ATMOSARENA: BENCHMARKING FOUNDATION MODELS FOR ATMOSPHERIC SCIENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has emerged as a powerful tool for atmospheric sciences, showing significant utility across various tasks in weather and climate modeling. In line with recent progress in language and vision foundation models, there are growing efforts to scale and finetune such models for multi-task spatiotemporal reasoning. Despite promising results, existing works often evaluate their model on a small set of non-uniform tasks, which makes it hard to quantify broad generalization across diverse tasks and domains. To address this challenge, we introduce `AtmosArena`, the first multi-task benchmark dedicated to foundation models in atmospheric sciences. `AtmosArena` comprises a suite of tasks that cover a broad spectrum of applications in atmospheric physics and atmospheric chemistry. To showcase the capabilities and key features of our benchmark, we conducted extensive experiments to evaluate two state-of-the-art deep learning models, ClimaX and Stormer on `AtmosArena`, and compare their performance with other deep learning and traditional baselines. By providing a standardized, open-source benchmark, we aim to facilitate further advancements in the field, much like open-source benchmarks have driven the development of foundation models for language and vision.

1 INTRODUCTION

Modeling of large-scale atmospheric systems is an omnipresent challenge for science and society. Traditionally, numerical methods are the dominating approach in atmospheric sciences, which operationalize rigorous systems of differential equations to simulate such phenomena (Lynch, 2008; Bauer et al., 2015). Despite their widespread use in practice, numerical methods suffer from many challenges, such as inadequate resolution of important small-scale physical processes and substantial computational demands (Balaji et al., 2017; Lavers et al., 2022; Leung et al., 2003; Rauscher et al., 2010). Deep learning has emerged as a powerful complement due to its ability to learn complex systems from historical data and produce fast predictions within seconds. Deep learning methods have proven great utility and performance across various atmospheric tasks, including but not limited to precipitation nowcasting (Ravuri et al., 2021b; Sønderby et al., 2020; Andrychowicz et al., 2023), medium-range weather forecasting (Weyn et al., 2020; Rasp & Thuerey, 2021; Keisler, 2022; Pathak et al., 2022b; Bi et al., 2022; Lam et al., 2023; Nguyen et al., 2023c; Chen et al., 2023b;a; Kochkov et al., 2023), climate projection (Watson-Parris et al., 2022b), climate downscaling (Baño Medina et al., 2020; Liu et al., 2020; Nagasato et al., 2021; Rodrigues et al., 2018; Sachindra et al., 2018; Vandal et al., 2019), air pollution forecasting (Ayturan et al., 2018; Bekkar et al., 2021; Tao et al., 2019; Bui et al., 2018; Heydari et al., 2022), and greenhouse gas emission prediction (Hamrani et al., 2020; Bakay & Ağbulut, 2021; Altikat, 2021).

Recent years have witnessed a paradigm shift from training task-specific models to developing foundation models for atmospheric sciences (Nguyen et al., 2023a; Bodnar et al., 2024), similar to models such as GPT-x (Brown et al., 2020; Achiam et al., 2023) in natural language processing, or CLIP (Radford et al., 2021) in computer vision. These foundation models are trained on large-scale and diverse datasets, enabling them to develop a rich, general understanding of the atmosphere. Once pre-trained, they can adapt efficiently to various downstream tasks, ranging from weather nowcasting to long-term climate projections, via lightweight finetuning. This approach is particularly attractive for atmospheric sciences, where there is an increasing availability of high-quality datasets and tasks have non-trivial global and regional structure.

Table 1: Comparisons between AtmosArena and existing works that consider multiple atmospheric tasks. AtmosArena offers the most comprehensive set of tasks, data, and evaluation metrics.

Benchmark	Tasks	Data	Metrics
AtmosArena	Weather forecasting	ERA5	RMSE, ACC
	S2S forecasting	ERA5	RMSE, ACC, Spectral Div
	Climate data infilling	ERA5, Berkeley Earth	Bias, RMSE
	Climate model emulation	ClimateBench	Spatial, Global, Total, RMSE
	Climate downscaling	ERA5	RMSE, Bias, Pearson
	Extreme weather events detection	ClimateNet	IoU, Precision, Recall, F-1
ClimateLearn	Weather forecasting	ERA5	RMSE, ACC
	Downscaling	ERA5	RMSE, Bias, Pearson
	Projection	ClimateBench	Spatial, Global, Total, RMSE
ClimaX	Weather forecasting	ERA5	RMSE, ACC
	S2S forecasting	ERA5	RMSE, ACC
	Climate model emulation	ClimateBench	Spatial, Global, Total, RMSE
	Climate downscaling	ERA5	RMSE, Bias, Pearson
Aurora	Weather forecasting	HRES Analysis	RMSE, ACC
	Air composition forecasting	CAMS Analysis	RMSE, ACC

Standardized open-source benchmarks are crucial for the advancement of foundation models. In language, benchmarks such as HeLM (Liang et al., 2022), LLM Foundry, LM Evaluation Harness (Gao et al., 2023), and Big Bench (Srivastava et al., 2022) have aided researchers to systematically evaluate the performance of large language models. Similarly, for perception, comprehensive benchmarks such as VQA (Antol et al., 2015), SciBench (Wang et al., 2023), MMMU (Yue et al., 2023), and MathVista (Lu et al., 2023), have significantly accelerated research in multimodal foundation models. In stark contrast, there is no standardized multi-task benchmark for benchmarking atmospheric foundation models and existing works (Nguyen et al., 2023a; Bodnar et al., 2024) limit their evaluation to a relatively small set of non-overlapping tasks, which creates challenges in objective assessment of progress in the field.

To address this gap, we introduce AtmosArena, an open-source benchmark for foundation models in atmospheric sciences. To the best of our knowledge, AtmosArena is the first of its kind to offer a comprehensive evaluation framework tailored for this domain. AtmosArena encompasses a suite of tasks that span a wide spectrum of problems from both atmospheric and machine learning perspectives. Each task within AtmosArena is supported by datasets, fine-tuning protocols, evaluation code, standardized evaluation metrics, and a collection of deep learning and traditional baselines. This suite not only facilitates a fair assessment of model performance but also serves as a crucial tool for identifying opportunities for future development in the field. AtmosArena aims to set a new standard in the evaluation of atmospheric models, providing a solid foundation for the development of new methodologies. Table 1 summarizes the tasks, datasets, and metrics supported by AtmosArena.

To showcase the utility of AtmosArena, we conduct extensive experiments across all tasks included in the benchmark. We test and compare three representative classes of models: (1) deep learning with no pretraining, (2) single-source pretraining, and (3) multi-source pretraining. We also include traditional methods as simple baselines. To ensure fairness, we maintained consistent fine-tuning and evaluation settings across all models. The experimental results indicate that pretrained models generally outperform baselines without pretraining in most tasks. However, no single model consistently dominates across all tasks. This underscores the comprehensiveness of AtmosArena and highlights potential opportunities for future model development. In line with our commitment to openness and reproducibility, we will make all our data, code, and model checkpoints publicly available.

2 Related Work

Deep Learning for Atmospheric Sciences Deep learning has revolutionized atmospheric sciences in recent years in both speed and accuracy. In weather forecasting, notable models like Pangu (Bi et al., 2022), Graphcast (Lam et al., 2023), and Stormer (Nguyen et al., 2023c) have surpassed the accuracy of the gold-standard IFS HRES system. This progress spans from simple models like ResNet (Rasp & Thuerey, 2021) to advanced architectures such as Graph Neural Networks (Keisler, 2022; Lam et al., 2023), Fourier neural operators (Pathak et al., 2022a), and Transformers (Bi et al., 2022; Nguyen et al., 2023a; Chen et al., 2023c;a; Nguyen et al., 2023c). In addition to medium-range, other works focus on forecasting at different time scales, such as nowcasting (Sønderby et al., 2020; Ravuri et al., 2021a; Andrychowicz et al., 2023) or longer-term prediction tasks (Watt-Meyer et al., 2023; Mouatadid et al., 2023). To account for uncertainty, recent works have also proposed ensemble forecasting with hybrid-physics models (Kochkov et al., 2024) or diffusion (Price et al., 2024), which are particularly useful for extreme event prediction like heavy rainfall (Zhang et al., 2023) and floods (Nearing et al., 2024).

Foundation Models for Atmospheric Sciences ClimaX (Nguyen et al., 2023a) is the first foundation model for weather and climate, pretrained on five simulated datasets from CMIP6 and finetuned on four downstream tasks. Aurora (Bodnar et al., 2024) is the latest atmospheric foundation model which scaled up pretraining to larger models, more data, and finer grid resolutions. Aurora was shown to achieve state-of-the-art performance in operational weather forecasting and air composition forecasting. In addition to atmospheric sciences, the development of scientific foundation models for physical domains is growing quickly as a field. For example, recent works in Partial Differential Equations (PDEs) modeling have proposed to pretrain large-scale models for micro-scale dynamical systems that can transfer in a zero-shot or few-shot fashion to unseen equations (Sun et al., 2024; Herde et al., 2024; Alkin et al., 2024; McCabe et al., 2023).

Atmospheric Datasets and Benchmarks Standardized benchmarks fuel the growth of atmospheric deep learning. WeatherBench (Rasp et al., 2020a; 2023) provides data, metrics, baselines, and a leaderboard for medium-range weather forecasting. Another common data source for weather forecasting is CMIP6 (Eyring et al., 2016b) which provides a large collection of simulation runs from climate models. SubseasonalClimateUSA (Mouatadid et al., 2024) and ChaosBench (Nathaniel et al., 2024) are two recent benchmarks that have been proposed to push the forecasting capabilities to sub-seasonal and seasonal time scales. Beyond forecasting, standard datasets have been developed for a diverse set of tasks in weather and climate, including climate emulation (Kaltenborn et al., 2023), sub-resolution physics modeling (Yu et al., 2024), precipitation prediction (de Witt et al., 2020; Sit et al., 2021), extreme weather events detection and localization (Rahnemoonfar et al., 2021; Requena-Mesa et al., 2021; Minixhofer et al., 2021; Prabhat et al., 2021; Racah et al., 2017), natural disaster-related tasks (Proma et al., 2022), atmospheric radiative transfer (Cachay et al., 2021), long-term global trends prediction (Watson-Parris et al., 2022b), cloud classification (Rasp et al., 2020b), nowcasting (Franch et al., 2020), tropical cyclone intensity prediction (Maskey et al., 2020), air quality metrics prediction (Betancourt et al., 2021), hydrometeorological time series analysis (Villaescusa-Navarro et al., 2022), and river flow analysis (Godfried et al., 2020). Beyond plain datasets, libraries such as ClimateLearn (Nguyen et al., 2023b), Scikit-downscale Hamman & Kent (2020), CCdownscaling Polasky et al. (2023), and CMIP6-Downscaling CarbonPlan (2022) provide software for training deep learning methods for various tasks in atmospheric sciences.

3 Key Components of AtmosArena

As a first benchmark, we aim to build a comprehensive suite of tasks in atmospheric sciences, emphasizing diversity from both domain-specific and machine learning perspectives. Domain-wise, tasks are broadly classified into atmospheric physics or atmospheric chemistry. Atmospheric physics focuses on physical variables like temperature, humidity, and wind, essential for modeling weather patterns in the short-term and climate trends in the longer term. Atmospheric chemistry, on the other hand, focuses on the composition and transformation of atmospheric constituents, such as pollutants like carbon monoxide and dioxide, crucial for studying air quality and environmental health.

Due to space constraints, this section presents the six tasks under atmospheric physics: Medium-range Weather Forecasting, S2S Forecasting, Extreme Weather Events Detection, Climate Downscaling, Climate Data Infilling, and Climate Model Emulation. Tasks related to atmospheric chemistry are detailed in Appendix G. From a machine learning perspective, many common predictive tasks in atmospheric sciences can be mapped to well-defined problems in machine learning. Within this perspective, our benchmark can be seen as spanning five distinct categories of tasks defined on a grid: forecasting, segmentation, super-resolution, inpainting, and counterfactual prediction. This diverse suite of tasks allows us to obtain a holistic evaluation of atmospheric foundation models.

3.1 Tasks

Medium-range weather forecasting is the task of predicting the global weather conditions at a future time step $t + T$ given the weather conditions at or before the current step t, where the lead time T ranges from a few hours to two weeks. A deep learning model takes an input of shape $V \times H \times W$ and outputs a prediction of shape $V' \times H \times W$, in which V and V' are the numbers of input and output atmospheric variables, respectively, while $H \times W$ denotes the spatial resolution of the data.

Sub-seasonal-to-seasonal (S2S) forecasting is similar to medium-range forecasting but with a longer lead time range between 2 weeks and 2 months (Vitart & Robertson, 2018; Vitart et al., 2022). This task bridges the gap between weather forecasting and climate modeling and holds significant socioeconomic value in disaster mitigation, but has received much less attention than the other two well-established tasks. Since the weather becomes too chaotic for any model to perform accurate point prediction after two weeks, we instead task the models to forecast the average statistics of key variables over a two-week window.

Extreme weather events detection is the task of identifying weather patterns that may lead to extreme weather events, such as tropical cyclones and atmospheric rivers. Deep learning models are trained to perform pixel-level detection and segmentation of these events in climate data. Specifically, the input typically consists of key atmospheric variables, and the output is a segmented map where each pixel is classified as part of an extreme event or as background. This approach allows for precise quantification of the frequency, intensity, and spatial extent of extreme events under various climate scenarios, providing valuable insights for climate research and policy-making.

Climate downscaling is the task of improving the spatial resolution of climate model outputs, which typically operate on large grid cells due to their high computational demands. This refinement is crucial for accurately representing local phenomena and informing regional policy decisions. In this task, deep learning models transform an input grid of dimensions $V \times H \times W$ into a higher-resolution output $V' \times H' \times W'$, where $H' > H$ and $W' > W$.

Climate data infilling involves estimating missing or incomplete data in historical and current climate datasets. This task aims to provide a more comprehensive and continuous historical record of important atmospheric variables, such as near-surface air temperature, enabling robust climate analysis and modeling. In data infilling, deep learning models are trained to predict missing values by leveraging patterns found in available data. The typical input to these models includes incomplete datasets of dimensions $V \times H \times W$, and the output is a complete dataset of the same dimensions, where the previously missing values are estimated by the model.

Climate model emulation involves predicting the annual mean global distributions of crucial climate variables like surface temperature and precipitation indices, given different scenarios of anthropogenic forcing factors such as carbon dioxide (CO_2) and methane (CH_4). The input is a tensor of shape $T \times V \times H \times W$ which captures the forcing conditions over T consecutive years, and the output shape is $V' \times H \times W$. Unlike temporal forecasting, this task assesses a model's ability to predict the response of the climate system to varying levels of external factors, providing a foundation for long-term climate strategy and policy decisions.

3.2 Datasets

ERA5 maintained by ECMWF (Hersbach et al., 2020) is a common dataset for training and evaluating data-driven methods in atmospheric sciences (Bi et al., 2022; Lam et al., 2023; Nguyen et al., 2023c). ERA5 is a reanalysis dataset that provides the best guess of different climate variables at any point in time by integrating observational data with an advanced forecasting model known as the Integrated Forecasting System (IFS) (Wedi et al., 2015). ERA5 offers hourly data from 1979 to the present and at a $0.25°$ (721×1440) global grid, totaling nearly 400,000 data points at 37 different pressure levels. Given its extensive scale, we regrid the original data to $1.40625°$ (128×256) grid and consider data from 1979 to 2020 for training and evaluation. We use ERA5 for four tasks in `AtmosArena`, including medium-range weather forecasting, S2S forecasting, climate downscaling, and data infilling.

Berkeley Earth provides a variety of high-quality temperature data products that incorporate a large set of temperature observations (Rohde & Hausfather, 2020). In `AtmosArena`, we use the global monthly average temperature data at $1°$ (180×360) grid as an independent test dataset for the infilling task. We regrid the data to the common resolution of $1.40625°$.

ClimateBench is a benchmark for testing data-driven methods for climate model emulation (Watson-Parris et al., 2022b). ClimateBench consists of simulation outputs of the Norwegian Earth System Model (NorESM2) (Seland et al., 2020) from CMIP6 (Eyring et al., 2016a) that are run under different forcing scenarios for the period $2015 - 2100$. The dataset includes four input forcing factors – carbon dioxide (CO_2), sulfur dioxide (SO_2), black carbon (BC), and methane (CH_4), and the annual mean global distributions of four target variables – surface temperature, diurnal temperature range, precipitation, and the 90th percentile of precipitation.

ClimateNet is an expert-labeled dataset of tropical cyclones (TCs) and atmospheric rivers (ARs), two important weather patterns that may lead to extreme weather events (Prabhat et al., 2020). ClimateNet consists of 459 data points of simulation runs of the Community Atmospheric Model (CAM5.1) from $1996 - 2013$. Each data point has a spatial resolution of 768×1152 with a total of 16 atmospheric variables, and each pixel is labeled with one of three classes – TCs, ARs, and Background.

3.3 Models

We consider a state-of-the-art representative from three classes of models. Many other recent models would also benefit from this benchmark Bodnar et al. (2024); Price et al. (2024), but they are currently closed-source. Table 2 shows the inference FLOPs and the number of parameters of each baseline we consider in this paper. We maintain a public leaderboard at `https://atmosarena.github.io/leaderboard/` to allow open and fair evaluation of both open- and closed-source models.

Table 2: FLOPs and model size of different baselines considered in `AtmosArena`.

	ClimaX	Stormer	UNet
FLOPs	986.098B	7377.751B	969.404B
Parameter count	110.842M	468.752M	577.745M

Non-pretrained model We aim to provide state-of-the-art methods tailored to each specific task in `AtmosArena`. For tasks without an established baseline, we use UNet (Ronneberger et al., 2015) as the deep learning baseline. We chose UNet due to its excellent performance in various dense prediction tasks in computer vision, which resemble most of the atmospheric tasks in `AtmosArena`. The Unet models we train in the experiments have the same size of 500M parameters, for which we have performed extensive hyperparameters tuning to obtain a strong non-pretrained baseline.

Single-source pretrained model We include Stormer (Nguyen et al., 2023c), a state-of-the-art open-source deep learning model for medium-range weather forecasting. Stomer is a transformers-based architecture (Vaswani et al., 2017) trained on 6-hourly ERA5 data at $1.40625°$ resolution from 1979 to 2018. We chose Stormer since it was trained on the same spatial resolution as our datasets, and its simple architecture allows seamless finetuning on new tasks. Stormer has 400M parameters.

Multi-source pretrained model We include ClimaX (Nguyen et al., 2023a), the first large-scale atmospheric foundation model trained on multiple data sources. ClimaX was pretrained to perform temporal forecasting on five simulated datasets at $1.40625°$ from CMIP6 (Eyring et al., 2016a) and was shown to transfer well to various atmospheric tasks via finetuning. Since ClimaX and Stormer share similar transformer architectures and training objectives, comparing them helps examine if and when multi-source pretraining is beneficial to the model. ClimaX has 100M parameters.

3.4 Finetuning protocols

ClimaX and Stormer share a similar architecture, which consists of an embedding layer, a transformer backbone, and a prediction head. The embedding layer transforms an input of shape $V \times H \times W$ to a sequence of shape $(H/p \times W/p) \times D$, where $(H/p \times W/p)$ is the sequence length, p is the patch size, and D is the hidden dimension. The transformer backbone processes this sequence and outputs a sequence of the same shape, and finally the prediction head outputs a prediction of shape $V' \times H' \times W'$. We refer to the original papers for a detailed description of these models.

We consider two finetuning settings, one where we freeze the core transformer backbone, and the other where we finetune the entire network. The frozen setting helps examine the direct transferability of the pretrained backbone to new tasks without further training. In tasks where the input or target variables were unseen during pretraining, we replace the pretrained embedding layer and prediction head with newly initialized networks. For datasets having a different spatial resolution from pretraining data, we interpolate the pretrained positional embedding to match the new sequence length.

4 Benchmark Evaluation

This section evaluates different models on six atmospheric physics tasks described in Section 3.1. Through the experiments, we aim to showcase the breadth of `AtmosArena` and provide practical recommendations for finetuning atmospheric foundation models on new tasks. We refer to Appendix H for the atmospheric chemistry experiments. We also present infilling results on the Berkeley Earth dataset and regional case studies on S2S forecasting in Appendix H.

4.1 Medium-range weather forecasting

We compare ClimaX and Stormer with Graphcast (Lam et al., 2023) – a leading forecasting method, and Climatology – a simple baseline, on weather forecasting with lead times from 1 to 14 days. We consider six target variables: temperature at 2 meters (T2m), zonal (U10m) and meridional (V10m) wind at 10 meters, geopotential at 500hPa (Z500), temperature at 850hPa (T850), and specific humidity at 700hPa (Q700), which are commonly used to verify forecasting models in previous works. Since Stormer and Graphcast were trained specifically for forecasting, we roll-out the pretrained checkpoints to obtain forecasts at different lead times without further training. For ClimaX, we perform full finetuning for each specific lead time and target variable, following the protocol in the original paper. All deep learning methods are trained on ERA5 from 1979 to 2018 and tested on 2020. The same data split is used for other tasks unless noted otherwise.

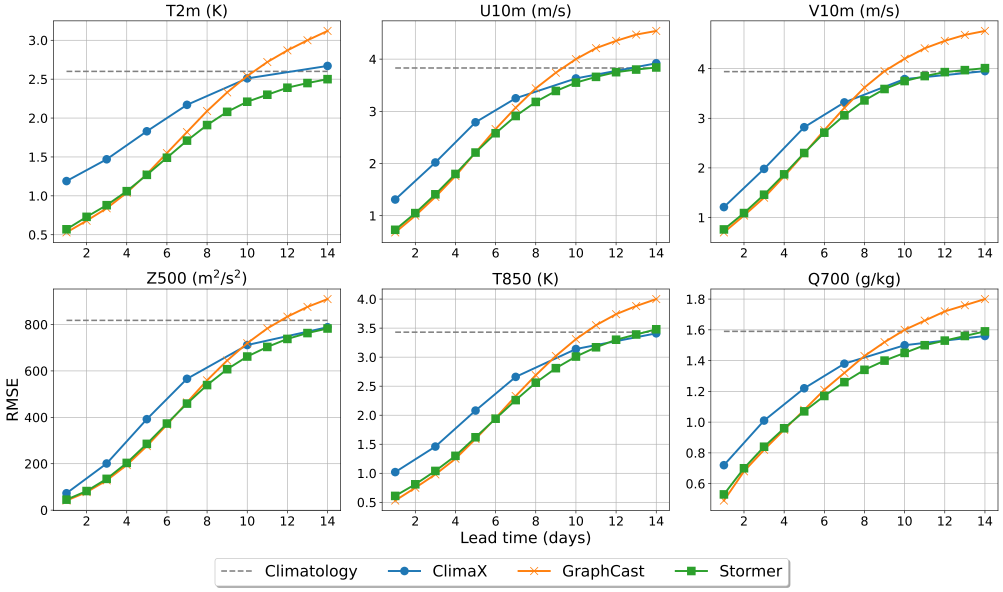

Figure 1: Medium-range weather forecasting performance measured by RMSE on six key variables at different lead times. Solid lines are deep learning models and the dashed line denotes the climatology baseline. Lower RMSE indicates better performance.

Figure 1 summarizes the RMSE results of this task (see Appendix for other metrics). Stormer is the best overall method, performing competitively with Graphcast at short lead times and much better at longer time scales. Graphcast works well for short lead times, but its performance degrades quickly and becomes worse than Climatology after day 10. ClimaX, on the other hand, performs poorly at small lead times but surpasses Graphcast at around day 10 and catches Stormer at day 14. This is because ClimaX performs direct forecasting which avoids error accumulation at long lead times.

4.2 Subseasonal-to-seasonal (S2S) forecasting

We evaluate ClimaX, Stormer, and Unet on forecasting the biweekly average statistics of four target variables – Z500, T850, T2m, and Q700. We consider two lead times of 2 weeks and 4 weeks, in which the average statistics are computed over weeks 3-4 and weeks 5-6, respectively. We construct the biweekly average data for training and evaluation from ERA5. For each baseline, we train two separate models to predict directly the average values at two different lead times. For ClimaX and Stormer, we consider two finetuning protocols where we either freeze (ClimaX frozen and Stormer frozen) or finetune (ClimaX finetuned and Stormer finetuned) the transformer backbone. Similar to medium-range weather forecasting, we include Climatology to examine if deep learning models achieve meaningful skills for S2S forecasting compared to this simple baseline.

Table 3: S2S performance measured by RMSE and ACC on four target variables at two lead times.

		Z500		T850		T2m		Q700	
		Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6
RMSE (↓)	ClimaX frozen	458.53	471.58	1.79	1.84	1.67	1.73	**0.69**	**0.70**
	ClimaX finetuned	**453.05**	469.92	**1.77**	**1.80**	1.65	1.70	**0.69**	0.71
	Stormer frozen	461.19	**467.37**	**1.77**	1.81	**1.56**	1.69	0.70	0.72
	Stormer finetuned	466.82	475.06	1.79	1.84	1.64	1.75	0.71	0.72
	Unet	498.46	521.32	1.90	2.09	1.63	2.29	0.74	0.75
	Climatology	475.58	475.58	2.00	2.00	1.61	**1.61**	0.76	0.76
ACC (↑)	ClimaX frozen	**0.84**	0.81	**0.92**	0.90	0.96	**0.95**	**0.86**	0.84
	ClimaX finetuned	**0.84**	0.81	**0.92**	0.90	0.95	0.94	**0.86**	0.84
	Stormer frozen	0.78	0.77	0.88	0.87	0.95	0.94	0.81	0.81
	Stormer finetuned	0.77	0.77	0.87	0.87	0.94	0.93	0.82	0.82
	Unet	**0.84**	**0.84**	**0.92**	**0.91**	**0.97**	0.93	0.85	**0.85**

Table 3 summarizes the results of S2S forecasting. In terms of RMSE, both ClimaX and Stormer have meaningful skills except for T2m, while Unet underperforms Climatology for most variables. Interestingly, the frozen version of ClimaX and Stormer performs competitively to their fully finetuned counterpart. This result highlights the importance of pretraining, which allows models to efficiently transfer to new forecasting tasks without further training of the transformer backbone. In terms of ACC, ClimaX and Unet perform similarly while Stormer lags behind. Overall, ClimaX outperforms Stormer in this task despite having a poorer performance on medium-range weather forecasting. This can be explained by the difference between the pretraining objective of the two models, where ClimaX was trained to perform forecasting at much longer horizons (6 hours to 1 week) compared to Stormer (6 hours to 1 day).

4.3 Climate downscaling

We consider the task of downscaling for six key variables: Z500, T850, T2m, Q700, U10m, and V10m. We use ERA5 at $5.625°$ as the low-resolution input, and ERA5 at $1.40625°$ as the high-resolution target, corresponding to $4\times$ upsampling. We include Unet as a deep learning baseline in addition to the two finetuning versions of ClimaX and Stormer. We report RMSE and Absolute Mean Bias, which is the absolute difference between the spatial mean of predictions and ground-truths.

Table 4: Downscaling performance measured by RMSE and Absolute Mean Bias on six variables.

		Z500	T850	T2m	Q700	U10m	V10m
RMSE (↓)	ClimaX frozen	105.49	0.93	1.16	0.70	1.02	1.01
	ClimaX finetuned	74.62	0.78	0.94	0.61	0.83	0.83
	Stormer frozen	104.26	0.95	1.12	0.76	1.07	1.05
	Stormer finetuned	**38.84**	**0.57**	**0.62**	**0.55**	**0.64**	**0.64**
	Unet	47.65	0.66	0.73	0.56	0.70	0.70
Absolute Mean Bias (↓)	ClimaX frozen	28.660	0.167	0.054	**0.001**	0.032	0.009
	ClimaX finetuned	13.830	0.153	0.119	0.002	**0.007**	**0.001**
	Stormer frozen	17.540	**0.046**	0.048	**0.001**	0.019	0.011
	Stormer finetuned	**0.090**	0.051	**0.031**	**0.001**	0.011	0.017
	Unet	8.790	0.140	0.040	0.005	0.011	0.006

Table 4 shows the performance of the considered methods. Unlike the forecasting tasks, there is a significant gap between the frozen and the fully finetuned models of ClimaX and Stormer. This

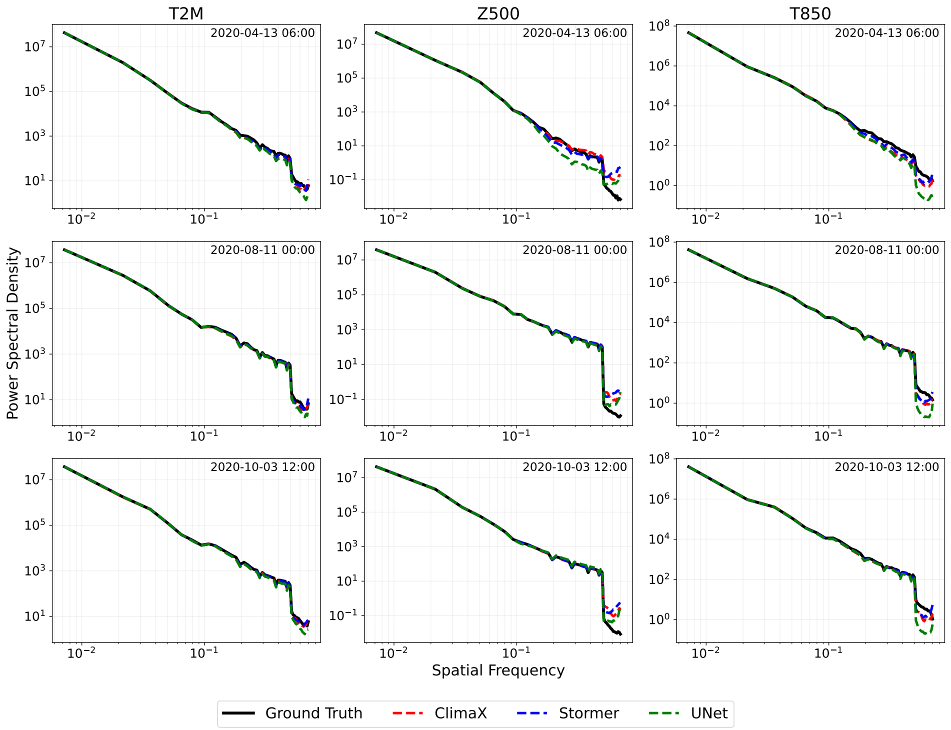

Figure 2: PSD plots of the baselines in comparison with the ground truths across three variables and three random test data points.

indicates that the transformer backbone pretrained for temporal forecasting might be sub-optimal for spatial downscaling and further finetuning is required to achieve good performance. Stormer is the best model in this task with the lowest RMSE and Absolute Mean Bias for most variables, followed by the Unet baseline. Since ClimaX has the lowest parameter count, we hypothesize that larger models tend to perform better in this task. This observation was also suggested by the scaling analysis in the original ClimaX paper.

In addition to quantitative metrics, we also plot the Power Spectral Density (PSD) to examine how well each model preserves the power spectrum across different spatial scales of the ground truth. To create these plots, we computed the 2D Power Spectral Density using the Fast Fourier Transform (FFT) for each spatial field, then performed radial averaging to obtain 1D PSD curves that show how power varies with spatial frequency. For each variable we consider (T2M, Z500, T850), we plotted the PSD curves of the ground truth and predictions from the three models on a log-log scale.

Figure 2 shows the PSD plots for three randomly selected data points in the test set. All three models display excellent agreement with the ground truth across low to medium spatial frequencies for all variables, indicating they accurately capture large-scale spatial patterns. However, there are notable differences at high spatial frequencies (> 0.2): UNet tends to underestimate the power at these frequencies, suggesting it may smooth out fine-scale details, while ClimaX and Stormer better preserve these high-frequency components. The results suggest that two pretrained models, ClimaX and Stormer, have an advantage in preserving fine-scale spatial details compared to UNet.

4.4 Data infilling

We test the ability of foundation models to fill in missing temperature data, which is a common issue due to gaps in the coverage of observation stations. We construct training and validation data for this task from ERA5. During training, we generate a random mask for each training data point, with the mask ratio (missing ratio) drawn from a uniform distribution $r \sim \mathcal{U}[0.1, 0.9]$. We test each model to perform infilling with a set of mask ratios $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, where a fixed set of masks for each ratio is pre-generated and saved to disk to maintain evaluation consistency across models.

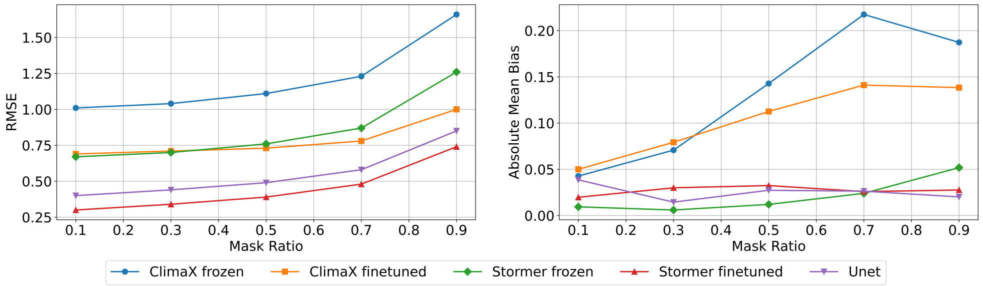

Figure 3: Infilling performance for surface temperature measured by RMSE and Absolute Mean Bias with different missing ratios.

Figure 3 shows the performance of the considered models for different mask ratios. Similar to downscaling, fully finetuned models work much better than frozen counterparts, and Stormer is the best method for this task. This result again highlights the difference between temporal and spatial tasks and the need for full finetuning to achieve good performance.

4.5 Climate model emulation

We aim to predict the annual mean global distributions of four target variables: surface air temperature, diurnal temperature range (difference between daily maximum and minimum surface air temperature), precipitation, and the 90th percentile precipitation. The input variables are four forcing factors: carbon dioxide (CO_2), sulfur dioxide (SO_2), black carbon (BC), and methane (CH_4). Following ClimateBench, we report NRMSE_s, NRMSE_g, and $\text{NRMSE}_t = \text{NRMSE}_s + 5\times \text{NRMSE}_g$ as the evaluation metrics. We use the best method in ClimateBench, namely ClimateBench-NN, as the baseline in addition to ClimaX and Stormer. We note that in this task, both the input and target variables were unseen during the pretraining of ClimaX and Stormer, so we replaced their embedding layer and prediction head with randomly initialized networks. Therefore, the transformer backbone essentially serves as a feature extractor. We finetune a separate model for each target variable.

Table 5: Climate model emulation performance measured by NRMSE_s, NRMSE_g, and NRMSE_t.

	Surface air temperature			Diurnal temperature range			Precipitation			90th percentile precipitation		
	NRMSE_s	NRMSE_g	NRMSE_t	NRMSE_s	NRMSE_g	NRMSE_t	NRMSE_s	NRMSE_g	NRMSE_t	NRMSE_s	NRMSE_g	NRMSE_t
ClimaX frozen	**0.085**	**0.043**	**0.297**	**6.688**	**0.810**	**10.739**	**2.193**	0.183	3.110	**2.681**	0.342	**4.389**
ClimaX finetuned	0.086	**0.043**	0.300	7.148	0.961	11.952	2.360	0.206	3.390	2.739	0.332	4.397
Stormer frozen	0.117	**0.043**	0.334	9.123	0.980	14.022	6.159	0.210	7.211	6.773	0.296	8.254
Stormer finetuned	0.126	0.047	0.361	8.598	0.834	12.767	6.180	0.391	8.136	6.797	0.316	8.376
ClimateBench-NN	0.123	0.080	0.524	7.465	1.233	13.632	2.349	**0.151**	**3.104**	3.108	**0.282**	4.517

Table 5 shows the superior performance of ClimaX in this task, outperforming Stormer and the ClimateBench-NN baseline by a large margin. This result highlights a unique benefit of multi-source pretraining in acquiring a general-purpose backbone that allows for easy transferability to downstream tasks and datasets significantly different from pretraining. Moreover, frozen models generally work better than the fully finetuned counterparts for this task. This can be explained by the small data size of ClimateBench (754 data points), so further finetuning of the backbone can lead to overfitting and hurt the test performance. A similar result was observed in the ClimaX paper.

4.6 Extreme weather detection

Finally, we consider the task of detecting Tropical Cyclones (TCs) and Atmospheric Rivers (ARs), two atmospheric phenomena highly correlated with extreme weather events. We use the ClimateNet dataset for finetuning and evaluation, in which we use data from 1996 to 2010 for training and validation, and 2011 to 2013 for testing. We finetune ClimaX and Stormer to classify each pixel into one of three classes: TC, AR, and Background (BG). Similar to climate model emulation, we replace the pretrained embedding and prediction layer with randomly initialized networks. Since ClimateNet data is of much higher resolution, we increase the patch size to 8 for both ClimaX and Stormer, and interpolate the pretrained positional embedding to match the new sequence length.

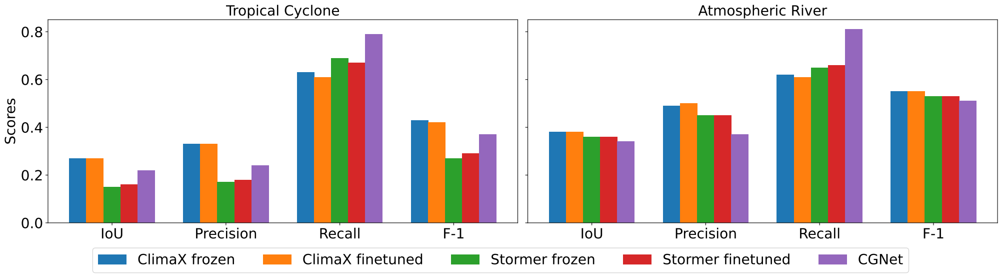

Figure 4: Extreme weather detection performance measured by IoU, Precision, Recall, and F-1.

Figure 4 compares the performance of ClimaX and Stormer with CGNet (Wu et al., 2020), a lightweight segmentation architecture based on CNN specifically designed for this task. Since the BG class dominates other classes, we adopt the weighted Jaccard loss function (Lacombe et al., 2023) to counter this class imbalance. The two finetuned versions of ClimaX work best in this task with respect to IoU and F-1, significantly outperforming its counterpart Stormer. This again demonstrates the importance of multi-source pretraining in obtaining higher transferable backbones. ClimaX also outperforms CGNet in $3/4$ metrics, showing the benefit of foundation models over specialized architectures.

5 Conclusion

We presented AtmosArena, the first benchmark dedicated to foundation models in atmospheric sciences. AtmosArena offers a diverse suite of tasks, datasets, and evaluation metrics to evaluate a foundation model holistically. AtmosArena not only provides a standard benchmark for comparing model performance but also serves as a crucial tool for identifying future research works. In addition, we release all our data, code, and model checkpoints, facilitating reproducible research and broadening collaborations. Given the vast development of scientific foundation models, we believe our contribution is timely and useful for both machine learning and atmospheric communities.

Limitations and Future Work With academic resource constraints, we acknowledge that there are various directions to improve AtmosArena in each of four dimensions – datasets, tasks, models, and evaluations. One such direction involves integrating regional datasets and expanding the collection of supported data sources. On the task side, we plan to include probabilistic tasks that are an important aspect of modeling weather and climate. For models and evaluations, we plan to find platforms for hosting atmospheric foundation models, along with an accompanying leaderboard.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers. *arXiv preprint arXiv:2402.12365*, 2024.

S Altikat. Prediction of co2 emission from greenhouse to atmosphere with artificial neural networks and deep learning neural networks. *International Journal of Environmental Science and Technology*, 18(10):3169–3178, 2021.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alex Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Yasin Akın Ayturan, Zeynep Cansu Ayturan, and Hüseyin Oktay Altun. Air pollution modelling with deep learning: a review. *International Journal of Environmental Pollution and Environmental Modelling*, 1(3):58–62, 2018.

J. Baño Medina, R. Manzanas, and J. M. Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020. doi: 10.5194/gmd-13-2109-2020. URL `https://gmd.copernicus.org/articles/13/2109/2020/`.

Melahat Sevgül Bakay and Ümit Ağbulut. Electricity production based forecasting of greenhouse gas emissions in turkey with deep learning, support vector machine and artificial neural network algorithms. *Journal of Cleaner Production*, 285:125324, 2021.

V. Balaji, E. Maisonnave, N. Zadeh, B. N. Lawrence, J. Biercamp, U. Fladrich, G. Aloisio, R. Benson, A. Caubel, J. Durachta, M.-A. Foujols, G. Lister, S. Mocavero, S. Underwood, and G. Wright. CPMIP: measurements of real computational performance of Earth system models in CMIP6. *Geoscientific Model Development*, 10(1):19–34, 2017. doi: 10.5194/gmd-10-19-2017. URL `https://gmd.copernicus.org/articles/10/19/2017/`.

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

Abdellatif Bekkar, Badr Hssina, Samira Douzi, and Khadija Douzi. Air-pollution prediction in smart city, deep learning approach. *Journal of big Data*, 8:1–21, 2021.

C. Betancourt, T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler. Aq-bench: a benchmark dataset for machine learning on global air quality metrics. *Earth System Science Data*, 13(6):3013–3033, 2021. doi: 10.5194/essd-13-3013-2021. URL `https://essd.copernicus.org/articles/13/3013/2021/`.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tien-Cuong Bui, Van-Duc Le, and Sang-Kyun Cha. A deep learning approach for forecasting air pollution in south korea using lstm. *arXiv preprint arXiv:1804.07891*, 2018.

Salva Rühling Cachay, Venkatesh Ramesh, Jason NS Cole, Howard Barker, and David Rolnick. Climart: A benchmark dataset for emulating atmospheric radiative transfer in weather and climate models. *arXiv preprint arXiv:2111.14671*, 2021.

CarbonPlan. CMIP6-Downscaling. `https://github.com/carbonplan/cmip6-downscaling`, 2022.

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*, 2023a.

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, 2023b. doi: 10.1038/s41612-023-00512-1. URL `https://doi.org/10.1038/s41612-023-00512-1`.

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. *arXiv preprint arXiv:2306.12873*, 2023c.

Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Freddie Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards global precipitation forecasting from satellite imagery, 2020.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016a. doi: 10.5194/gmd-9-1937-2016. URL `https://gmd.copernicus.org/articles/9/1937/2016/`.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016b.

Gabriele Franch, Valerio Maggio, Luca Coviello, Marta Pendesini, Giuseppe Jurman, and Cesare Furlanello. Taasrad19, a high-resolution weather radar reflectivity dataset for precipitation nowcasting. *Scientific Data*, 7, 07 2020. doi: 10.1038/s41597-020-0574-8.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL `https://zenodo.org/records/10256836`.

Andrew Geiss, Sam J Silva, and Joseph C Hardin. Downscaling atmospheric chemistry simulations with physically consistent deep learning. *Geoscientific Model Development*, 15(17):6677–6694, 2022.

Isaac Godfried, Kriti Mahajan, Maggie Wang, Kevin Li, and Pranjalya Tiwari. Flowdb a large scale precipitation, river, and flash flood dataset, 2020.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

Joseph Hamman and Julia Kent. Scikit-downscale: an open source python package for scalable climate downscaling. In *2020 EarthCube Annual Meeting*, 2020.

Abderrachid Hamrani, Abdolhamid Akbarzadeh, and Chandra A Madramootoo. Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of The Total Environment*, 741:140338, 2020.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes. *arXiv preprint arXiv:2405.19101*, 2024.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 0035-9009. doi: https://doi.org/10.1002/qj.3803.

Azim Heydari, Meysam Majidi Nezhad, Davide Astiaso Garcia, Farshid Keynia, and Livio De Santoli. Air pollution forecasting application based on deep learning model and optimization algorithm. *Clean Technologies and Environmental Policy*, pp. 1–15, 2022.

Stephan Hoyer and Joe Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1):10, April 2017. doi: 10.5334/jors.148.

Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36: 21757–21792, 2023.

Ryan Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022.

Christoph A Keller, K Emma Knowland, Bryan N Duncan, Junhua Liu, Daniel C Anderson, Sampa Das, Robert A Lucchesi, Elizabeth W Lundgren, Julie M Nicely, Eric Nielsen, et al. Description of the nasa geos composition forecast modeling system geos-cf v1. 0. *Journal of Advances in Modeling Earth Systems*, 13(4):e2020MS002413, 2021.

K Emma Knowland, Christoph A Keller, Pamela A Wales, Krzysztof Wargan, Lawrence Coy, Matthew S Johnson, Junhua Liu, Robert A Lucchesi, Sebastian David Eastham, E Fleming, et al. Nasa geos composition forecast modeling system geos-cf v1. 0: Stratospheric composition. *Journal of advances in modeling earth systems*, 14(6):e2021MS002852, 2022.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, James Lottes, Stephan Rasp, Peter Düben, Milan Klöwer, et al. Neural general circulation models. *arXiv preprint arXiv:2311.07222*, 2023.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate, 2024.

Romain Lacombe, Hannah Grossman, Lucas Hendren, and David Lüdeke. Improving extreme weather events detection with light-weight neural networks. *arXiv preprint arXiv:2304.00176*, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 0(0):eadi2336, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

David A. Lavers, Adrian Simmons, Freja Vamborg, and Mark J. Rodwell. An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165, 2022. doi: https://doi.org/10.1002/qj.4351. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4351.

L. Ruby Leung, Linda O. Mearns, Filippo Giorgi, and Robert L. Wilby. REGIONAL CLIMATE RESEARCH: Needs and Opportunities. *Bulletin of the American Meteorological Society*, 84 (1):89–95, 2003. ISSN 00030007, 15200477. URL http://www.jstor.org/stable/26215433.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Yumin Liu, Auroop R. Ganguly, and Jennifer Dy. Climate Downscaling Using YNet: A Deep Convolutional Network with Skip Connections and Fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, pp. 3145–3153, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403366. URL https://doi.org/10.1145/3394486.3403366.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Peter Lynch. The origins of computer weather prediction and climate modeling. *Journal of computational physics*, 227(7):3431–3444, 2008.

Manil Maskey, Rahul Ramachandran, Muthukumaran Ramasubramanian, Iksha Gurung, Brian Freitag, Aaron Kaulfus, Drew Bollinger, Daniel Cecil, and J. Miller. Deepti: Deep-learning-based tropical cyclone intensity estimation system. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, 07 2020. doi: 10.1109/JSTARS.2020.3011907.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

Christoph Minixhofer, Mark Swan, Calum McMeekin, and Pavlos Andreadis. Droughted: A dataset and methodology for drought forecasting spanning multiple climate zones. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Judah Cohen, Miruna Oprescu, Ernest Fraenkel, and Lester Mackey. Adaptive bias correction for improved subseasonal forecasting. *Nature Communications*, 14(1), June 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38874-y. URL http://dx.doi.org/10.1038/s41467-023-38874-y.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester Mackey. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking, 2024.

Takeyoshi Nagasato, Kei Ishida, Ali Ercan, Tongbi Tu, Masato Kiyama, Motoki Amagasaki, and Kazuki Yokoo. Extension of convolutional neural network along temporal and vertical directions for precipitation downscaling. *arXiv preprint arXiv:2112.06571*, 2021.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *arXiv preprint arXiv:2402.00712*, 2024.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, and et al. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, Mar 2024. doi: 10.1038/s41586-024-07145-1.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023a.

Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. *arXiv preprint arXiv:2307.01909*, 2023b.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Sandeep Madireddy, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *arXiv preprint arXiv:2312.03876*, 2023c.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035. Curran Associates, Inc., 2019.

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*, 2022a.

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*, 2022b.

Andrew D Polasky, Jenni L Evans, and Jose D Fuentes. Ccdownscaling: A python package for multivariable statistical climate model downscaling. *Environmental Modelling & Software*, 165: 105712, 2023.

Prabhat, Karthik Kashinath, Mayur Mudigonda, Sol Kim, Lukas Kapp-Schwoerer, Andre Graubner, Ege Karaismailoglu, Leo von Kleist, Thorsten Kurth, Annette Greiner, et al. Climatenet: An expert-labelled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development Discussions*, 2020:1–28, 2020.

Prabhat, K. Kashinath, M. Mudigonda, S. Kim, L. Kapp-Schwoerer, A. Graubner, E. Karaismailoglu, L. von Kleist, T. Kurth, A. Greiner, A. Mahesh, K. Yang, C. Lewis, J. Chen, A. Lou, S. Chandran, B. Toms, W. Chapman, K. Dagon, C. A. Shields, T. O'Brien, M. Wehner, and W. Collins. Climatenet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, 14(1):107–124, 2021. doi: 10.5194/gmd-14-107-2021. URL `https://gmd.copernicus.org/articles/14/107/2021/`.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather, 2024.

Adiba Mahbub Proma, Md Saiful Islam, Stela Ciko, Raiyan Abdul Baten, and Ehsan Hoque. Nadbenchmarks–a compilation of benchmark datasets for machine learning tasks related to natural disasters. *arXiv preprint arXiv:2212.10735*, 2022.

Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr. Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,

2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/519c84155964659375821f7ca576f095-Paper.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

Stephan Rasp and Nils Thuerey. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405, 2021.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020a.

Stephan Rasp, Hauke Schulz, Sandrine Bony, and Bjorn Stevens. Combining crowd-sourcing and deep learning to explore the meso-scale organization of shallow convection, 2020b.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2023.

Sara A. Rauscher, Erika Coppola, Claudio Piani, and Filippo Giorgi. Resolution effects on regional climate model simulations of seasonal precipitation over Europe. *Climate Dynamics*, 35(4):685–711, Sep 2010. ISSN 1432-0894. doi: 10.1007/s00382-009-0607-7. URL https://doi.org/10.1007/s00382-009-0607-7.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, Sep 2021a. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL https://doi.org/10.1038/s41586-021-03854-z.

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, 2021b.

Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1132–1142, 2021.

Eduardo Rocha Rodrigues, Igor Oliveira, Renato Cunha, and Marco Netto. Deepdownscale: A deep learning strategy for high-resolution weather forecast. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 415–422. IEEE, 2018.

Robert A Rohde and Zeke Hausfather. The berkeley earth land/ocean temperature record. *Earth System Science Data*, 12(4):3469–3479, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

DA Sachindra, Khandakar Ahmed, Md Mamunur Rashid, S Shahid, and BJC Perera. Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, 212: 240–258, 2018.

Øyvind Seland, Mats Bentsen, Dirk Jan Leo Olivié, Thomas Toniazzo, Ada Gjermundsen, Lise Seland Graff, Jens Boldingh Debernard, Alok Kumar Gupta, Yan-Chun He, Alf Kirkevåg, et al. Overview of the norwegian earth system model (noresm2) and key climate response of cmip6 deck, historical, and scenario simulations. 2020.

Muhammed Sit, Bong-Chul Seo, and Ibrahim Demir. Iowarain: A statewide rain event dataset based on weather radars and quantitative precipitation estimation. *arXiv preprint arXiv:2107.03432*, 2021.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equation: Multi-operator learning and extrapolation. *arXiv preprint arXiv:2404.12355*, 2024.

Qing Tao, Fang Liu, Yong Li, and Denis Sidorov. Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. *IEEE access*, 7:76690–76698, 2019.

Thomas Vandal, Evan Kodra, and Auroop R Ganguly. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137:557–570, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, David N. Spergel, Rachel S. Somerville, Jose Manuel Zorrilla Matilla, Faizan G. Mohammad, Sultan Hassan, Helen Shao, Digvijay Wadekar, Michael Eickenberg, Kaze W. K. Wong, Gabriella Contardo, Yongseok Jo, Emily Moser, Erwin T. Lau, Luis Fernando Machado Poletti Valle, Lucia A. Perez, Daisuke Nagai, Nicholas Battaglia, and Mark Vogelsberger. The camels multifield data set: Learning the universe's fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, April 2022. ISSN 1538-4365. doi: 10.3847/1538-4365/ac5ab0. URL http://dx.doi.org/10.3847/1538-4365/ac5ab0.

F. Vitart, A. W. Robertson, A. Spring, F. Pinault, R. Roškar, W. Cao, S. Bech, A. Bienkowski, N. Caltabiano, E. De Coning, B. Denis, A. Dirkson, J. Dramsch, P. Dueben, J. Gierschendorf, H. S. Kim, K. Nowak, D. Landry, L. Lledó, L. Palma, S. Rasp, and S. Zhou. Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. *Bulletin of the American Meteorological Society*, 103(12):E2878–E2886, December 2022. doi: 10.1175/bams-d-22-0046.1.

Frédéric Vitart and Andrew W Robertson. The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate and Atmospheric Science*, 1(1):1–7, 2018.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

D. Watson-Parris, Y. Rao, D. Olivié, Ø. Seland, P. Nowack, G. Camps-Valls, P. Stier, S. Bouabid, M. Dewey, E. Fons, J. Gonzalez, P. Harder, K. Jeggle, J. Lenhardt, P. Manshausen, M. Novitasari, L. Ricard, and C. Roesch. ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022a. doi: https://doi.org/10.1029/2021MS002954. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002954. e2021MS002954 2021MS002954.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10): e2021MS002954, 2022b.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. Ace: A fast, skillful learned global atmospheric model for climate prediction, 2023.

NP Wedi, P Bauer, W Denoninck, M Diamantakis, M Hamrud, C Kuhnlein, S Malardel, K Mogensen, G Mozdzynski, and PK Smolarkiewicz. *The modelling infrastructure of the Integrated Forecasting System: Recent advances and future challenges*. European Centre for Medium-Range Weather Forecasts, 2015.

Jonathan A Weyn, Dale R Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020.

Ross Wightman. PyTorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

T Wu, S Tang, R Zhang, J Cao, and Y Zhang Cgnet. A light-weight context guided network for semantic segmentation., 2020, 30. *DOI: https://doi. org/10.1109/TIP*, pp. 1169–1179, 2020.

Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus Christopher Will, Gunnar Behrens, Julius Busecke, Nora Loose, Charles I Stern, Tom Beucler, Bryce Harrop, Benjamin R Hillman, Andrea Jenney, Savannah Ferretti, Nana Liu, Anima Anandkumar, Noah D Brenowitz, Veronika Eyring, Nicholas Geneva, Pierre Gentine, Stephan Mandt, Jaideep Pathak, Akshay Subramaniam, Carl Vondrick, Rose Yu, Laure Zanna, Tian Zheng, Ryan Abernathey, Fiaz Ahmed, David C Bader, Pierre Baldi, Elizabeth Barnes, Christopher Bretherton, Peter Caldwell, Wayne Chuang, Yilun Han, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Karthik Kashinath, Marat Khairoutdinov, Thorsten Kurth, Nicholas Lutsko, Po-Lun Ma, Griffin Mooers, J. David Neelin, David Randall, Sara Shamekh, Mark A Taylor, Nathan Urban, Janni Yuval, Guang Zhang, and Michael Pritchard. Climsim: A large multi-scale dataset for hybrid physics-ml climate emulation, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

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, Jul 2023. doi: 10.1038/s41586-023-06184-4.

A Appendix

B Licenses and Terms of Use

The source code will be available online under the MIT License upon acceptance. The licenses of the datasets we use in AtmosArena are as follows:

- ERA5 is curated and provided by WeatherBench2 which is licensed under Apache License 2.0 (https://github.com/google-research/weatherbench2/blob/main/LICENSE).
- Berkeley Earth (https://berkeleyearth.org/data/), ClimateBench (https://zenodo.org/record/7064308), ClimateNet (https://gmd.copernicus.org/articles/14/107/2021/) are available under the CC BY 4.0 license.
- CAMS Analysis provided by Copernicus Atmosphere Monitoring Service (CAMS) is free of charge, worldwide, non-exclusive, royalty-free and perpetual (https://atmosphere.copernicus.eu/sites/default/files/repository/CAMS_data_license.pdf).
- GEOS-CF (https://portal.nccs.nasa.gov/datashare/gmao/geos-cf/) provided by NASA is free for public access.

C Datasets

C.1 Dataset details

Table 6: Summary of the datasets used to finetune and evaluate baselines in AtmosArena.

Name	Resolution	Temporal coverage	Surface Variables	Multi-level Variables	Num levels	Size (GB)	Num frames
ERA5	128x256	1979-2020	T2m, U10, V10, MSLP	Z, T, U, V, Q	13	1600	61,324
Berkeley Earth	128x256	1850-2023	T2m	N/A	N/A	0.26	2,088
ClimateBench	32x64	2015-2100	CO2, SO2, CH4, BC, TAS, DTR, PR, PR90	N/A	N/A	0.12	839
ClimateNet	768x1152	1996-2013	TMQ, UBOT, VBOT, PS, PSL, PRECT, TS, TREFHT, ZBOT	U850, V850, QREFHT, T200, T500, Z1000, Z200	N/A	28	459
CAMS Analysis	128x256	2017-2022	T2m, U10, V10, MSLP, TC CO, TC NO, TC NO2, TC SO2, TC O3, PM1, PM2.5, PM10	U, V, T, Q, Z, CO, NO, NO2, SO2, O3	13	59	3774
GEOS-CF	128x256	2018-2023	NO2, SO2, CO, O3, PM2.5	N/A	N/A	363	52,584

Table 6 details the datasets in AtmosArena, including their spatial resolution, temporal coverage, variables, and size. The full names of the abbreviated variables are:

- T2m, U10, V10, MSLP: 2-meter temperature, 10-meter zonal wind, 10-meter meridional wind, Mean sea level pressure.
- Z, T, U, V, Q: Geopotential, Temperature, Zonal wind, Meridional wind, Specific humidity at different pressure levels.
- CO2, SO2, CH4, BC: Carbon dioxide, Sulfur Dioxide, Methane, Black carbon.
- TAS, DTR, PR, PR90: Surface air temperature, Diurnal temperature range, Precipitation, 90th percentile precipitation.
- TMQ, UBOT, VBOT, PS, PSL, PRECT, TS, TREFHT, ZBOT: Total Precipitable Water, Lowest Model Level Zonal Wind, Lowest Model Level Meridional Wind, Surface Pressure, Sea Level Pressure, Total Precipitation Rate, Surface Temperature, Reference Height Temperature, Lowest Model Level Height.
- U850, V850, QREFHT, T200, T500, Z1000, Z200: Zonal Wind at 850 mb, Meridional Wind at 850 mb, Specific Humidity at Reference Height, Temperature at 200 mb, Temperature at 500 mb, Geopotential Height at 1000 mb, Geopotential Height at 200 MB.

- TC CO, TC NO, TC NO2, TC SO2, TC O3, PM1, PM2.5, PM10: Total Column Carbon Monoxide, Total Column Nitric Oxide, Total Column Nitrogen Dioxide, Total Column Sulfur Dioxide, Total Column Ozone, Particulate Matter 1um, Particulate Matter 2.5um, Particulate Matter 10um.
- CO, NO, NO2, SO2, O3: Zonal Wind, Meridional Wind, Temperature, Specific Humidity, Geopotential Height, Carbon Monoxide, Nitric Oxide, Nitrogen Dioxide, Sulfur Dioxide, Ozone.

For ERA5, following WeatherBench2 (Rasp et al., 2023), we used the 6-hourly subsampled data from the original ERA5 at 00:00, 06:00, 12:00, and 18:00, and used the 13 pressure levels for the multi-level variables: 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000. We use the same pressure levels for CAM Analysis.We also note that the resolutions of ERA5, Berkeley Earth, ClimateBench, CAMS Analysis, and GEOS-RF used in our paper are different from their original resolutions. We used bilinear interpolation to regrid the original data to the resolutions in Table 6.

C.2 Train, validation, and test split

Table 7: Summary of train, validation, and test split of the datasets in AtmosArena.

Name	Train time frame	Validation time frame	Test Year(s)
time frame	1979-2018	2019	2020
Berkeley Earth	N/A	N/A	2000-2024
ClimateBench	2015-2100	2015-2100	2015-2100
ClimateNet	1996-2007	2008-2010	2011-2013
CAMS Analysis	2018-2020	2021	2022
GEOS-CF	2017-2020	2021	2022

Tabel 7 summarizes the train, validation, and test split of the datasets we included in AtmosArena. Most datasets are split according to time, where training, validation, and test data belong to non-overlapping time periods. For ClimateBench, which we used for the climate model emulation task, however, the data is split according to different future emission scenarios. We refer to ClimateBench (Watson-Parris et al., 2022b) for a detailed discussion of these scenarios.

D Evaluation metrics

This section presents the formulation of the evaluation metrics we included in AtmosArena. We use the following notations across the metrics:

- N is the number of data points
- H is the number of latitude coordinates.
- W is the number of longitude coordinates.
- X and $\tilde{X}$ are the ground-truth and prediction, respectively.

Each equation below is computed for one single variable. To account for the non-uniformity of the grid cell areas when gridding a round Earth, most metrics are latitude-weighted to give more weight to the cells closer to the equator. The latitude weighting function is given by

$$L(i) = \frac{\cos(H_i)}{\frac{1}{H}\sum_{i=1}^{H}\cos(H_i)} \tag{1}$$

D.1 Forecasting metrics

Root Mean Square Error (RMSE)

$$\text{RMSE} = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}L(i)(\tilde{X}_{k,i,j} - X_{k,i,j})^2}. \tag{2}$$

Anomaly Correlation Coefficient (ACC) is the spatial correlation between prediction anomalies $\tilde{X}^{'}$ relative to climatology and ground truth anomalies $X^{'}$ relative to climatology:

$$\text{ACC} = \frac{\sum_{k,i,j} L(i)\tilde{X}^{'}_{k,i,j}X^{'}_{k,i,j}}{\sqrt{\sum_{k,i,j} L(i)\tilde{X}^{'2}_{k,i,j} \sum_{k,i,j} L(i)X^{'2}_{k,i,j}}}, \tag{3}$$

$$\tilde{X}^{'} = \tilde{X} - C, X^{'} = X - C, \tag{4}$$

in which climatology C is the temporal mean of the ground truth data over a fixed period. We used the climatology data from WeatherBench2 (Rasp et al., 2023) in our all experiments.

Spectral Divergence (SpecDiv) is inspired by KL divergence, which computes the expectation of the logarithmic ratio between the ground truth and predicted spectra. This metric emphasizes the relative error between the frequency components of the ground truth and prediction:

$$\text{SpecDiv} = \sum_k S'(k) \cdot \log\left(\frac{S'(k)}{\tilde{S}'(k)}\right) \tag{5}$$

where $S'(k)$ and $\hat{S}'(k)$ represent the spectral components of the ground truth and predictions, respectively, and k denotes the spectral component.

D.2 Climate downscaling and infilling metrics

Root Mean Square Error (RMSE) This is the same as Equation (2).

Mean Bias measures the mean difference between the prediction and the ground truth. A positive mean bias shows overestimation, while a negative mean bias shows underestimation:

$$\text{Mean bias} = \frac{1}{N \times H \times W} \sum_{k=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} (\tilde{X}_{k,i,j} - X_{k,i,j}) \tag{6}$$

Anomaly Pearson Coefficient measures the Pearson correlation between the prediction and the ground truth anomalies. We first flatten the prediction and ground truth anomalies, and compute the metric as follows:

$$\rho_{\tilde{X}',X'} = \frac{\text{cov}(\tilde{X}', X')}{\sigma_{\tilde{X}'}\sigma_{X'}} \tag{7}$$

NOTE: For the Climate data infilling task, we compute the metrics over the masked cells only.

D.3 Climate model emulation metrics

Normalized spatial root mean square error (NRMSE$_s$) measures the spatial discrepancy between the temporal mean of the prediction and the temporal mean of the ground truth:

$$\text{NRMSE}_s = \sqrt{\left\langle \left(\frac{1}{N}\sum_{k=1}^{N}\tilde{X} - \frac{1}{N}\sum_{k=1}^{N} X \right)^2 \right\rangle} \Bigg/ \frac{1}{N}\sum_{k=1}^{N} \langle X \rangle, \tag{8}$$

in which $\langle A \rangle$ is the global mean of A:

$$\langle A \rangle = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} L(i) A_{i,j} \tag{9}$$

Normalized global root mean square error (NRMSE$_g$) measures the discrepancy between the global mean of the prediction and the global mean of the ground truth:

$$\text{NRMSE}_g = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\langle \tilde{X} \rangle - \langle X \rangle \right)^2} \Bigg/ \frac{1}{N}\sum_{k=1}^{N} \langle X \rangle. \tag{10}$$

Total normalized root mean square error (Total) is the weighted sum of NRMSE$_s$ and NRMSE$_g$:

$$\text{Total} = \text{NRMSE}_s + \alpha \cdot \text{NRMSE}_g, \tag{11}$$

where α is chosen to be 5 as suggested by Watson-Parris et al. (2022a).

D.4 Extreme weather events detection metrics

Each pixel in the $H \times W$ grid is classified into one of three classes, leading to a confusion matrix per class (AR, TC, and BG). The performance metrics, calculated for each class, are defined as follows using the elements of the confusion matrix—True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN):

Intersection over Union (IoU)

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \tag{12}$$

Precision

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{13}$$

Recall

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{14}$$

F-1 Score

$$\text{F-1}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{15}$$

Specificity

$$\text{Specificity}_c = \frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c} \tag{16}$$

E Experiment details

This section details the experiments we conducted in Section 3, including model architectures and hyperparameters, training objectives, and optimization.

E.1 Model architectures

Unet We borrow our Unet implementation from PDEArena (Gupta & Brandstetter, 2022). Table 8 shows hyperparameters of Unet we use in all our experiments. The Unet model has a total of 500M parameters.

Table 8: Default hyperparameters of UNet

Hyperparameter	Meaning	Value
Padding size	Padding size of each convolution layer	1
Kernel size	Kernel size of each convolution layer	3
Stride	Stride of each convolution layer	1
Channel multiplications	Determine the number of output channels for Down and Up blocks	$[1, 2, 2, 4]$
Blocks	Number of Resnet blocks	2
Use attention	If use attention in Down and Up blocks	False

ClimaX and Stormer For ClimaX and Stormer, we borrow the implementation from their original papers (Nguyen et al., 2023a;c), which we refer to for a detailed description of their architectures. Table 9 shows hyperparameters of ClimaX and Stormer we use in all our experiments. The parameter count for ClimaX and Stormer is 100M and 400M, respectively.

Table 9: Default hyperparameters of ClimaX and Stormer

Hyperparameter	Meaning	ClimaX	Stormer
p	Patch size	4	2
D	Embedding dimension	1024	1024
Depth	Number of ViT blocks	8	24
# heads	Number of attention heads	16	16
MLP ratio	Determine the hidden dimension of the MLP layer in a ViT block	4	4
Prediction depth	Number of layers of the prediction head	2	1
Hidden dimension	Hidden dimension of the prediction head	1024	N/A

E.1.1 Extensions for climate model emulation

We modify the architecture of ClimaX and Stormer for this task to account for the time dimension T in the input. Each time slice of the input goes through the embedding layer and the transformer blocks independently, resulting in an output tensor of shape $T \times h \times w \times D$ where D is the embedding dimension. This tensor then goes through a global pooling layer along the spatial dimensions h and w, outputting a tensor of shape $T \times D$. This sequence of tensors is aggregated by a cross-attention layer over the time dimension to a single vector of D dimensions. Finally, a linear layer predicts the output of shape $V \times H \times W$. The cross-attention layer along the time dimension is randomly initialized and trained together with the new embedding and prediction layer, as well as the transformer backbone.

E.1.2 Extensions for extreme weather events detection

Since the spatial resolution of ClimateNet is 768×1152, training the original ClimaX and Stormer with patch sizes of 4 and 2, respectively, is too computationally expensive. To address this issue, we use a stack of 6 convolutional layers to embed the input before the attention blocks which outputs a tensor of shape $96 \times 144 \times D$, reducing the spatial resolution by 8. This tensor goes through the transformer blocks and a linear prediction head which outputs a tensor of shape $3 \times 96 \times 144$ where 3 is the number of classes. Finally, this output is bilinearly interpolated to the original spatial resolution of 768×1152. The bilinear interpolation module is also used by the baseline CGNet (Wu et al., 2020).

E.2 Training details

E.2.1 Data normalization

For tasks that utilize ERA5 for training and evaluation, including medium-range weather forecasting, S2S forecasting, climate downscaling, and climate data infilling, we normalize both input and output variables to have mean 0 and standard deviation 1. The normalization constants are computed across the entire training set. During evaluation, predictions and ground-truths are de-normalized to the original scale before computing the metrics.

For the extreme weather events detection task that uses ClimateNet, we normalize the input variables similarly to ERA5, but not the output variables since they are discrete labels.

For the climate model emulation task that uses ClimateBench, we normalize the input variables similarly to ERA5, but not the output variables since we predict each target variable separately.

E.2.2 Training objectives

Regression For the five regression tasks, including medium-range weather forecasting, S2S forecasting, climate downscaling, climate data infilling, and climate model emulation, we use the same latitude-weighted mean-squared error loss for training:

$$\mathcal{L}(\theta) = \frac{1}{V' \times H \times W} \sum_{v=1}^{V'} \sum_{i=1}^{H} \sum_{j=1}^{W} L(i)(\tilde{X}^{v,i,j} - X^{v,i,j})^2. \tag{17}$$

Classification For the extreme weather events detection task, we utilize the weighted Jaccard loss proposed in Lacombe et al. (2023) to prioritize the TC and AR classes:

$$\mathcal{L}(\theta) = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(1 - w_c \frac{\tilde{X}^{c,i,j} X^{c,i,j}}{(\tilde{X}^{c,i,j} + X^{c,i,j}) - \tilde{X}^{c,i,j} X^{c,i,j}} \right), \tag{18}$$

in which w_c is the weight of class c. Following Lacombe et al. (2023), we set w_c to 0.678, 31.08, and 2.9 for BG, TC, and AR, respectively.

E.2.3 Optimization

For all tasks, we used AdamW with parameters ($\beta_1 = 0.9, \beta_2 = 0.95$) and weight decay of $1e-5$ for all parameters except for the positional embedding in ClimaX and Stormer. We trained each model for 50 epochs with a batch size of 32, using a linear warmup schedule for 5 epochs, followed by a cosine-annealing schedule for 45 epochs. Table 10 shows the peak learning rate for each task.

Table 10: Learning rate for finetuning ClimaX in different downstream tasks

Task	Finetuning LR	Scratch Training LR
Medium-range weather forecasting	$5e-6$	$5e-4$
S2S forecasting	$5e-5$	$5e-4$
Climate downscaling	$5e-5$	$5e-4$
Climate data infilling	$1e-4$	$5e-4$
Climate model emulation	$5e-4$	$5e-4$
Extreme weather events detection	$5e-4$	$5e-4$

For finetuning ClimaX and Stormer, we used a smaller learning rate for tasks that are similar to pretraining and a larger learning rate for tasks that are more different.

E.2.4 Software and hardware stack

We use PyTorch (Paszke et al., 2019), `numpy` (Harris et al., 2020) and `xarray` (Hoyer & Hamman, 2017) to manage our data and model training. We also use `timm` (Wightman, 2019) for implementations of ClimaX and Stormer. All training is done on 8 NVIDIA RTX A6000 GPUs. We leverage `fp16` floating point precision in our experiments.

F Visualizations

F.1 S2S forecasting

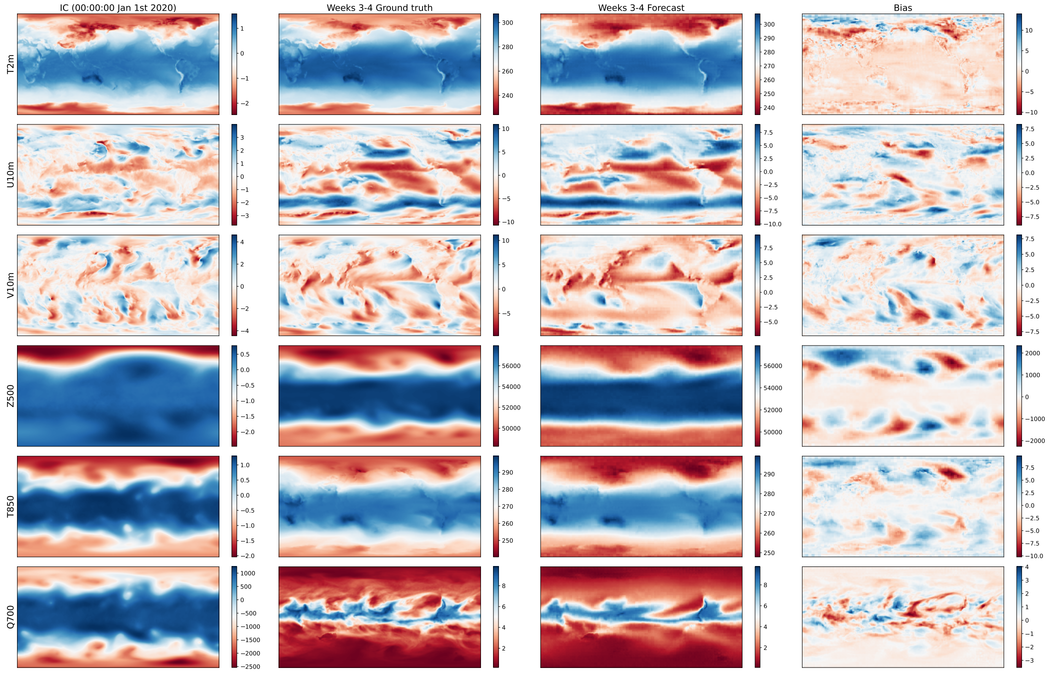

Figure 5: ClimaX forecasts for weeks 3-4 of six target variables.

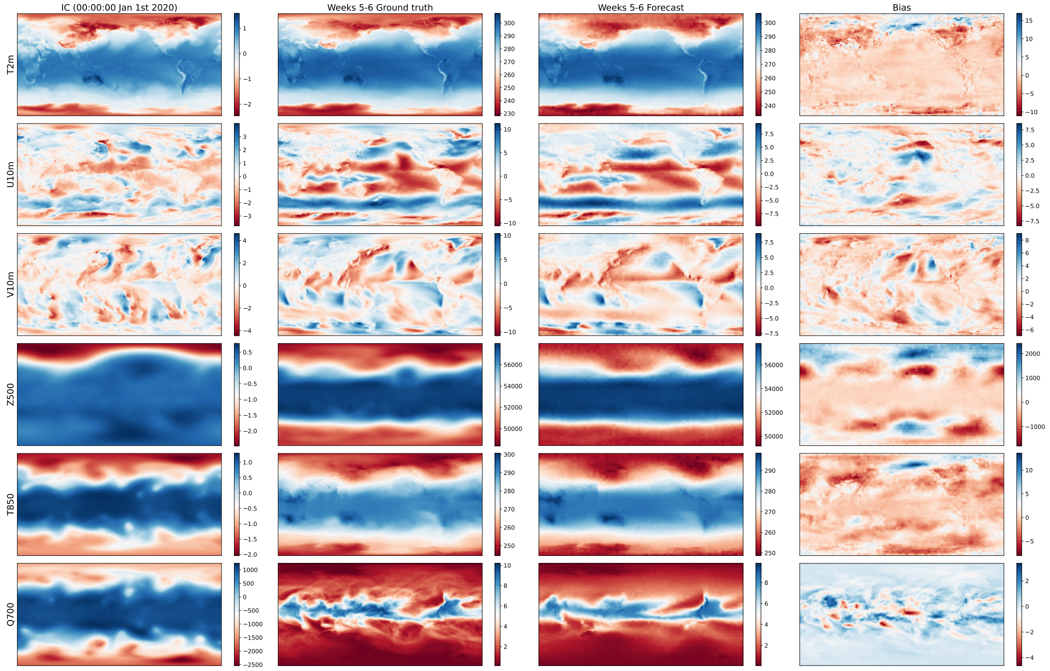

Figure 6: ClimaX forecasts for weeks 5-6 of six target variables.

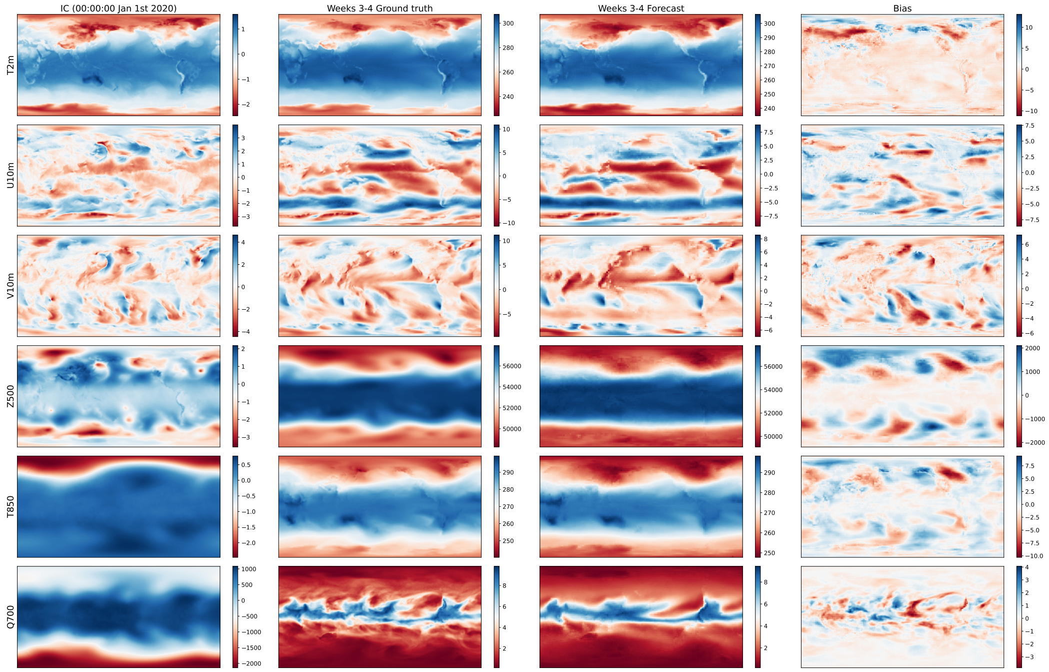

Figure 7: Stormer forecasts for weeks 3-4 of six target variables.

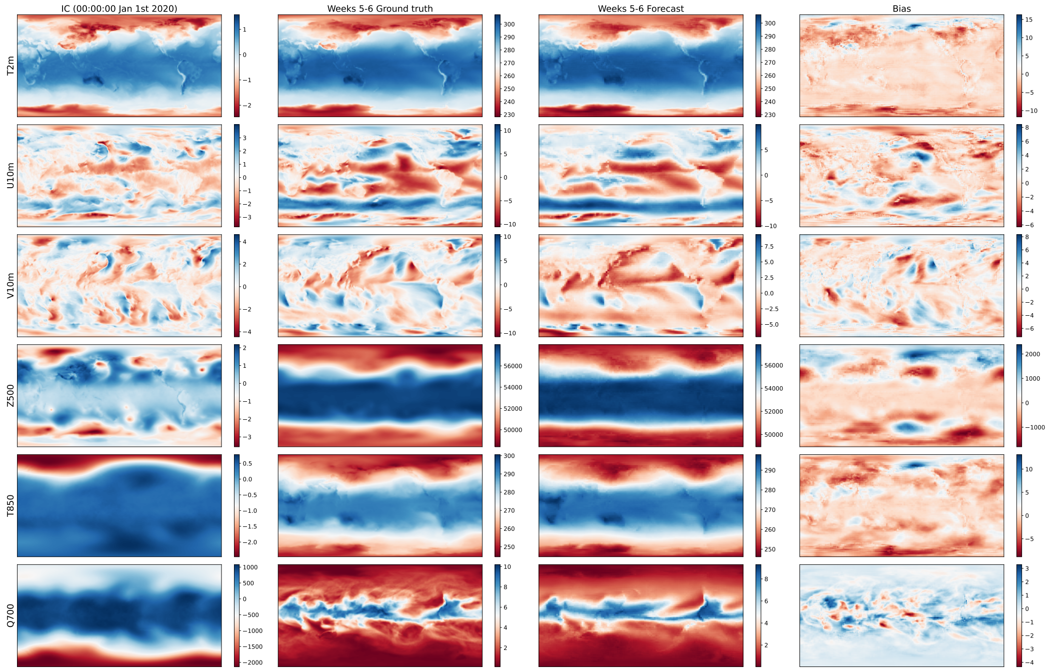

Figure 8: Stormer forecasts for weeks 5-6 of six target variables.

F.2 Climate downscaling

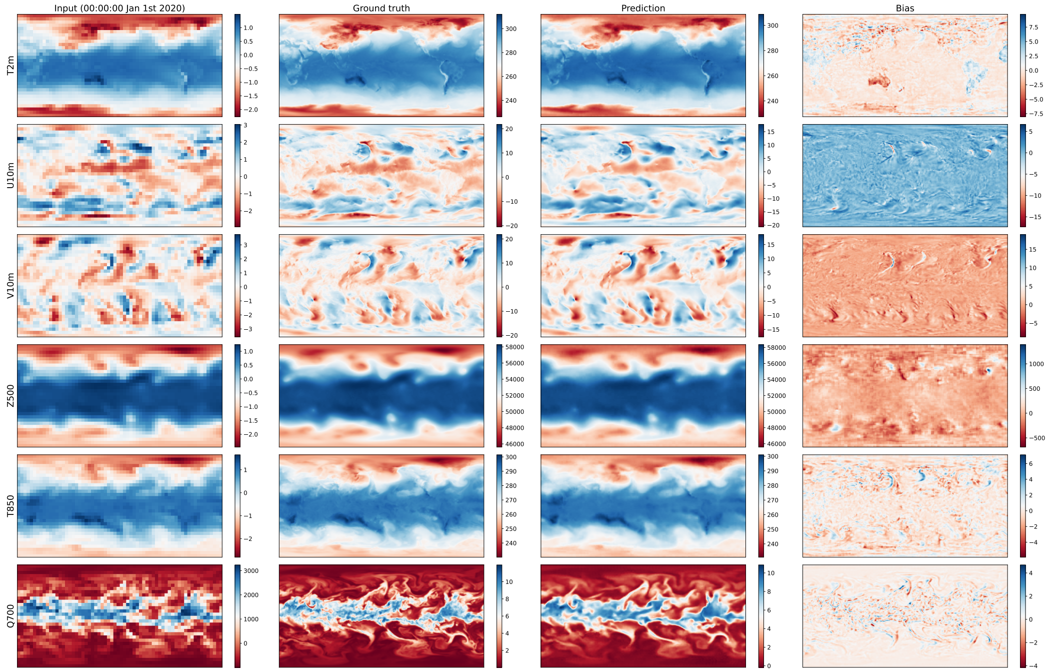

Figure 9: ClimaX downscaling predictions of six target variables.

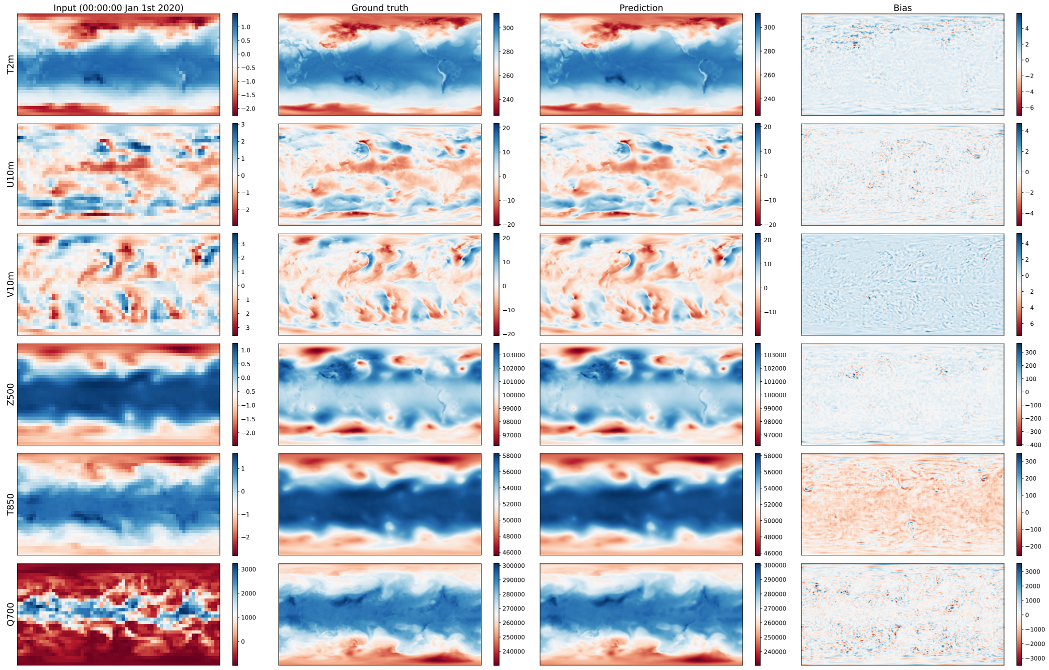

Figure 10: Stormer downscaling predictions of six target variables.

G Atmospheric chemistry

This section presents the atmospheric chemistry tasks that AtmosArena includes.

G.1 Atmospheric chemistry downscaling

Atmospheric chemistry simulations are essential for understanding various global processes such as air pollution, biogeochemical cycles, and climate change. High-resolution models can capture fine-scale chemical interactions, providing insights into local pollution levels and their health impacts. However, these models are computationally intensive. Deep learning offers a solution by transforming coarse-resolution inputs into finer-resolution outputs (Geiss et al., 2022). Specifically, the input is a grid of dimensions $V \times H \times W$, and the output is a higher-resolution grid $V' \times H' \times W'$, where $H' > H$ and $W' > W$. This allows for precise monitoring of atmospheric pollutants and their effects on human health and the environment, enabling more informed policy decisions and scientific research.

Dataset We utilize GEOS-CF, a simulated dataset from the NASA GEOS Composition Forecast (GEOS-CF) system (Knowland et al., 2022). GEOS-CF combines the NASA GEOS model with the GEOS-Chem chemical transport model to simulate the atmospheric composition (Keller et al., 2021). The dataset offers outputs on a $0.25°$ grid, which we downsample to $5.625°$ for the low-resolution input and $1.40625°$ for the high-resolution output. For our benchmark, we use the meteorological replay simulation ("das" files), covering the years 2018 to the present. We focus on downscaling the five near-surface atmospheric pollutants: NO2, SO2, CO, O3, and PM2.5, averaged over a 1-hour window ("chm_tavg_1hr" files).

G.2 Atmospheric composition forecasting

This task involves predicting the global atmospheric composition of important air pollutants such as carbon monoxide and carbon dioxide at different lead times. This task is crucial for understanding air quality, which directly impacts human health by influencing the prevalence of non-communicable diseases. This task presents a significant challenge to data-driven models due to the complexity of atmospheric dynamics and the influence of human activities on emission levels. The task formulation and input and output shapes are similar to weather forecasting.

Dataset We use CAMS Analysis maintained by ECMWF for the atmospheric composition forecasting task in AtmosArena. As part of the Copernicus Atmosphere Monitoring Service (CAMS), this dataset integrates meteorological variables with concentrations of air pollutants such as carbon monoxide and carbon dioxide, providing a comprehensive overview of global atmospheric composition. The dataset offers 12-hourly data at a $0.4°$ (450×900 grids) resolution from 2017 to the present. Similar to ERA5, we regrid this dataset to the common resolution of $1.40625°$ for easier training and evaluation.

H Additional experiments

H.1 Atmospheric chemistry experiments

H.1.1 Atmospheric chemistry downscaling

We consider the task of downscaling for five near-surface variables: NO2, SO2, CO, O3, and PM2.5. We use GEOS-CF at $5.625°$ as the low-resolution input, and GEOS-CF at $1.40625°$ as the high-resolution target, corresponding to $4\times$ upsampling. We use 2018-2020 for training, 2021 for validation, and 2020 for testing. Due to time and compute constraints, we only consider ClimaX finetuned and Unet as baselines.

Table 11 reports the MAE metric in the log space of the five target variables. ClimaX finetuned and Unet perform competitively. Given the results in climate downscaling, we believe fully finetuned Stormer will outperform Unet in this task.

Table 11: MAE of ClimaX finetuned and Unet for downscaling five target near-surface pollutants.

	NO2	SO2	CO	O3	PM2.5
ClimaX finetuned	0.069	0.049	0.405	**0.0065**	**0.100**
Unet	**0.064**	**0.047**		0.0071	0.104

H.1.2 Atmospheric composition forecasting

We compare ClimaX with Unet on forecasting eight near-surface pollutants: Total Column Carbon Monoxide (TC CO), Total Column Nitric Oxide (TC NO), Total Column Nitrogen Dioxide (TC NO2), Total Column Sulfur Dioxide (TC SO2), Total Column Ozone (TC O3), Particulate Matter 1um (PM1), Particulate Matter 2.5um (PM2.5), and Particulate Matter 10um (PM10), with lead times from 1 to 3 days. For each baseline method, we finetune a separate model for each specific lead time and target variable. We use CAMS Analysis from 2017 to 2020 for training, 2021 for validation, and 2022 for testing.

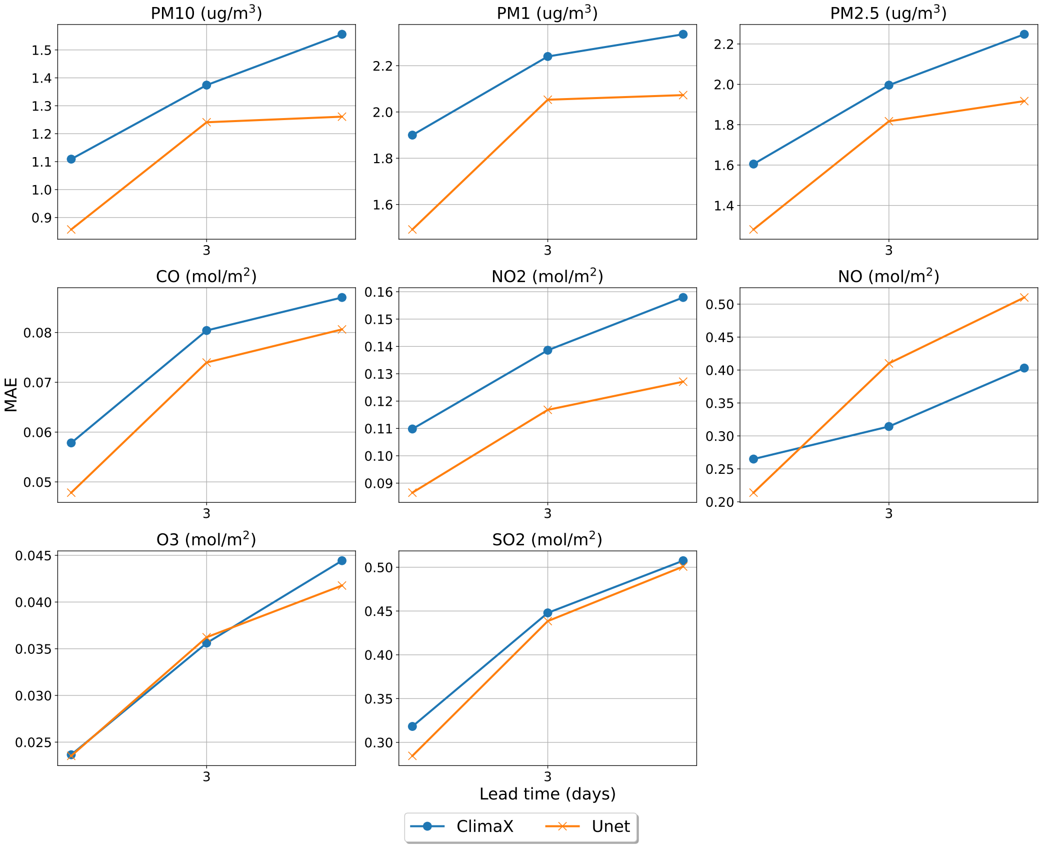

Figure 11: Air composition forecasting performance.

Figure 11 shows the performance of ClimaX and Unet on forecasting eight key pollutants from 1 day to 5 days. Unet outperforms ClimaX for almost all variables. This result shows that the temporal forecasting capabilities of pretrained models may not transfer well to new tasks in a new domain.

H.2 Additional metrics for atmospheric physics tasks

S2S forecasting In addition to RMSE and ACC, we report Spectral Divergence as a physics-based metric, which measures the discrepancy between the frequency components of the ground truth and prediction. Table 12 shows the superior performance of ClimaX frozen across all variables and lead times. This highlights the effectiveness of multi-source pretraining in obtaining a general-purpose backbone that can adapt to forecasting tasks with unseen time scales only via lightweight finetuning.

Table 12: S2S performance measured by Spectral Div on four target variables at two lead times.

		Z500		T850		T2m		Q700	
		Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6	Weeks 3-4	Weeks 5-6
Spectral Div (↓)	ClimaX frozen	0	0	**0.3153**	**0.2894**	**0.1671**	**0.1805**	**0.0789**	**0.0903**
	ClimaX finetuned	0	0	0.3224	0.3180	0.2298	0.2093	0.0930	0.0937
	Stormer frozen	0	0	0.3307	0.4161	0.4705	0.5971	0.5188	0.7513
	Stormer finetuned	0	0	0.3275	0.3024	0.6603	0.6105	0.4337	0.3468
	Unet	0	0	0.3863	0.5110	0.2065	0.4647	0.0809	0.8157

Downscaling Table 13 shows the Anomaly Pearson Coefficient of different baselines on the climate downscaling tasks. Stormer finetuned is the best method for all four variables. However, all baselines achieve very similar performances, suggesting Anomaly Pearson Coefficient may not be the best metric for distinguishing different models in this task. A similar result was observed in ClimaX (Nguyen et al., 2023a).

Table 13: Downscaling performance measured by Anomaly Pearson Coefficient on six variables.

		Z500	T850	T2m	Q700	U10	V10
Anomaly Pearson (↑)	ClimaX frozen	0.9963	0.9879	0.9833	0.9388	0.9690	0.9716
	ClimaX finetuned	0.9977	0.9907	0.9869	0.9532	0.9802	0.9813
	Stormer frozen	0.9956	0.9856	0.9821	0.9240	0.9654	0.9689
	Stormer finetuned	**0.9993**	**0.9951**	**0.9946**	**0.9626**	**0.9886**	**0.9894**
	Unet	0.9987	0.9931	0.9917	0.9613	0.9850	0.9861

Extreme weather events detection Table 14 shows the Specificity metrics of different methods in the extreme weather events detection tasks. ClimaX frozen is the best-performing method, showing the effectiveness of multi-source pretraining in transferring the backbone to a completely new task. However, the baselines perform very similarly for this metric, suggesting it may not be the best to evaluate methods in this task.

Table 14: Specificity Metrics of different methods for TC and AR detection.

	ClimaX frozen	ClimaX finetuned	Stormer frozen	Stormer finetuned	CGNet
TC	0.99	0.99	0.98	0.98	0.99
AR	0.96	0.96	0.95	0.95	0.92

H.3 Climate data infilling on Berkeley Earth

We test the models trained to perform infilling for ERA5 in Sections 4.4 on the Berkeley Earth dataset to examine their transferability between datasets. Similarly to ERA5, we generate a fixed set of masks during testing, with the mask ratio $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We test the models on infilling for data from 2020 to 2023. Figure 12 shows that all methods perform similarly for this dataset, and the performances do not get worse as we increase the mask ratio. We hypothesize that because of the distribution shift from ERA5 to Berkeley Earth, the best thing the models can do is to predict the average, leading to very similar performances among models and mask ratios.

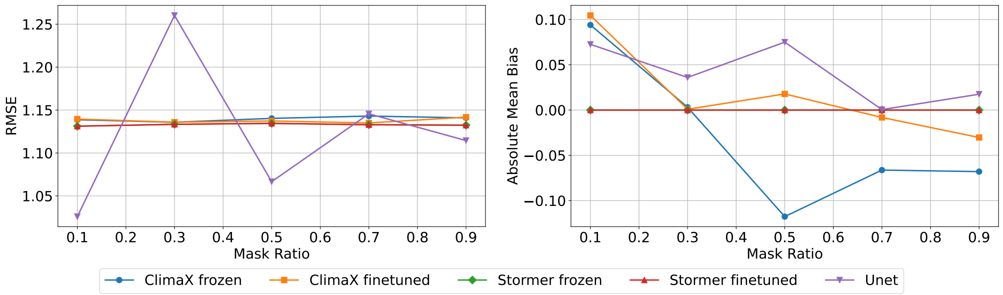

Figure 12: Performance of different models measured by RMSE and Absolute Mean Bias on infilling the Berkeley Earth temperature data with different mask ratios.