 8c$
	`SamplePosterior`	81	$8c \rightarrow 4c$
	`Linear`	81	$4c \rightarrow 768$
Trnasformer Block $\times$ 12	`LayerNorm`	81	768
	`MaskAttention`	81	768
	`LayerNorm`	81	768
	`FFN`	81	768
Out	`Rearrange`	$81 \rightarrow 9 \times 9$	768
	`Conv1` $\times$ `1`	9×9	$768 \rightarrow 256c$
	`Rearrange`	$9 \times 9 \rightarrow 144 \times 144$	$256c \rightarrow c$
	`Conv3` $\times$ `3`	144×144	c

H Metrics defination

H.1 CSI and FAR

For the evaluation of global precipitation variables, the metrics include the Critical Success Index (CSI) and False Alarm Ratio (FAR). They are core binary classification evaluation metrics that quantify the detection accuracy and reliability of precipitation events. In the field of meteorology, these metrics assess the consistency and accuracy between precipitation predictions and observed results, quantitatively evaluating the performance of models. To measure the accuracy of prediction for precipitation with different intensities. Before calculating these metrics, we transform the predicted pixel values and ground truth into binary values (0 or 1) using a given threshold τ. The value is set to 0 if it is less than τ; otherwise, it is set to 1. These binary values enable us to determine the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) counts. CSI, HSS, and FSS are calculated by these counts as follows:

1) Critical Success Index. CSI is a metric that evaluates the proportion of correctly predicted events of hits among conditions, including hits (TP), false alarms (FN), and misses (FP). The formulation of CSI is:

$$CSI = \frac{TP}{TP + FN + FP} \tag{3}$$

The value of CSI ranges from 0 to 1. Higher values indicate better prediction accuracy.

2) False Alarm Ratio. The FAR metric quantifies the proportion of predicted positive events that were actually negative in meteorological verification, emphasizing the reliability of alarm triggers. It is

Table 11: The details of the MMAE on encoded multimodal tokens within a sub-image in a time window 12. `Conv1` $\times$ 1 is the 2D convolutional layer with 1×1 kernel. (c_1, c_2, c_3, c_4) means a multimodal input list with input channels c_1, c_2, c_3, and c_4. The MaskTokens is similar to the function of random_mask in [34] , while adding the [EOS] tokens to keep the sequences from different samples the same length. The operator of PaddingTokens fills the feature map as [34] did. The `FFN` consists of two `Linear` layers separated by a `GeLU` activation layer [57].

Module	Layer	Resolution	Channels
Multimodal Input	-	$9 \times 9 \times 12$	(c_1, c_2, c_3, c_4)
Multimodal PatchEmbed	`Conv1` $\times$ 1	$9 \times 9 \times 12$	$(c_1, c_2, c_3, c_4) \rightarrow (768, 768, 768, 768)$
	`Flatten`	$9 \times 9 \times 12 \rightarrow 81 \times 12$	$(768, 768, 768, 768)$
	`PosEmbed`	81×12	$(768, 768, 768, 768)$
	`TemporalEmbed`	81×12	$(768, 768, 768, 768)$
	`Rearrange`	$81 \times 12 \rightarrow 3888$	$(768, 768, 768, 768) \rightarrow 768$
Random Masking	`MaskTokens`	$3888 \rightarrow 128$	768
Trnasformer Block $\times$ 12	`LayerNorm`	128	768
	`MaskAttention`	128	768
	`LayerNorm`	128	768
	`FFN`	128	768
Feature Map Filling	`Linear`	128	$768 \rightarrow 512$
	`PaddingTokens`	$128 \rightarrow 3888$	512
	`Rearrange`	$3888 \rightarrow 81 \times 12 \times 4$	512
	`PosEmbed`	$81 \times 12 \times 4$	512
	`TemporalEmbed`	$81 \times 12 \times 4$	512
	`Rearrange`	$81 \times 12 \times 4 \rightarrow 3888$	512
Trnasformer Block $\times$ 8	`LayerNorm`	3888	512
	`Attention`	3888	512
	`LayerNorm`	3888	512
	`FFN`	3888	512
Multimodal Out	`Rearrange`	$3888 \rightarrow 972$	$(512, 512, 512, 512)$
	`LayerNorm`	972	$(512, 512, 512, 512)$
	`Linear`	972	(c_1, c_2, c_3, c_4)
	`Rearrange`	$972 \rightarrow 9 \times 9 \times 12$	(c_1, c_2, c_3, c_4)

Table 12: The details of our AIWP module training in the assimilated space. The inputs of this module are sub-images with 8 neighbours in a multimodal way. (c_1, c_2, c_3, c_4) means a multimodal input list with input channels c_1, c_2, c_3, and c_4. $Tview$ and $Sview$ indicate the temporal dimension and the spatial dimension as the sequence, respectively. The `FFNwithSwiGLU` consists of two `Linear` layers separated by a `SwiGLU` activation layer [58].

Module	Layer	Resolution	Channels
Input with Conditions	-	$27 \times 27 \times 12$	(c_1, c_2, c_3, c_4)
PatchEmbed	`Concat`	$27 \times 27 \times 12$	$(c_1, c_2, c_3, c_4) \rightarrow c_1 + c_2 + c_3 + c_4$
	`Linear`	$27 \times 27 \times 12$	$c_1 + c_2 + c_3 + c_4 \rightarrow 768$
	`Rearrange`	$27 \times 27 \times 12 \rightarrow 729 \times 12$	768
	`PosEmbed`	729×12	768
TS Block $\times$ 12	`Tview`	$729 \times 12 \rightarrow (729\times)12$	768
	`Attention`	$(729\times)12$	768
	`LayerNorm`	$(729\times)12$	768
	`FFNwithSwiGLU`	$(729\times)12$	768
	`LayerNorm`	$(729\times)12$	768
	`Sview`	$(729\times)12 \rightarrow (12\times)729$	768
	`Attention`	$(12\times)729$	768
	`LayerNorm`	$(12\times)729$	768
	`FFNwithSwiGLU`	$(12\times)729$	768
	`LayerNorm`	$(12\times)729$	768
Multimodal Out	`Rearrange`	$(12\times)729 \rightarrow 7 \times 27 \times 12$	768
	`LayerNorm`	$7 \times 27 \times 12$	768
	`Linear`	972	(c_1, c_2, c_3, c_4)

Table 13: Computation cost during inference.

	Inference time(ms)	Parameters(MB)	Memory(MB)	Batch size(per GPU)
Mask-ViT-VAE	53	96	7262	50
AIDA	310	105	18242	12
AIWP	491	216	47134	2

defined as:

$$FAR = \frac{FP}{FP + TP} \tag{4}$$

where FP denotes false positive predictions (e.g., forecasted rainfall with no ground observation) and TP represents true positives (correctly predicted rainfall events). FAR ranges from 0 (perfect reliability) to 1 (all alarms are false), with lower values indicating better prediction specificity.

H.2 MAE

To evaluate the accuracy of direct observation predictions, we use a pointwise Mean Absolute Error (MAE) as the metric to calculate errors on the ground truth with variable missing values. It is worth noting that the MAE is calculated with the raw observation point by point to ignore the influence of missing values. It is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{5}$$

N is the total number of points with observation, y_i is the ground truth at the i^{th} location, and $\hat{y}_i$ is the prediction.

I More results of direct observation predictions

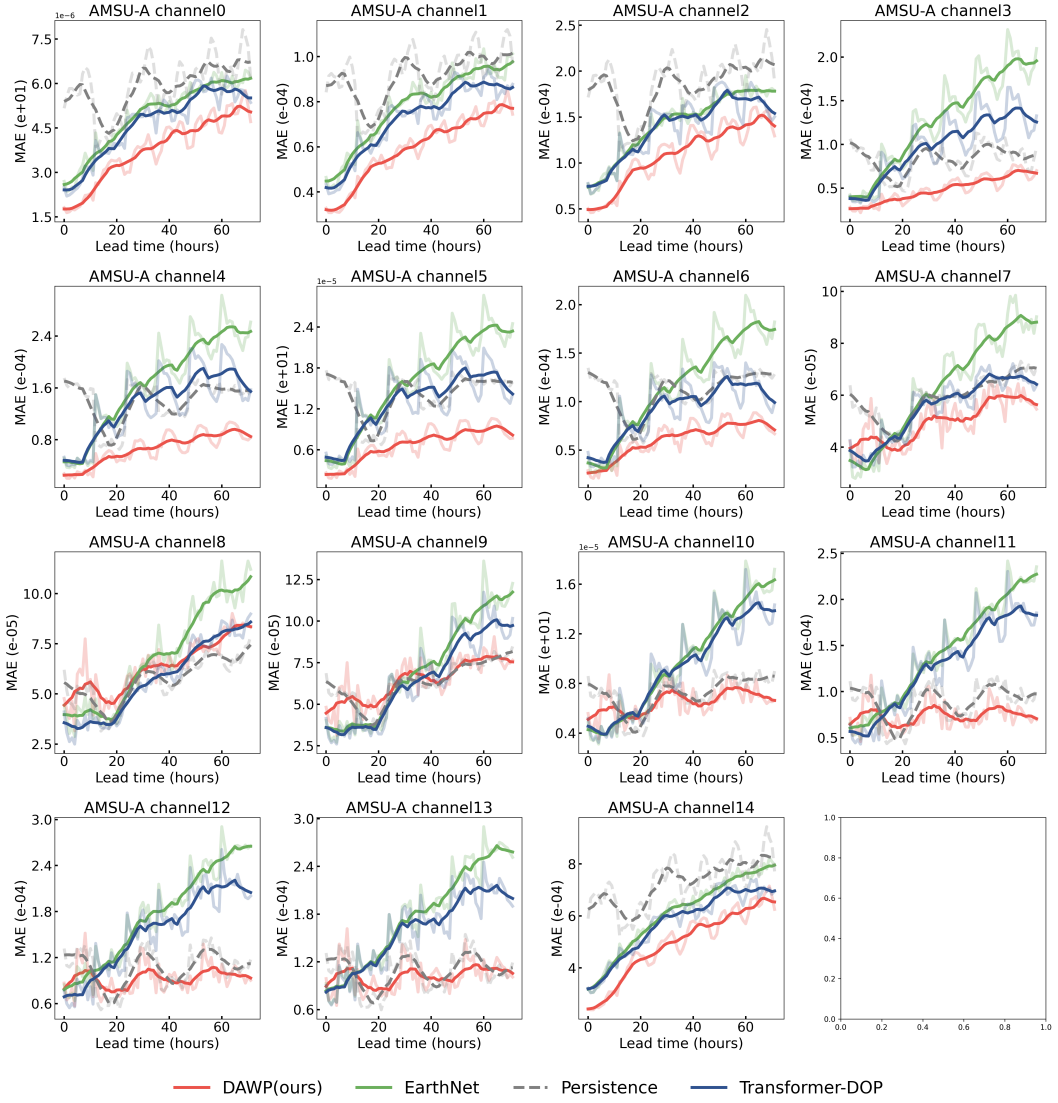

Figure 10: Curves of MAE for the prediction of different channels in sensor AMSU-A. The max leadtime is 72h with a 1h temporal resolution.

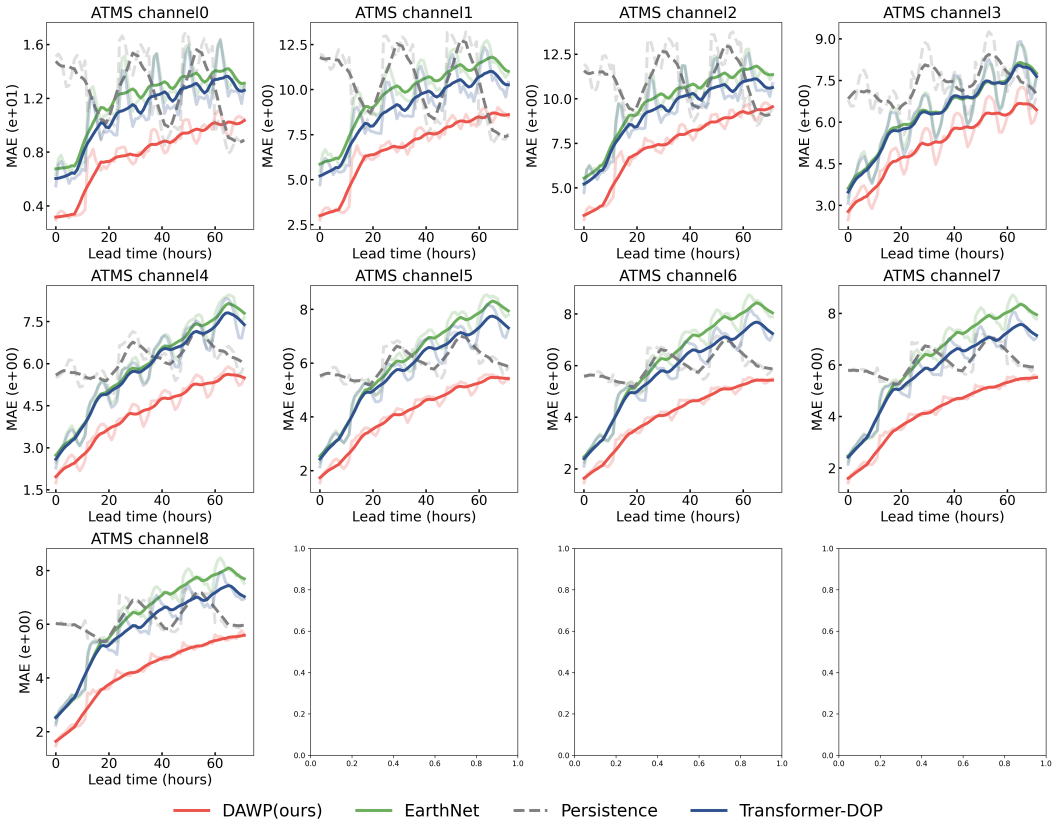

Figure 11: Curves of MAE for the prediction of different channels in sensor ATMS. The max leadtime is 72h with a 1h temporal resolution.

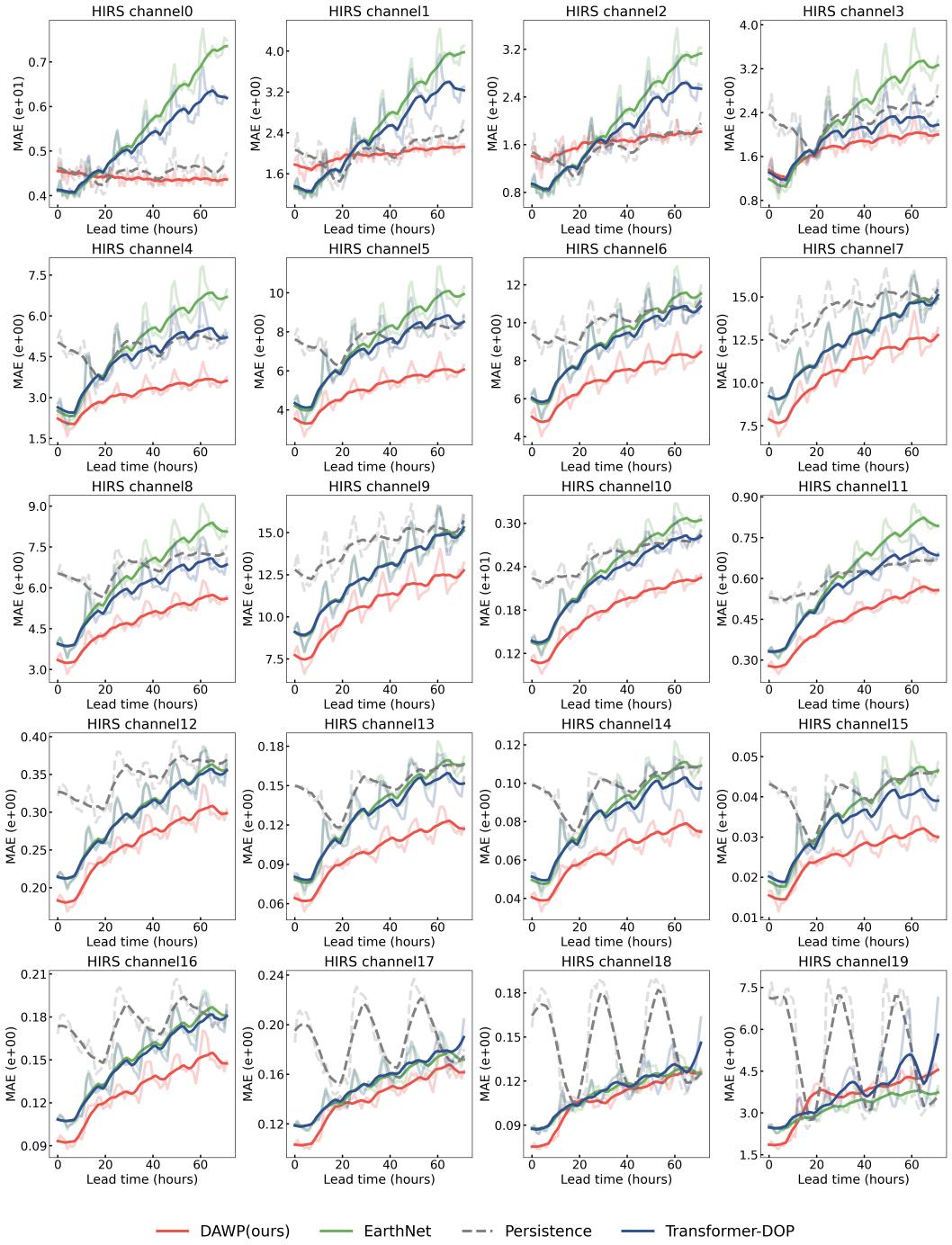

Figure 12: Curves of MAE for the prediction of different channels in sensor HIRS. The max leadtime is 72h with a 1h temporal resolution.

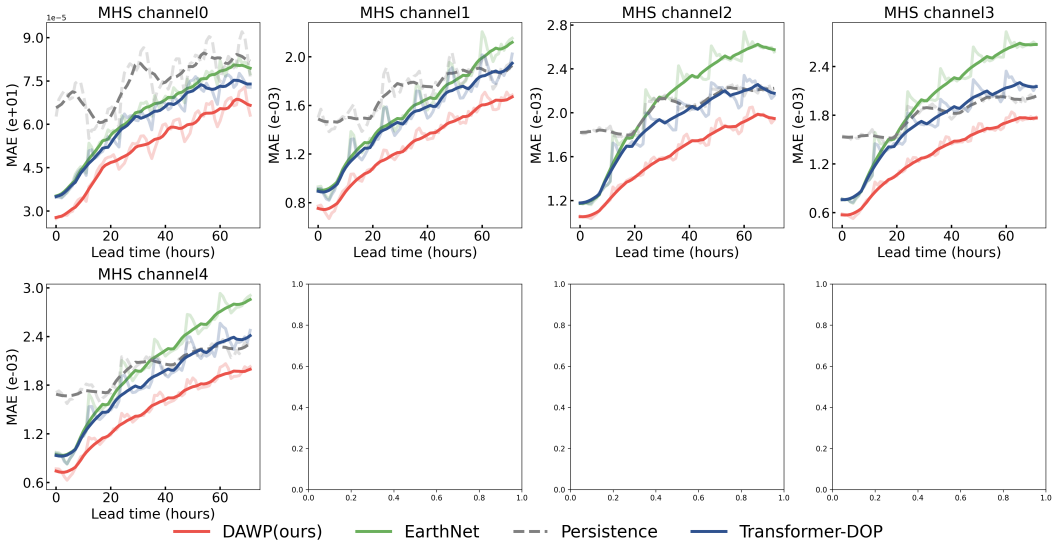

Figure 13: Curves of MAE for the prediction of different channels in sensor MHS. The max leadtime is 72h with a 1h temporal resolution.

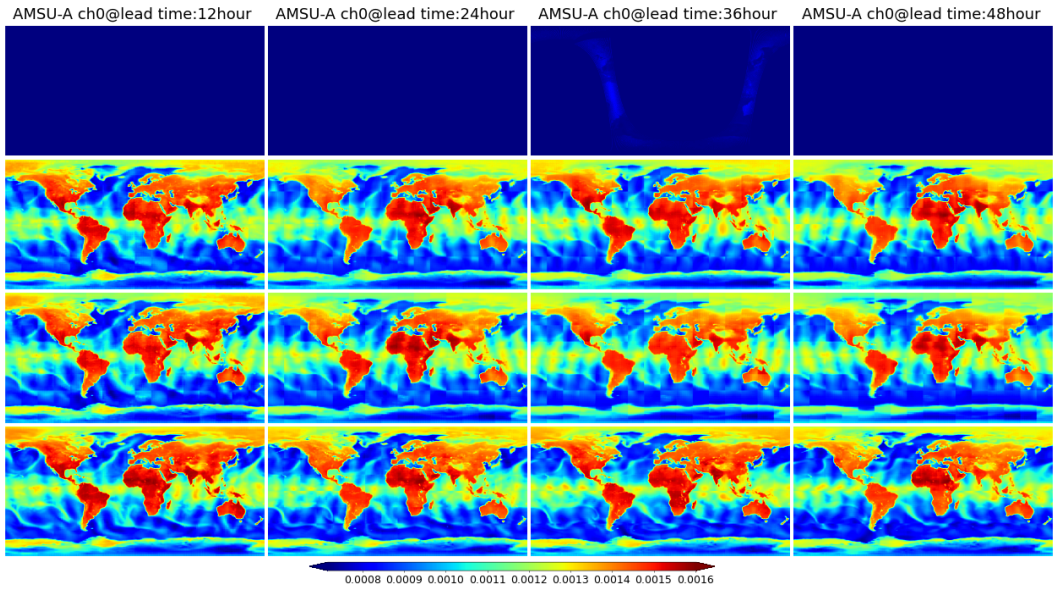

Figure 14: A visualization of rollout predictions for channel 0 of AMSU-A. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

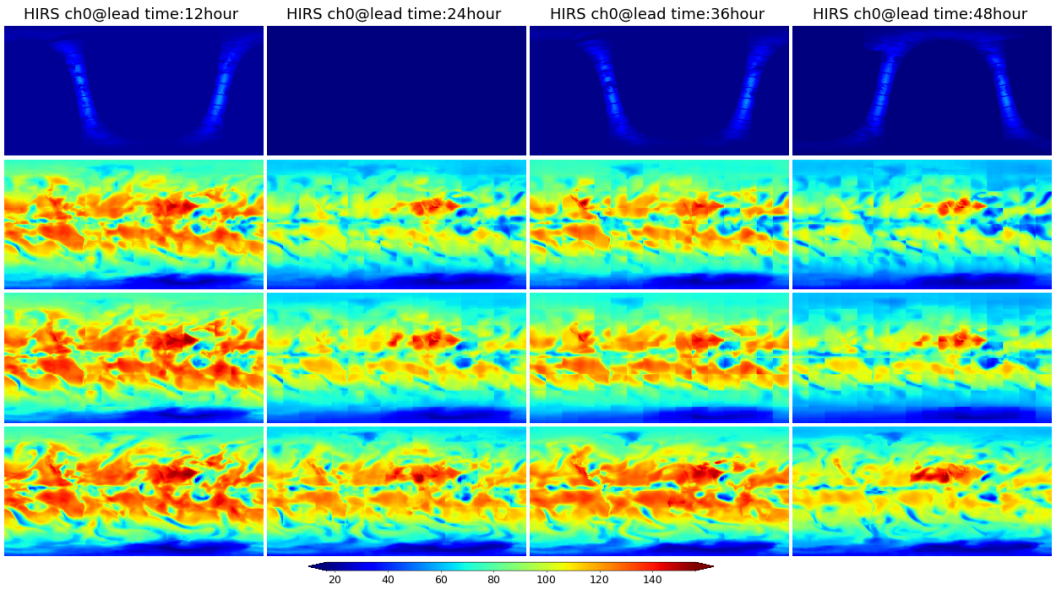

Figure 15: A visualization of rollout predictions for channel 9 of HIRS. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

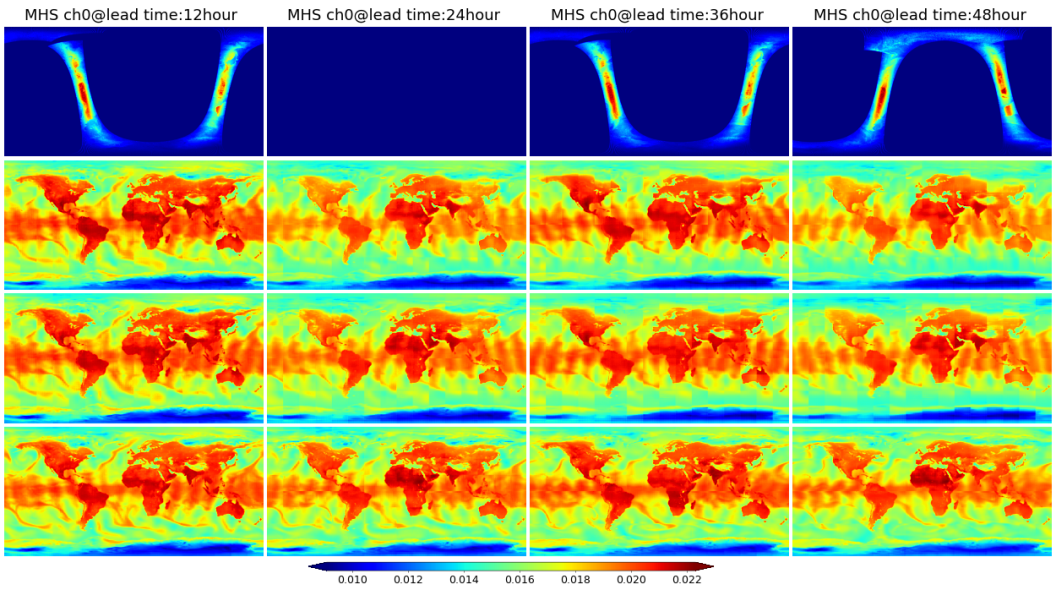

Figure 16: A visualization of rollout predictions for channel 0 of MHS. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

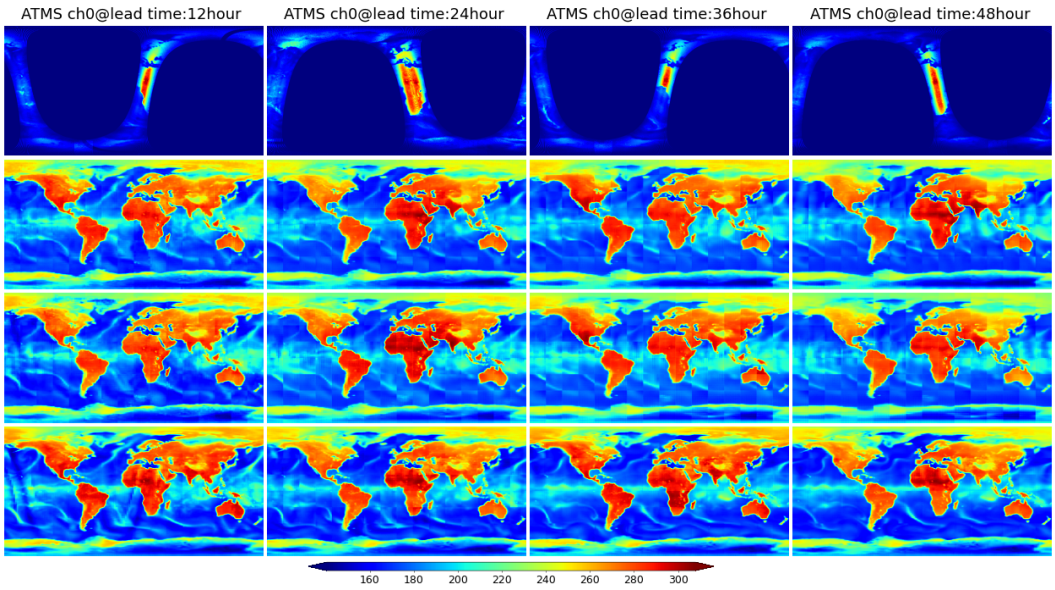

Figure 17: A visualization of rollout predictions for channel 1 of ATMS. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

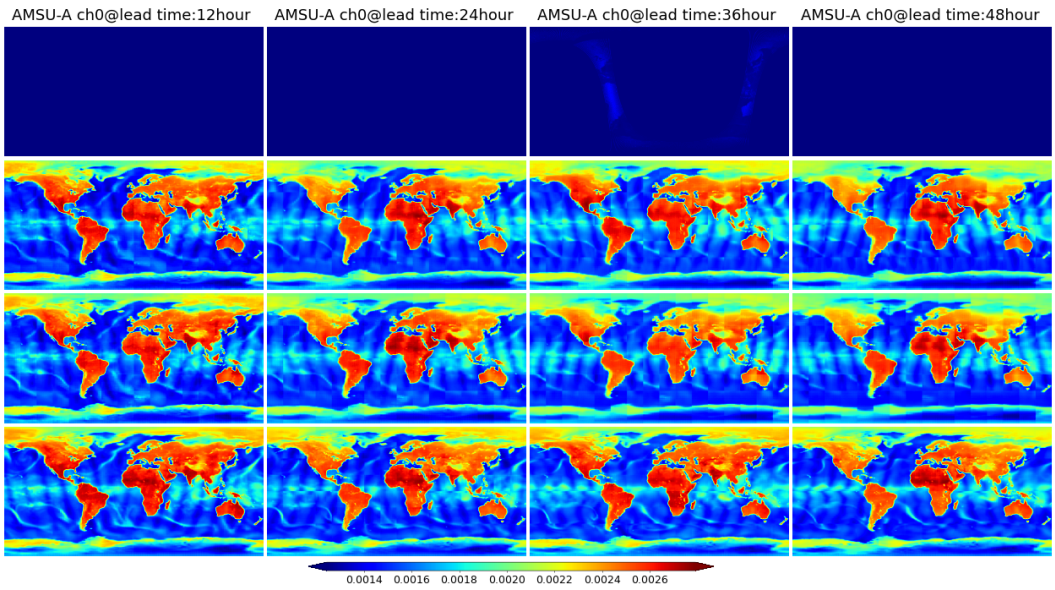

Figure 18: A visualization of rollout predictions for channel 1 of AMSU-A. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

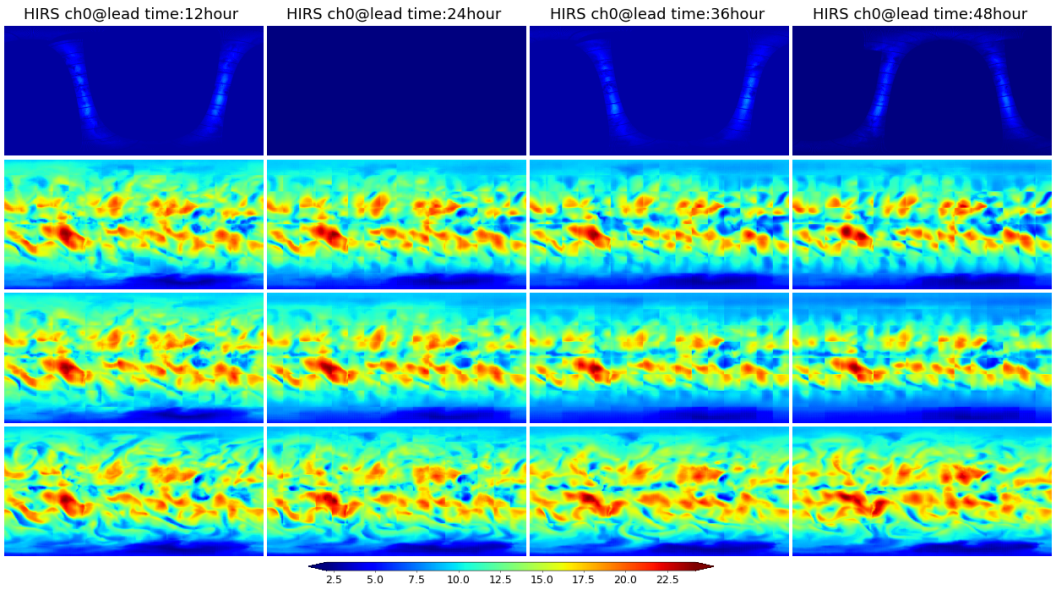

Figure 19: A visualization of rollout predictions for channel 10 of HIRS. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).

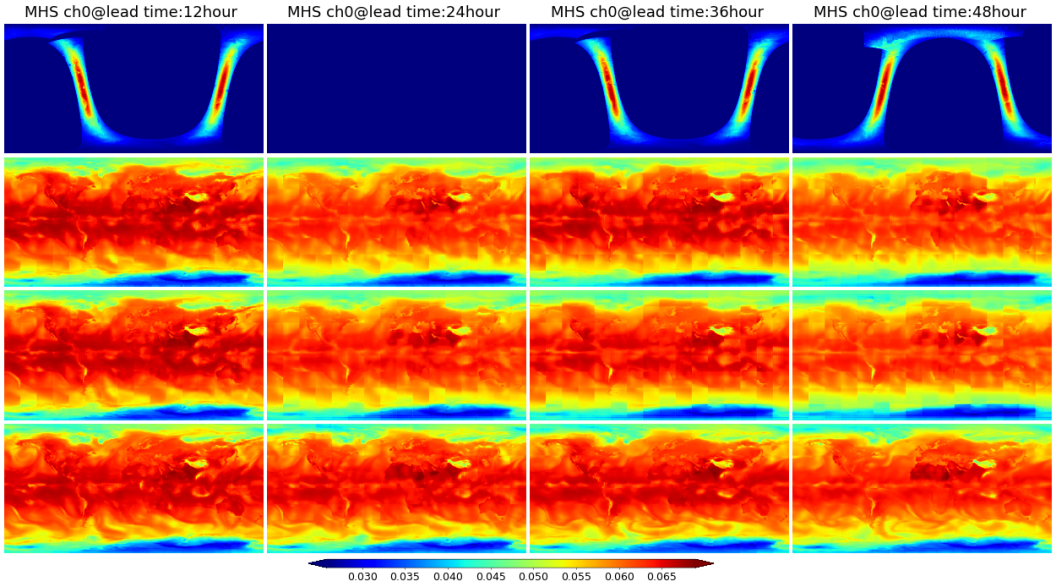

Figure 20: A visualization of rollout predictions for channel 1 of MHS. From top to bottom are the results of ground truth, Transformer-DOP, EarthNet, and DAWP (ours).