NOAA NCEP

NCEP OFFICE NOTE 521
https://doi.org/10.25923/xd3y-wy31

GFS-Powered Machine Learning Weather Prediction: A Comparative Study on Training GraphCast with NOAA's GDAS Data for Global Weather Forecasts

Sadegh Sadeghi Tabas,
Jun Wang, Wei Li, Mallory Row,
Zhan Zhang, Lin Zhu, Jiayi Peng,
Jacob R. Carley

US DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
College Park, MD
March 2025

GFS-Powered Machine Learning Weather Prediction: A Comparative Study on Training GraphCast with NOAA's GDAS Data for Global Weather Forecasts

Sadegh Sadeghi Tabas[2] (https://orcid.org/0000-0001-9157-3397)

Jun Wang[1] (https://orcid.org/0009-0007-9030-3417)

Wei Lei[3] (htpps://orcid.org/0000-0003-2251-3731)

Mallory Row[1]

Zhan Zhang[1] (https://orcid.org/0009-0003-7116-1530)

Lin Zhu[3] (https://orcid.org/0009-0000-8374-9671)

Jiayi Peng[2] (https://orcid.org/0000-0003-2251-3731)

Jacob R. Carley[1] (https://orcid.org/0000-0003-4763-6666)

[1]NOAA National Weather Service, National Centers for Environmental Prediction, College Park, MD, USA (https://ror.org/00ndyev54)

[2]Axiom at Environmental Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, MD, USA

[3]SAIC at Environmental Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, MD, USA

National Centers for Environmental Prediction Office Note 521
March 2025

Recommended citation
Tabas, S. S., J. Wang, W. Li, M. Row, Z. Zhang, L. Zhu, J. Peng, and J. R. Carley. (2025). GFS-Powered Machine Learning Weather Prediction: A Comparative Study on Training GraphCast with NOAA's GDAS Data for Global Weather Forecasts. U.S. Dept. of Commerce, NOAA NCEP Office Note 521, xx p. doi: https://doi.org/10.25923/xd3y-wy31

Acknowledgements
The Parallel Works team is acknowledged for their support in providing cloud computing resources and resolving cloud infrastructure issues. The authors would like to thank the Google DeepMind team for their constructive comments on the methodology and the Weather Bench 2 team for providing Zarr databases of HRES and ERA5 reanalysis and climatology data. The authors acknowledge the high performance computing resources provided by the NOAA Research and Development High Performance Computing Program. The project described in this article was supported by the Inflation Reduction Act as well as the NOAA Software Engineering for Novel Architectures (SENA) project. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

Copies of this report are available from
https://repository.library.noaa.gov/

Abstract

Accurate medium-range global weather forecasts serve as a critical cornerstone in informing decision-making processes across various societal and economic sectors. The development of machine learning (ML)-based models has significantly changed the field of weather prediction in recent years, demonstrating previously unheard-of levels of effectiveness when contrasted with conventional numerical weather prediction (NWP) models. These cutting-edge models leverage diverse ML architectures, such as Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), Fourier Neural Operators (FNOs), and Transformers. Among these advancements, GraphCast, a pioneering ML-based approach developed by Google DeepMind, has received particular attention in the community. Leveraging direct training from reanalysis data, GraphCast expedites global weather predictions across numerous variables within minutes. Impressively, GraphCast forecasts show improved accuracy in predicting severe weather events, including phenomena such as tropical cyclones (TC), atmospheric rivers, and extreme heat. The performance of the current version of the GraphCast relies on high-quality historical weather data for training, typically sourced from the European Center for Medium Range Weather Forecast (ECMWF)'s ERA5 reanalysis. Concurrently, the National Centers for Environmental Prediction (NCEP) has initiated collaborative endeavors with the research community to develop machine learning weather prediction (MLWP) models. Within this context, our study represents the efforts to advance the state-of-the-art by devising a methodology for parallel training of GraphCast using Global Data Assimilation System (GDAS) data obtained from NCEP’s current operational Global Forecast System (GFS version 16). GDAS provides real-time initial conditions to make the experimental real-time MLWP global forecasts possible. Our study includes a framework that includes model training, validation, and testing processes, along with a performance comparison of GraphCast. In addition to this comparative analysis, we examine the benefits and drawbacks of GraphCast's forecasting ability using GDAS data and suggest possible ways to improve subsequent iterations of this research.

Keywords: Global Data Assimilation System (GDAS), GraphCast, Machine Learning-based Weather Prediction, Graph Neural Networks

1. Introduction

Weather forecasting is a critical application of scientific computing, providing invaluable insights into future weather patterns and the potential occurrence of extreme weather events such as floods, droughts, hurricanes, and more. These forecasts play a pivotal role in various aspects of society, including daily activities, agriculture, energy production, transportation, and industrial operations. In the realm of weather forecasting, traditional numerical weather prediction (NWP) models have long been the cornerstone, leveraging High-Performance Computing (HPC) architectures extensively to simulate atmospheric conditions. Over the past decade, the rapid advancement of HPC systems has catalyzed significant progress in this field (Bauer et al., 2015).

Conventional NWP methods typically adopt a simulation-based approach. This approach uses numerical methods to solve partial differential equations (PDEs) that mathematically represent the physical laws governing atmospheric conditions (Ritchie et al., 1995; Molteni et al., 1996; Skamarock et al., 2008). However, the computational demands of solving these PDEs are substantial. For instance, with a spatial resolution of 0.25 degrees, a single simulation for a 10-day weather forecast can require hours of computation across multiple nodes on an HPC cluster (Bauer et al., 2020). This computational burden limits the higher resolution real-time capability of daily weather forecasts and restricts the number of ensemble members feasible for probabilistic weather prediction. Additionally, parameterized numerical representations of atmospheric processes are a major component of conventional NWP models, but they are frequently considered inadequate (Palmer et al., 2005; Allen et al., 2002) despite their complexity (Bauer et al., 2015). These parametric models introduce errors due to the simplification of unresolved phenomena, posing challenges to the accuracy and reliability of weather forecasts. The weather and climate history, however, is detailed in the reanalysis and reforecasts generated by the NWP models (Kalnay et al., 1996; Saha et al., 2010; Hersbach et al., 2020; Hamill et al., 2022; Guan et al., 2022). The NWP models, however, mainly rely on advancements in model design, numerical schemes and algorithms, bias correction, and calibrations based on these historical data to enhance the model rather than directly using these data to increase accuracy.

Machine learning-based weather prediction (MLWP) offers a promising alternative and enhancement to traditional NWP. MLWP harnesses historical data to train forecast models directly with billions of neural network parameters. The purpose is to resolve the physics laws hidden in the data and to capture intricate patterns and scales. This new approach leverages the advancements in computing power, data storage, and Artificial Intelligence (AI) algorithm breakthroughs to build ML models for weather and climate predictions. The first cutting-edge data-driven MLWP model was introduced in 2022 by Keisler (2022). Keisler conducted a groundbreaking study that introduced a data-driven approach for global weather forecasting using graph neural networks (GNNs). This innovative system was designed to predict the future 3D atmospheric state every six hours autoregressively, with successive steps combined to generate accurate forecasts spanning multiple days ahead. When compared to previous data-driven approaches, the model demonstrated significant performance improvements on important metrics like Z500 (geopotential height at 500 hPa) and T850 (temperature at 850

hPa) after being trained on ERA5 reanalysis data (Hersbach et al., 2020). Most importantly, the model's performance was found to be comparable to operational, full-resolution physical models from GFS and ECMWF. This research marks a milestone in weather forecasting, demonstrating the potential of data-driven approaches to rival traditional physical models in accuracy and forecast skill.

NVIDIA, Pathak et al. (2022) developed the data-driven MLWP model FourCastNet, which is especially effective at forecasting small-scale variables like precipitation and surface wind speed and producing high-resolution forecasts remarkably fast and efficiently. Huawei's Pangu-Weather model (Bi et al., 2022) achieves superior forecasting accuracy across various factors, surpassing conventional NWP methods. Similarly, GraphCast from Google DeepMind (Lam et al., 2023) offers efficient, high-resolution forecasts and excels in predicting severe weather events. Chen et al. (2023) introduced the FuXi model, which offers ensemble forecasts to handle uncertainty and increases the accurate lead time for Z500 and T2M ((temperature at 2 meters) forecasts. Lastly, GenCast by Google (Price et al., 2023) delivers skillful ensemble forecasts for up to 15 days, outperforming ECMWF ensemble system (ENS) in multiple verification metrics. More recently, ECMWF introduced the Artificial Intelligence Forecasting System (AIFS) in real-time operations (Lang et al., 2024), based on a graph neural network (GNN) that has a sliding window transformer processor, an encoder, and a decoder. AIFS is trained on ECMWF's ERA5 reanalysis and ECMWF's operational NWP analyses. It has a flexible and modular design and supports several levels of parallelism to enable training on high-resolution input data. Results indicated that AIFS produces highly skilled forecasts for upper-air variables, surface weather parameters, and Tropical Cyclone (TC) tracks. These advancements collectively enhance the accuracy, speed, and reliability of weather forecasting, improving early warning systems and disaster preparedness.

The majority of MLWP models have been trained and verified primarily with ECMWF ERA5 reanalysis data (Hersbach et al., 2020). One question is whether these MLWP models can be trained to produce good performance forecasts with different initial conditions generated at other operational centers and if these models can learn additional information from analysis data generated from these centers. To address this question, researchers at the NOAA National Centers for Environmental Prediction (NCEP) are collaborating with the broader research community to advance MLWP methodologies. This study represents efforts aimed at introducing a new method for distributed and parallel training of the GraphCast model using different sources as ground truth, including the NCEP Global Data Assimilation System (GDAS) data from the operational Global Forecast System (GFSv16) model. The GDAS data used in this study is 6 hourly 0.25 degrees latitude-longitude products post-processed from 13 km high resolution runs. We proposed multiple scenarios to fine-tune the pre-trained GraphCast model on ERA5 reanalysis leveraging GDAS data. It includes initializing GraphCast with GDAS data, fine-tuning and training Graphcast with GDAS, and evaluating the forecast results from these experiments. Once trained, the new MLWP model runs remarkably faster than the operational NWP model; it produces 10-day forecasts in 2 minutes using a single Nvidia A100 Graphical Processing Unit (GPU) core. In conclusion, we tried to clarify the benefits and drawbacks of

using GraphCast to leverage GDAS data in addition to developing the framework for MLWP training processes. We also suggested possible directions for improving this study in the future.

Through the use of GDAS data in the MLWP framework, our study pushes the boundaries of weather forecasting at NCEP. Additionally, it clarifies the effectiveness of various initial conditions and training datasets and paves the way for improved MLWP forecast performance in operational centers. The remainder of this paper is organized as follows: Section 2 provides a description of several datasets utilized in this study, the GraphCast model structure, as well as the experimental design and model training procedure. In Section 3, we elaborate on the results obtained from the fine-tuned and trained GraphCast model using GDAS data. The model state vertical structure of forecasts is analyzed, the tropical cyclone forecast errors are verified, and the model forecasts are evaluated using standard metrics. Finally, Section 4 concludes the study by summarizing key findings and insights gained, as well as outlining potential directions for future research endeavors in the field of MLWP.

2. Methodology

2.1. Datasets

We utilized several datasets throughout our experiments with GraphCast training and fine-tuning. This section provides a detailed inventory of all the datasets used in the process. We built the datasets from GDAS analysis (Kleist et al., 2023) and ERA5 reanalysis (Hersbach et al., 2020). We got these datasets from NOAA and ECMWF archives, a large corpus of data representing the global weather at 0.25-degree latitude/longitude resolution for hundreds of static, surface, and atmospheric variables. We extracted a subset from these datasets from March 21, 2021 to January 1, 2024 at 6-hourly time steps. In order to facilitate the efficient storage and retrieval of large-dimensional datasets for model training and verification, all of the datasets, along with the ERA5 climatology data, are stored in a cloud-based Zarr format (Miles et al., 2020).

GDAS analysis data (Kleist et al., 2023) is an extensive meteorological dataset produced by NCEP. Through data assimilation in the GFS system, this dataset combines data from multiple observational sources, such as satellite measurements, surface observations, aircraft data, and radiosonde data. GDAS analysis provides a consistent and continuous record of atmospheric conditions, offering detailed information on the whole globe for variables such as temperature, humidity, wind speed, and pressure across model vertical levels. The operational GFSv16 GDAS data is accessible from March 20, 2021, to the present through platforms like NOAA's National Centers for Environmental Information (NCEI) or NOAA’s Amazon Web Services (AWS) S3 bucket.

This study also uses forecast data from the operational GFS model (NOAA, 2024a). The model provides operational forecasts with a base resolution of 13 km. Here, we use the 0.25-degree post-processed products for evaluation.

ERA5 reanalysis (Hersbach et al., 2020) is a global reanalysis dataset from ECMWF for weather and climate. Similar to GDAS, since reanalysis data combines observational data with simulation data, it provides the best estimate of the total state of the weather at 0.25 degree resolution. We used ERA5 climatology to compute anomaly metrics such as the anomaly correlation coefficient (ACC). In this study, the climatology data was gathered from WeatherBench2, and computed for 1990-2019 using a running window for smoothing (Rasp et al., 2024) for each day of the year and the sixth hour of the day. The climatologies are computed for 1990-2019.

The operational high-resolution global analysis data from ECMWF (HRES-T0 Analysis; we refer to it as HRES in this study) that starts the ECMWF high-resolution run is also utilized as the ground truth in addition to ERA5 reanalysis. Both data sets are used to evaluate the quality of the machine learning model forecasts (Rasp et al., 2024).

2.2. GraphCast Model

In this study, we fine-tuned GraphCast, a novel MLWP approach developed by Lam et al. (2023). The fine-tuned model is designed for global medium-range weather forecasting, leveraging NOAA's GDAS data as well as ECMWF HRES and ERA5 reanalysis data. GraphCast supports various applications, including the prediction of TC tracks, atmospheric rivers, and extreme temperatures (Lam et al., 2023). In the inference stage, GraphCast produces a 10-day forecast in under a minute on a single-core GPU (or TPU) with at least 32GB of memory. GraphCast predicts future atmospheric states by using the current time and previous six-hour model states as inputs. The resolution used in this study is a 0.25-degree latitude/longitude grid. A set of surface and atmospheric variables (Table 1 in Lam et al., 2023) are located on each grid point. GraphCast produces forecasts through autoregressive steps. The model uses the forecasts from previous steps as inputs to generate forecasts at the next step, as shown in Figures 1.b and 1c in Lam et al. (2023).

2.2.1. Model Architecture

GraphCast is a neural network architecture with 36.7 million parameters that is based on GNNs and is implemented in an encode-process-decode configuration (Battaglia et al., 2018). The encoder, processor, and decoder parts of GraphCast are described in the following subsections. For further information, please see Lam et al., 2023.

A single GNN layer is used by the encoder component of GraphCast (Lam et al., 2023) to translate model variables on the input latitude-longitude grid into node attributes on internal meshes that the GNN processor can use. The internal multi-meshes are located on resolution regular icosahedron grids on the sphere created iteratively six times. As a result, the multi-mesh facilitates high-resolution spatial representation over the globe with less amount of total data than the original latitude-longitude input data. The processor, which contains 16 separate GNN layers to execute message passing across the multi-meshes, allows effective transmissions of information at different scales through a few number of message-passing iterations. The features learned by the processor at the last processor layer in the latent multi-meshes space are transformed back into forecasts on the latitude-longitude output grid by the GraphCast decoder component. This output can be used as input for the next step forecast.

2.2.2. Available GraphCast Versions

Currently, there are two versions of GraphCast weights available, both trained (and developed) by Google DeepMind. The first version (GraphCast-operational) has 13 pressure levels trained on 1979-2017 ERA5 reanalysis data and fine-tuned on 2016-2021 using HRES analysis data. The second version (GraphCast) has 37 pressure levels trained on 1979-2017 ERA5 reanalysis data. This study focused on GraphCast-13 due to its more accurate forecasts as it is fine-tuned on most recent HRES data (for more information, please see https://sites.research.google/weatherbench/) as well as its lower computational cost for fine-tuning.

2.3. Experimental Design and Model Training

In this study, we looked at the same structure with precisely the same number of parameters as DeepMind offered for the 13-pressure-level GraphCast operational model. In the following subsections, we provide details of our training setup to fine-tune the GraphCast model with GDAS data.

2.3.1. Training/Fine-tuning Data and Schedule

As we discussed in Section 2.1, we considered multiple sources of data for fine-tuning GraphCast. We prepared Zarr databases and generated NetCDF batch files; each file includes 16 weather state time steps that match the GraphCast inputs (current and past 6-hour weather states) as well as the upcoming 14 weather states as truth data. The input states in all cases were provided from GDAS data (from March 20, 2021, to Jan 1, 2023), while the truth data were selected from multiple sources, including GDAS analysis, ERA5 reanalysis, and HRES analysis data. We split the data into training and validation sets from March 20, 2021, to September 1, 2022, and September 1, 2022, to Jan 1, 2023, respectively (4 cycles per day). For the verification step, we verified the model forecasts against the GDAS analysis and ERA5 reanalysis data, and compared them with GFS forecasts for the entire year 2023 (two cycles per day; 00z and 12z). Table 1 explains different scenarios considered in this study. We call these models GCGFS (GraphCastGFS) for simplicity. Please note Scenario #1 and #2 are the same model fine-tuned with GDAS but verified against different data sets (ERA5 reanalysis for Scenario #1 or GDAS analysis for Scenario #2).

Scenario #	Training/Fine-Tuning	Input Dataset	Truth Dataset	Verification Dataset	# of AR Training Steps
1	Fine-Tuning	GDAS	GDAS/HRES/ERA 5	GDAS	14
2	Fine-Tuning	GDAS	GDAS/HRES/ERA 5	ERA5	14
3	Training	GDAS	GDAS	GDAS	12

Table 1. GCGFS training and verification configurations

2.3.3. Training Setup

The GraphCast training setup code was not supplied by the DeepMind team because it was closely linked to their internal infrastructure. We obtained the core model from the DeepMind GraphCast repository and developed our own training setup code and were able to train and fine-tune GraphCast using multiple GPU cores.

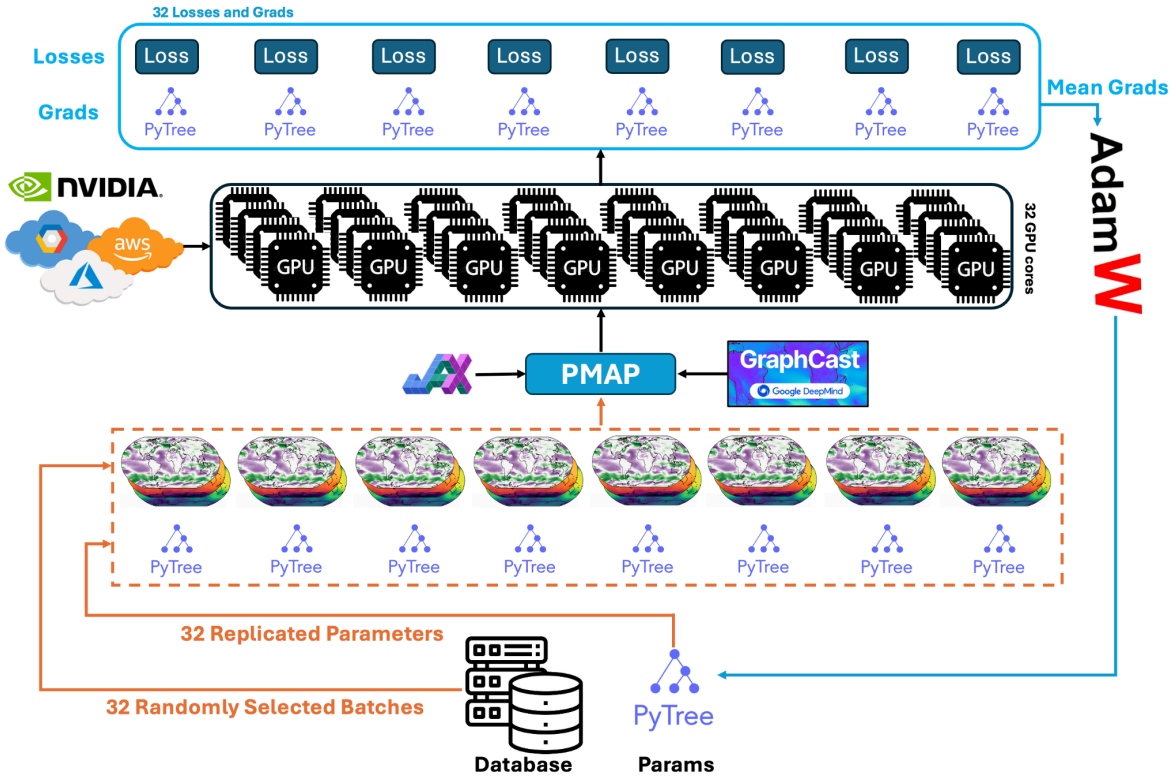

Figure 1. EMC training setup for fine-tuning GraphCast model

To train or fine-tune GraphCast, we leveraged pre-trained model weights provided by DeepMind as the initial weights. Similar to the original setup of GraphCast, we considered processing 32 batches (mini-batch) for every weight update. We utilized 32 GPU cores of NVIDIA H100 nodes, each with 80 GB of memory, provided through NOAA Parallel Works with AWS cloud nodes (4 nodes of P5 instances, each with 8 GPU cores). For fine-tuning (scenario #1 and #2), we trained the model for 2 more AR steps starting from step 12, where the Google Deepmind stops to let the model learn different initial conditions and to improve the longer lead time forecasts. For scenario #3, we trained the model following the curriculum training schedule (Lam et al., 2023). We started from AR step 1 to let the model learn the impact of different inputs from the beginning of the forecast. Due to computational costs and limited GPU node availability, we trained the model for 12AR steps as Deepmind did. To train the model with parallelization, we followed a Data Parallel (DP) paradigm. The GraphCast model, as well as its parameters, is replicated on 32 GPU cores while we randomly selected 32 batches with replacement from training data in batch files prepared in advance for the whole training and validation cycles. Using Jax's PMAP mechanism (Sapunov, 2024), the batch files were

distributed parallelly on 32 GPU cores. The GraphCast model and its parameters are copied over and run with the decomposed data. Next, the losses, as well as the gradients, were aggregated from 32 mini-batches, and the model weights were updated using the AdamW optimization method (Loshchilov, 2017). The training objective in the loss function is computed following Lam et al. (2023). To enhance the training process and prevent overfitting, we implemented early stopping steps and monitored convergence. Specifically, for every fine-tuning step explained above, the training process is stopped when the loss value ceases to improve for 25 consecutive epochs. This approach ensures that the model maintains generalization capabilities by stopping the training at the optimal point before overfitting occurs. Figure 2 shows the loss change during the fine-tuning (Scenario #1).

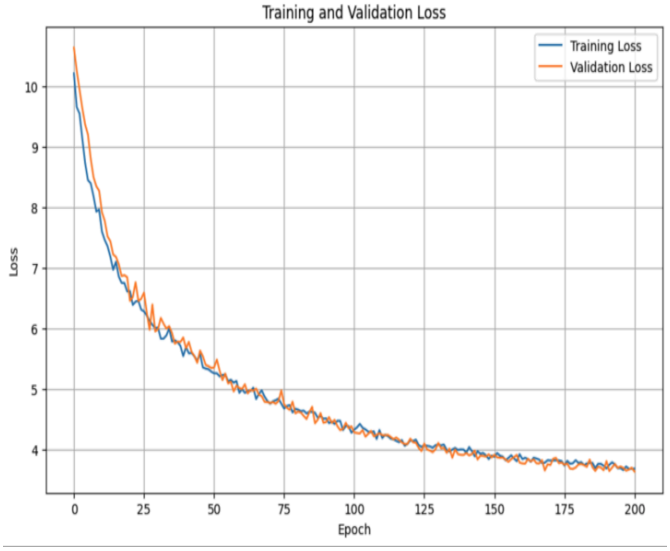

Figure 2: Training and validation loss for fine-tuning GraphCast with GDAS as input and ERA5 for truth.

Given that there were four AWS P5 nodes available, each with 32 GPU cores, the entire fine-tuning and training procedure took about four weeks. It is worth noting that sometimes we were only able to reserve two (or even one) nodes of AWS P5 instances, so we applied a loop of two (or four) steps prior to combining the gradients and losses to ensure consistency with the available time for each of the four nodes.

2.4. Verification and Evaluation Metrics

We compared the results with operational GFS forecast validation after using WeatherBench2 (Rasp et al., 2024) to validate the forecasts for the full year 2023 against GDAS analysis (Scenarios #1 and #3) and ERA5 reanalysis (Scenario #2). The goal is to show the improvement relative to operational GFS as well as the impact from the different verification data sets. The forecasts were verified over the global and regional domains, including North America, the Northern Hemisphere, the Southern Hemisphere, and Tropics. To quantify the skillfulness of GraphCast, we utilized the root mean square error (RMSE) and the anomaly correlation coefficient (ACC) metrics. The RMSE measures the magnitude of the differences between forecasts and ground truth for a given variable and a given lead time. The RMSE is defined in WeatherBench2 Equation (2) (Rasp et al., 2024). We list it here as a reference:

$$RMSE_{\ l} = \sqrt{\frac{1}{TIJ}\sum_{t}^{T}\sum_{i}^{I}\sum_{j}^{J}\omega(i)(f(t,l,i,j) - o(t,i,j))^{2}}$$

The ACC is defined following the equations (4) and (5) in Rasp et al. (2024). The ACC is listed here as a reference.

$$f'_{\ t,l,i,j} = f_{\ t,l,i,j} - c_{\ t,l,i,j};\ o'_{\ t,i,j} = o_{\ t,i,j} - c_{\ t,i,j}$$

$$ACC_{\ l} = \frac{1}{T}\sum_{t}^{T}\frac{\sum_{i}^{I}\sum_{j}^{J}\omega(i)f'_{\ t,l,i,j}o'_{\ t,i,j}}{\sqrt{\sum_{i}^{I}\sum_{j}^{J}\omega(i)f'^{2}_{\ t,l,i,j}\sum_{i}^{I}\sum_{j}^{J}\omega(i)o'^{2}_{\ t,,i,j}}}$$

After the training, we set up experimental real-time forecasts from Scenario #1, from April 28, 2024, and Scenario #3, from September 11, 2024. We evaluated the atmosphere states and surface fields against GDAS using the NCEP's Verification Statistics Data Base (VSDB, Zhou et al., 2015; Shafran et al., 2015) system for Boreal fall during September 11 to October 11, 2024. We also evaluated several hurricane cases and computed some statistics for TC track and intensity errors using NCEP's hurricane verification package (Marchok T., 2021; Franklin JL, 2010).

3. Results and Discussion

We compared the fine-tuned and trained GraphCast model results with GFS forecasts for the entire year 2023 for 2 cycles per day (00Z and 12Z; total 730 cycles). Just like the operational GFS, the refined model (Scenario #1) and the fully trained model (Scenario #3) are validated against GDAS analysis. We also validated these results using ERA5 reanalysis data (Scenario #2 in Table 1) to demonstrate the impact of training data sets on forecasting abilities. In the following subsections, we presented the verification results for global and the tropics region. As we got similar results as the globe for regions including North America, the Northern Hemisphere, the Southern Hemisphere, and the Tropics, we presented the results for these regions in the supplementary section. From now on, for simplicity, we use "verification" referring to "verification against GDAS analysis", unless we specifically mention the verifications that are against ERA5 reanalysis data.

3.1. Forecast verification against operational GFS

Figure 3 shows the global average RMSE values of the fine-tuned GCGFS (Scenario #1), trained GCGFS (Scenario #3), and fine-tuned GCGFS against ERA5 (Scenario #2), as well as operational GFS forecasts. The x-axis is 6 hourly forecast steps, up to 10 days. We utilized WeatherBench2 (RMSE is defined in Section 2.4) to calculate the metrics for the entire year of 2023 for surface and atmospheric variables.

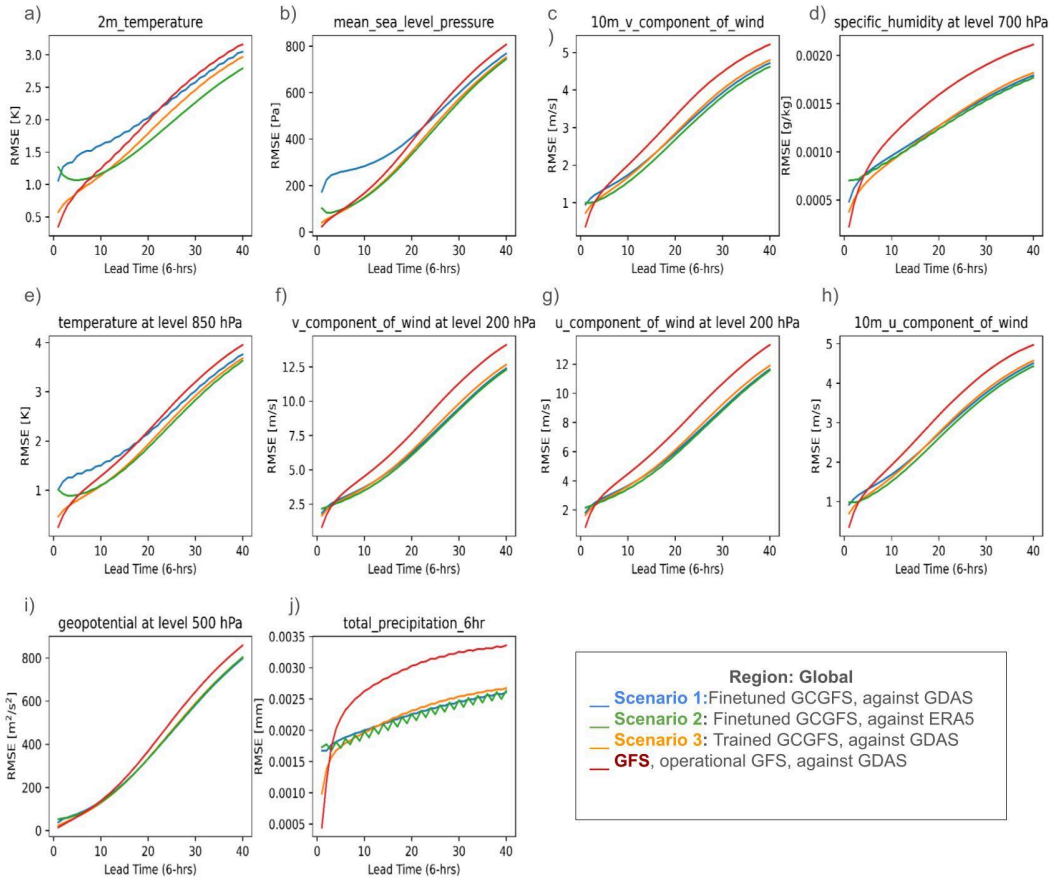

Figure 3. RMSEs of fine-tuned GCGFS (blue), trained GCGFS (orange), and fine-tuned GCGFS against ERA5 (green), and GFS (red) forecasts for the full year 2023 (00z and 12z). a) - j) are RMSEs for the following fields: 2 meter temperature (a) , mean sea level pressure (b), 10 meter v wind component (c), 700 hPa specific humidity (d), 850 hPa temperature (e), 200 hPa v wind component (f), 200 hPa v wind component (g), 10 meter u wind component (h), 500 hPa geopotential, (i) and 6 hourly total precipitation (j).

From the figure, it shows that GCGFS forecasts are similar in the three scenarios and have smaller RMSE compared to operational GFS at longer lead time (> 2 days) for fields including 10 meter and 200 hPa wind fields, 700 hPa specific humidity, 500 hPa geopotential, and 6 hourly total precipitation. For example, the GCGFS 500 hPa geopotential forecasts have about 10% reduced RMSE compared to GFS on day 10 (figure 3i). These fields have a larger RMSE than operation GFS at forecast time less than 2 days, especially for fine-tuned GCGFS (scenario 1 and 2). This is probably caused by the long AR step (12-14) training in the fine-tuning process when we intend to improve the long lead time forecast skills. In scenario 3, where the model was trained from AR steps 1 to 12, the RMSE is reduced compared to that in scenarios 1) and 2). It is still somewhat bigger than GFS. This may result from two causes: 1) the 12 AR steps training process that targets improving forecasts at long lead time may surrender short time forecast skills as the training weights are updated when the AR step increases; 2) insufficient training data (2 years of GDAS data) that results in inadequate learning

to effectively reduce the initial shock caused by different input data. It is worth noting that 2 meter and 850 hPa temperature fields and mean sea level pressure show larger RMSE up to 4-5 days lead time than GFS in scenario 1, while when verified against ERA5 (scenario 2) , those fields show smaller RMSE than GFS after GFS after day 1. This indicates that there are noticeable differences between ERA5 reanalysis and GDAS analysis for those fields; also, even though the model is fine-tuned with GDAS, the 2 meter temperature field still performs closer to ERA5 in longer lead time (see blue and green lines in Figures 3a, 3b, and 3e). In both scenarios 1 and 2, the initial shock is prominent. The two reasons for the AR steps in the fine-tuning process and inadequate training data may also apply here. We plan to redo the training later when we have enough GDAS data and additional resources. In this paper, we focus on the initial model fine-tuning and training process and verify the ML forecasts with common metrics.

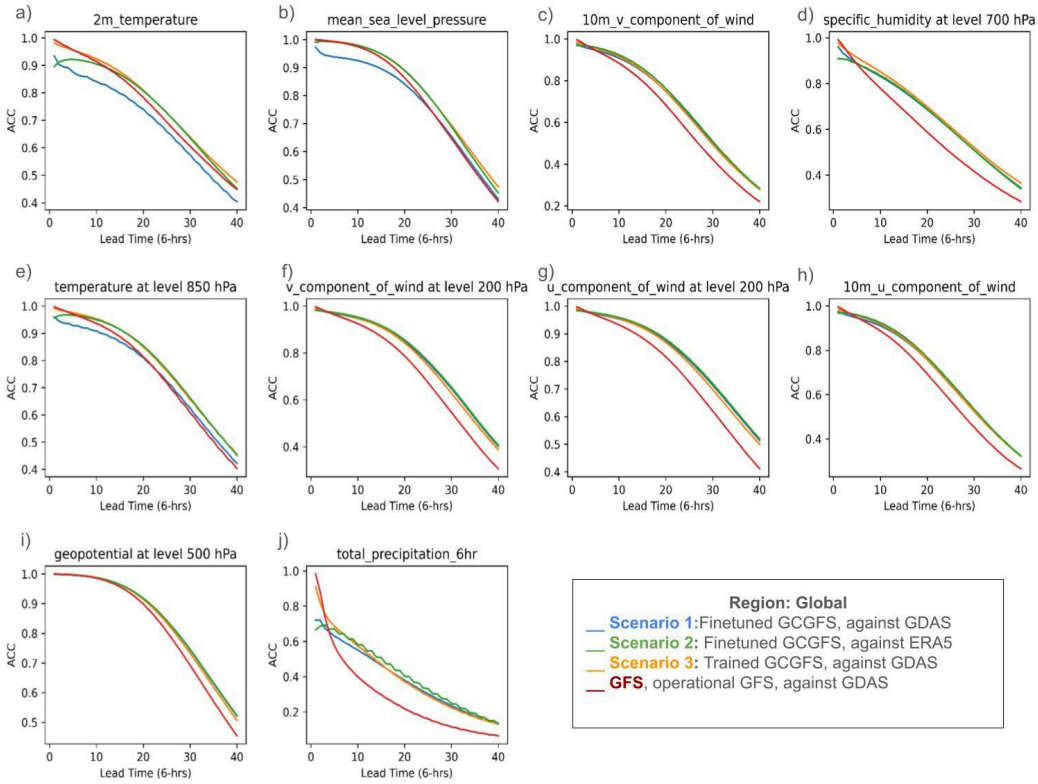

Figure 4. as in Figure 3, but for ACC metrics.

Figure 4 is the same as Figure 3 but shows the global average ACC metrics. Similar improvements have been seen in the ACC scores for both the atmospheric and surface fields at longer lead times. The ACC scores are very close in the three GCGFS scenarios except that the forecasts in scenario 3 have higher ACC scores at the short lead time, which is very close to GFS. It is worth noting that the 2 meter temperature ACC score in scenario 1 is noticeably below the GFS ACC score, while in scenario 3 it is about the same as GFS on day 1 and then higher than GFS after day 1. This indicates the training starting from the short lead time allows the GCGFS to learn adjusting predictions from different inputs at that lead time. The results confirm the training's efficacy even more.

3.2. Verification for boreal fall from NCEP VSDB verification system

The figures below show the performance of fine-tuned GCGFS and trained GCGFS compared to operational GFS (GFSv16). Figures 5 and 6 show the vertical structure differences of two model fields compared to GFSv16. The evaluations only show pressure levels up to 50 hPa, as the GraphCast operation version we used in this study only has 13 pressure levels with 50 hPa at the top. Figure 5 shows the boreal fall RMSE comparison of geopotential height (HGT, Fig. 5a-c) and temperature (T, Fig. 5d-f) averaged over September 11 to October 11, 2024, in three regions (Northern Hemisphere, Southern Hemisphere, and Tropics). It is clear that two GCGFS models reduce RMSEs for both fields in most of the troposphere in the three regions at long lead time (green area). However, both models show larger errors compared to GFSv16 for the two verified fields at the pressure levels above 100 hPa. The fine-tuned GCGFS has the largest error reduction in the two fields in the northern hemisphere among the three regions. The model has more than 9.6 gpm reduction in HGT RMSE at pressure levels between 300 and 200 hPa. The model also shows increased errors in the three regions at short lead times, especially for temperature. The RMSEs of HGT in the southern hemisphere and temperature in the northern hemisphere increase at the lower atmosphere, where pressure levels are below 850 hPa for all the lead time. Compared to fine-tuned GCGFS, in general the trained GCGFS has lower error in the lower atmosphere. This result could be attributed to the training process, which includes short lead time training steps.

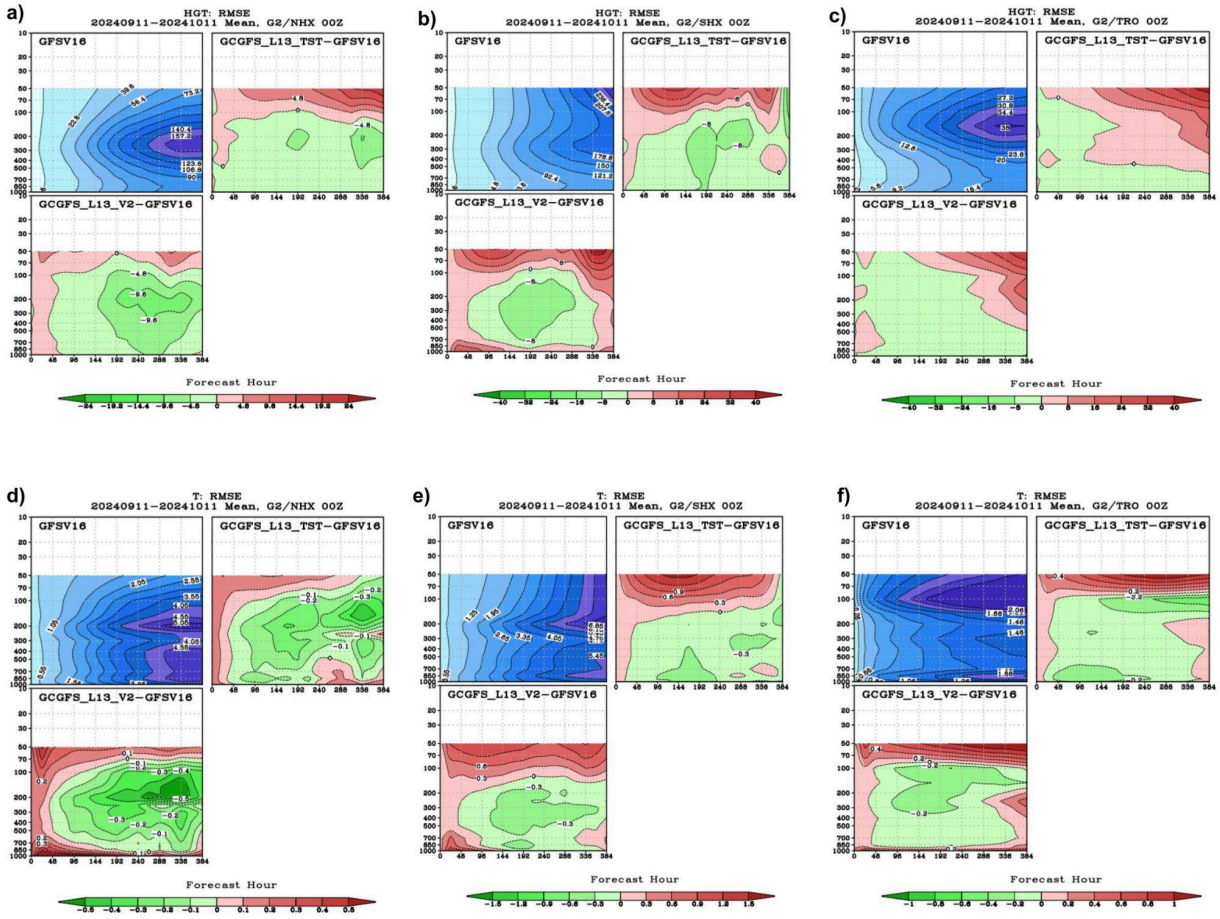

Figure 5. RMSE comparison of geopotential height on a) Northern Hemisphere, b) Southern Hemisphere, and c) Tropics and RMSE comparison of temperature on d) Northern Hemisphere, e) Southern Hemisphere, and f) Tropics. The x-axis is the forecast hour, and the y-axis is the pressure levels in hPa. In each plot of a-f), the top left is RMSE from GFSv16 against the reference, the bottom left is the RMSE difference between fine-tuned GCGFS and GFSv16, and the top right is the RMSE difference between trained GCGFS and GFSv16.

Figure 6 shows the bias comparison of the HGT and temperature fields on the northern hemisphere, the southern hemisphere, and the tropics. In general for HGT, the fine-tuned GCGFS shows a slight positive bias in the lower atmosphere, but it has a larger negative bias at the upper atmosphere and a longer lead time. The trained GCGFS starts to have negative bias earlier and from lower levels than the fine-tuned model, and it has higher negative bias at model top at long lead time. For temperature bias, the three models show different patterns. Unlike GFSv16, the fine-tuned GCGFS shows negative bias in lower atmosphere levels and positive bias in higher levels in the northern hemisphere, while it has positive bias in the near surface and negative bias in the upper atmosphere in the southern hemisphere. The trained GCGFS shows negative biases in all three regions. It shows the largest negative bias in the mid-troposphere in the northern hemisphere and the troposphere in the southern hemisphere. In summary, both models show comparable vertical structure to GFSv16 except at the top pressure layers. The trained GCGFS is slightly degraded compared to the fine-tuned model, while it has better temperature performance near the surface, which is consistent with the RMSE results.

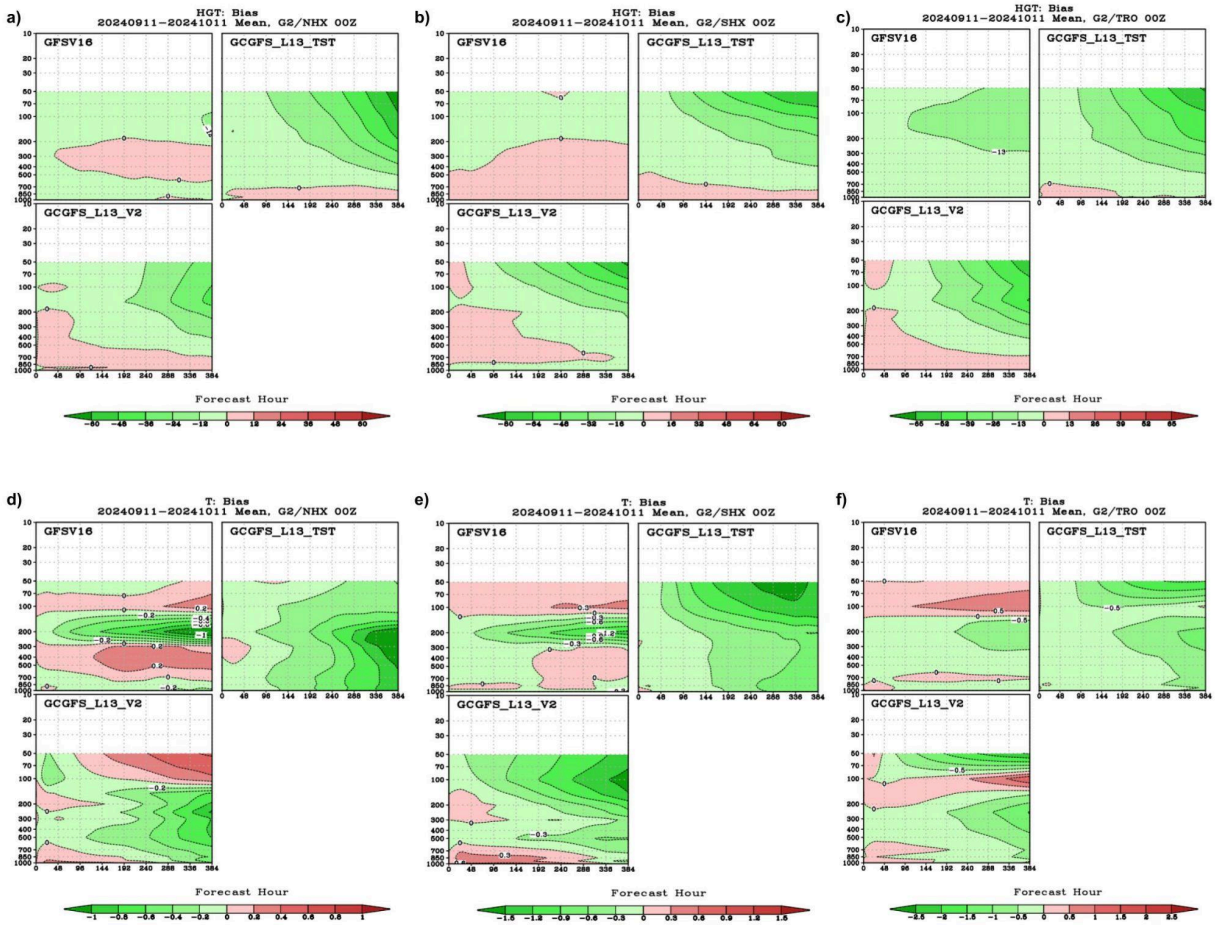

Figure 6. Bias comparison of geopotential height on a) Northern Hemisphere, b) Southern Hemisphere, and c) Tropics and bias comparison of temperature on d) Northern Hemisphere, e) Southern Hemisphere, and f) Tropics. In each plot of a-f), the top left is the bias in GFSv16, the bottom left is the bias in the fine-tuned GCGFS, and the top right is the bias in the trained GCGFS.

Figure 7 shows the 6-hourly total precipitation averaged over September 11, 2024, to October 11, 2024. It's clear that both the GCGFS models (Figs. 7c and 7d; Figs. 7k and 7j) capture the features in observations (Figs. 7a and 7i). The position and the shape of the Inter-Tropical Convergence Zone (ITCZ) from GCGFS models are found to be very close to the operational GFSv16 forecasts. It is also evident that the fields in the GCGFS models are smoother than those in GFSv16 and observations, especially at long lead time.

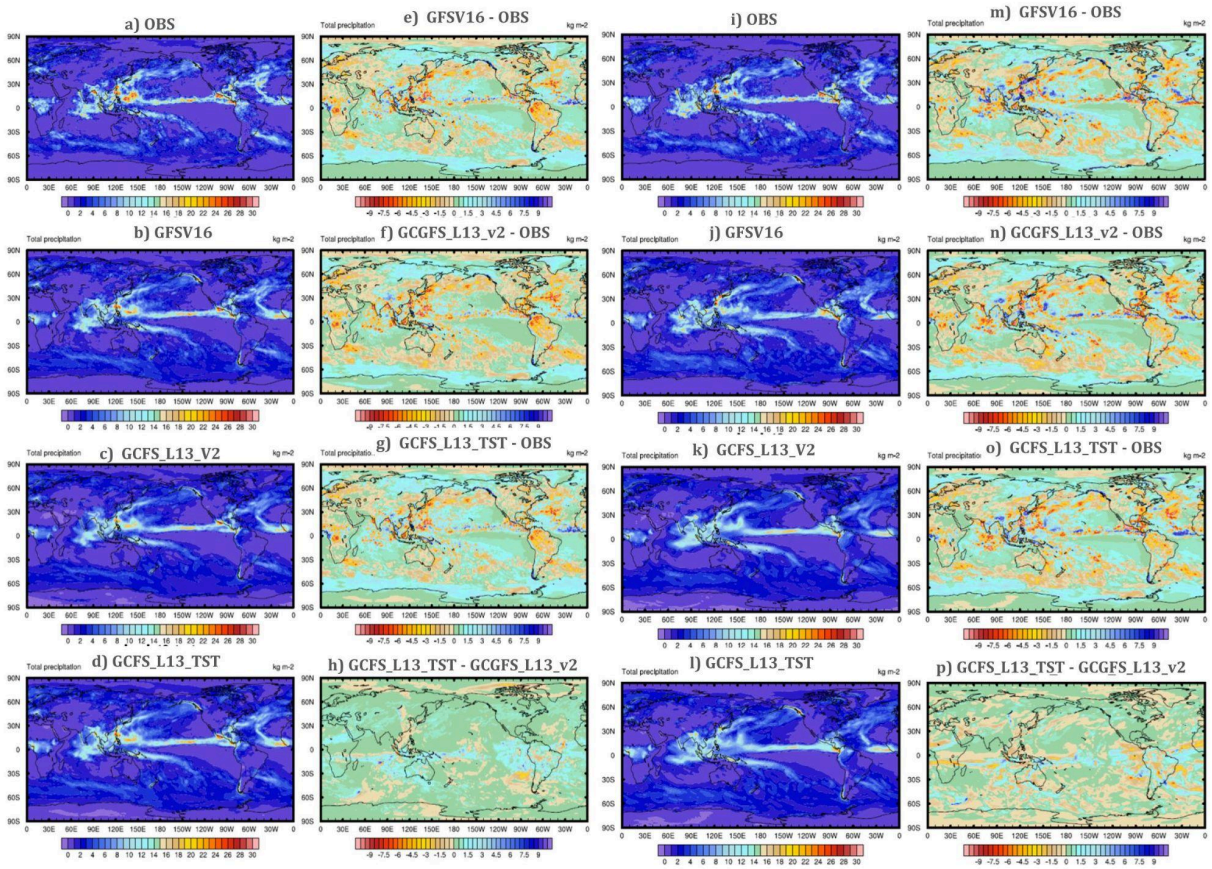

Figure 7. Global 6 hourly total precipitation at lead time 0 (0-24 hour forecast), (a-h) on left panel, and 5 day lead time (96-120 hour forecast), (i-p) on right panel. (a) and (i) represent observations; b) and j) represent GFSv16 forecasts; c) and k) represent fine-tuned GCGFS forecasts; d) and l) represent trained GCGFS forecasts; e) and m) represent the differences between GFSv16 and observations; f) and n) represent the differences between fine-tuned GCGFS and observations; g) and o) represent the differences between trained GCGFS and observations; and h) and p) represent the differences between trained GCGFS and fine-tuned GCGFS.

Figure 8 shows the time series of the 24-hour total precipitation averaged over several regions at 1-day and 5-day lead time. At the 1-day lead time, it's clear that the two GCGFS models match the GFS forecasts over the Maritime Continent (MC), Indian Ocean (IO), CONUS, and East Asia regions. Although the fine-tuned GCGFS is more in line with observations than the trained GCGFS, it tends to have the least amount of precipitation of the three models. At the 5-day lead time, the two ML models generally follow the monthly variability in the observations, and they are generally closer to observations than GFS over the tropics and globally.

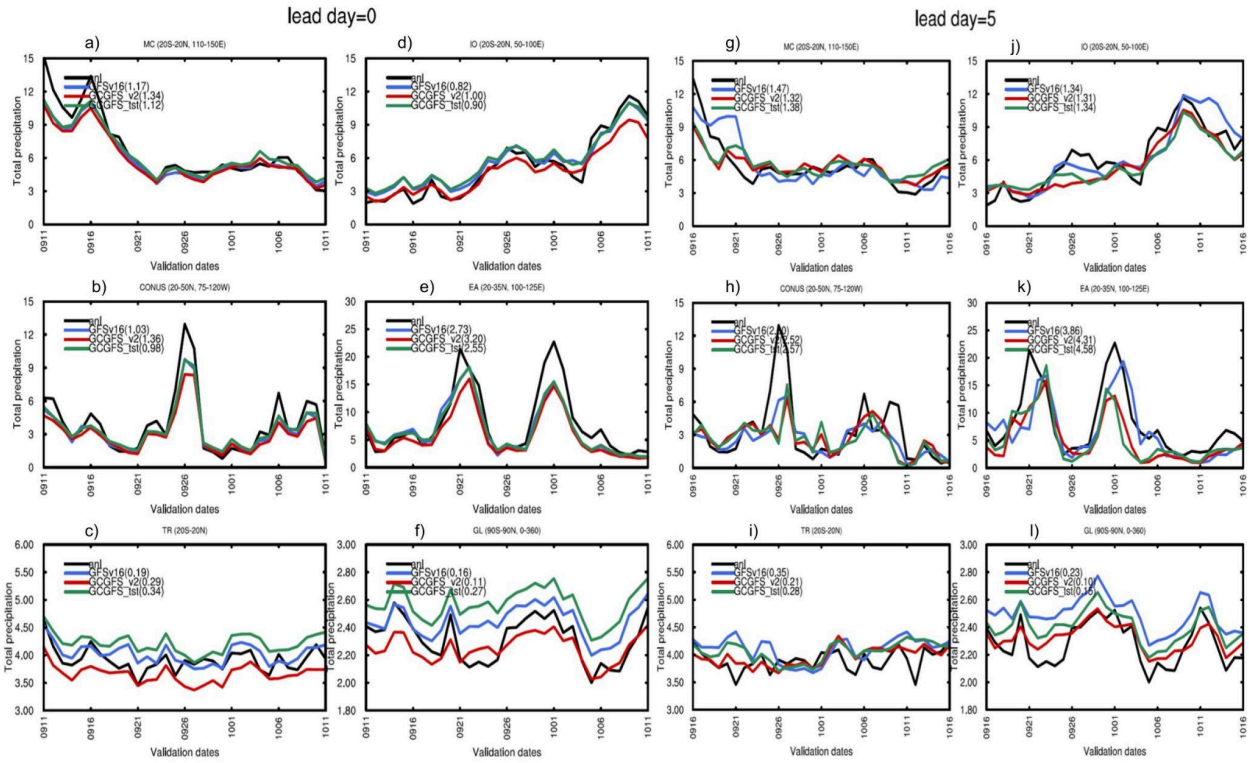

Figure 8. Time series of the total precipitation averaged over different regions for analysis, GFSv16 forecast, fine tuned GCGFS, and trained GCGFS at 1-day lead time (forecast hour 0-24) on the left panel and at 5-day lead time (forecast hour 96-120) on the right panel. a) and g) are for Maritime Continent (MC); b) and h) for CONUS region; c) and i) for tropical region; d) and i) for Indian Ocean (IO) region; e) and k) for East Asia; and f) and l) for global.

3.3 Tropical cyclones evaluations

The performance of the MLWP models in a TC case study was analyzed using the experimental real-time GCGFS outputs. Figure 9 shows a comparison of the composite tracks of the fine-tuned GCGFS and operational GFSv16 forecasts for Typhoon Gaemi, a category 4 equivalent tropic cyclone that occurred in July 2024. The fine-tuned GCGFS demonstrated improved accuracy and cycle-to-cycle consistency over GFS track forecasts, except for the track forecast at 00Z on 20240720 (track 1), which exhibited left of the track bias compared to the best track. It is worth noting that for track forecasts at 2024072006 (track 2) and 2025072012 (track 3), GFS forecasts had a right of the track bias, while GCGFS accurately predicted the landfall of Typhoon Gaemi along the northeastern coast of Taiwan. GCGFS continued performing better for the rest of forecast cycles, including after the typhoon's second landfall along the eastern China coast, where GFS incorrectly predicted the storm would move westward, while GCGFS correctly forecasted a northward direction. Overall, GCGFS demonstrated better forecast skills compared to GFS, particularly in capturing the storm's land interactions and directional shifts.

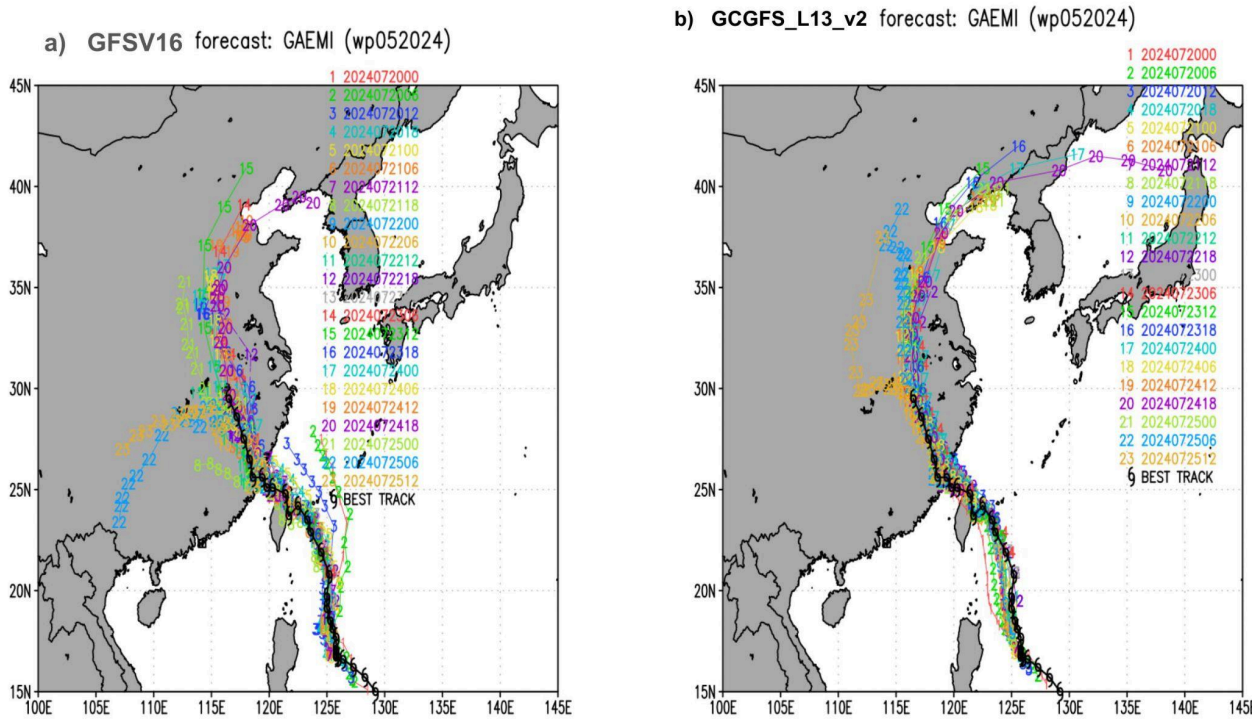

Figure 9. Composite tracks for Typhoon Gaemi in July 2024. The predicted tropical cyclone tracks from 00Z July 20, 2024, to 12Z July 25, 2024, are shown along with the best track from observations. a) is the GFSv16 forecasts, and b) is the fine-tuned GCGFS forecasts.

The forecast abilities of the GCGFS models for TC track and intensity were also assessed statistically. Figure 10 shows the mean absolute errors of TC track and intensity against TC's best track for North Atlantic Basin ((a), (b)) and North Western Pacific Basin ((c), (d)). Both GCGFS models perform similarly and better than operational GFS TC track forecasts, especially in the North Atlantic basin in TC track prediction. However, both GCGFS models show intensity degradation for all lead times. The fine-tuned GCGFS has even larger errors than the trained GCGFS. In the North Western Pacific Basin, both GCGFS models show slight track improvement but degraded intensity up to day 5. After day 5, track accuracy declines while intensity predictions improve, although the sample size is very small after day 5. In summary, both MLWP models show TC track improvement but intensity degradation in both basins. The Mean Square Error (MSE) in the training loss function is most likely the cause of this intensity degradation (Lam et al., 2023).

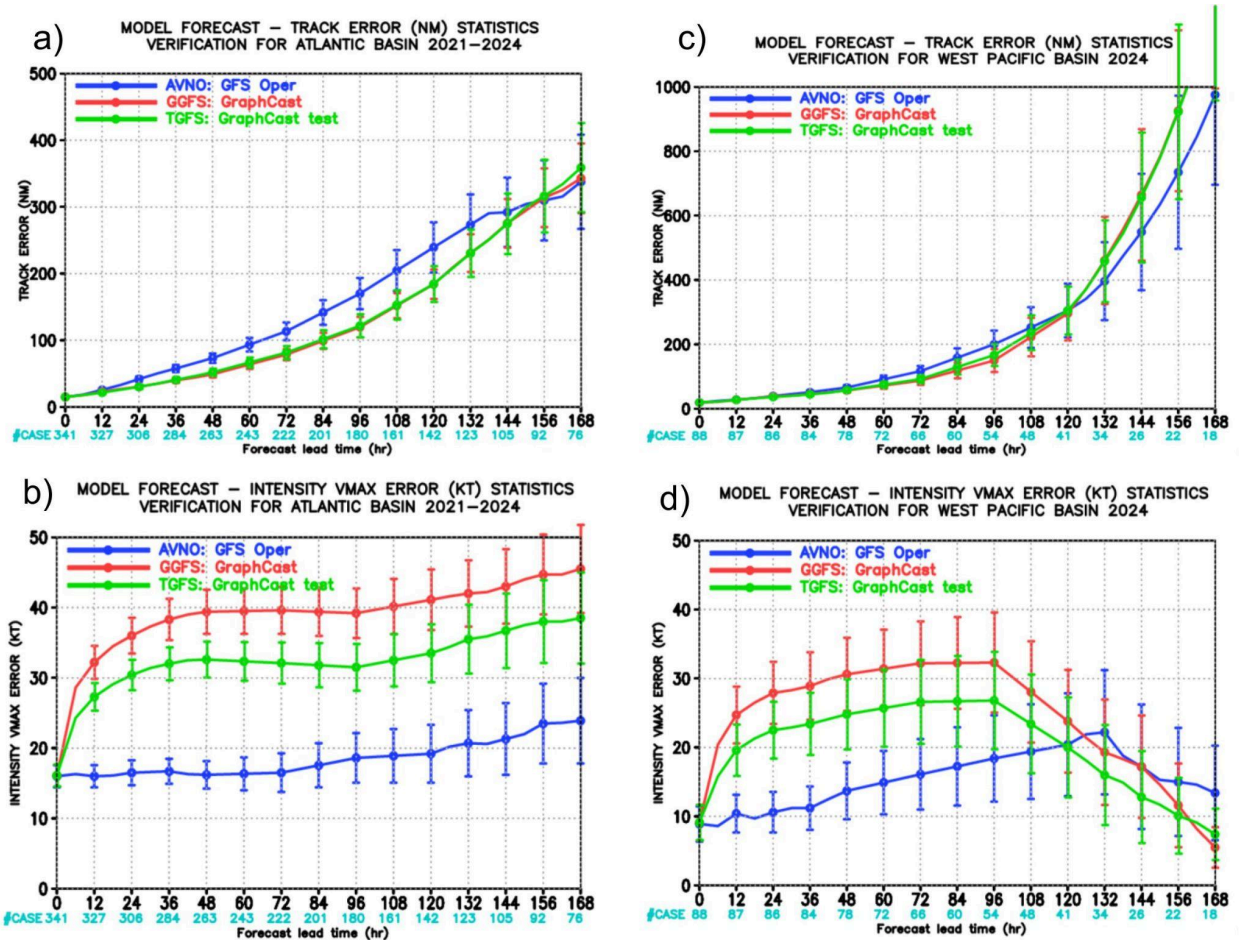

Figure 10. TC track errors in GFSv16, fine-tuned GCGFS, and trained GCGFS in a) North Atlantic and c) North Western Pacific basins. TC maximum wind errors in the b) North Atlantic and d) North Western Pacific basins.

4. Summary and Conclusions

In this paper, we present the work of integrating the data-driven weather prediction models based on GraphCast using NCEP's operational GDAS data. Specifically, we set up two versions of GCGFS based on Google DeepMind's GraphCast model with GDAS data as input, one that was fine-tuned the model with analysis data including GDAS, and another trained with GDAS analysis. The forecast results were evaluated using WeatherBench2 for 2023 model forecasts during the training and validation phase, and NCEP's operational evaluation packages for global weather and TC forecasts with experimental real-time data in 2024. Evaluation of the GCGFS forecast skills showed that both the trained and fine-tuned GCGFS models outperform the operational GFSv16 forecasts, especially at longer lead times. Both GCGFS models show reasonable vertical structure compared to operational GFS in most of the troposphere region. Even though the fine-tuned GCGFS has a larger overall improvement than the trained GCGFS, compared to the operational GFS, the fine-tuned GCGFS exhibits larger biases and errors at the model top layers and close to the surface, whereas the trained GCGFS exhibits smaller biases and errors close to the surface. Both GCGFS models show significant TC track

improvement in the North Atlantic Basin and North Western Pacific Basin at about a 5-6 day lead time while the intensity is degraded in both GCGFS models compared to the operational GFS. As for the computational resources, the GCGFS models run efficiently on a single Nvidia A100 GPU node, and in less than 4 minutes it produces 16-day global forecasts at 0.25 degree resolution for the selected 83 variables. The MLWP models, although don't solve atmospheric governing equations, can learn directly from the model states that are represented by the analysis data. We believe it opens new avenues to significantly improve weather and climate forecasts, providing accurate and accessible predictions to strengthen the breadth of weather-dependent decision-making with efficient utilization of resources.

There are several areas we are going to explore for future work. First, we will modify the loss function used in the GCGFS training. It is known that grid point MSE could lead the model to create less sharp features. Combining other evaluation criteria with the grid point MSE can help resolve the blurry issue as seen in the total precipitation forecasts and the intensity degradation in tropical cyclone forecasts. In addition, ongoing work is to update the loss weights on pressure levels to alleviate the significant error in the model top levels. Currently the weights at the top levels are close to zero, which could lead to unconstrained fields at the model top. Second, the data length of the consistent GDAS analysis is short, which has limited our training set. We will fine-tune GraphCast with more GDAS data that becomes available and also train the GraphCast with UFS replay data (NOAA, 2024) to improve the forecast skills. Third, we are going to address the forecast uncertainty as discussed in several MLWP model publications (Bi et al., 2022, and Lam et al., 2023). We also plan to develop hybrid MLWP and NWP multi-model ensembles for the global ensemble forecasting system (GEFS). The GCGFS models show different error growth characteristics than physical models, and we expect a wide range of probabilities that can be captured when training the ML models with different data sets. The hybrid ensembles can be used in conjunction with NWP models to address forecast uncertainty, increase model predictability, and decrease systematic errors in the ensemble mean.

5. Data and Code Availability Statement

The GDAS data used in this study is publicly available from NOAA S3 bucket storage (https://noaa-gfs-bdp-pds.s3.amazonaws.com/index.html). The HRES and ERA5 reanalysis data are publicly available in the Weather Bench 2 repository (https://weatherbench2.readthedocs.io/en/latest/data-guide.html). The GraphCast core model (forked from the DeepMind repository: https://github.com/google-deepmind/graphcast), as well as optimal weights and training scripts, are available at the NOAA-EMC GitHub repository (https://github.com/NOAA-EMC/graphcast). For further information, please contact Jun Wang (jun.wang@noaa.gov).

6. Acknowledgments

The Parallel Works team is acknowledged for their support in providing cloud computing resources and resolving cloud infrastructure issues. The authors would like to thank the Google DeepMind team for their constructive comments on the methodology and the Weather Bench 2

team for providing Zarr databases of HRES and ERA5 reanalysis and climatology data. The authors acknowledge the high performance computing resources provided by the NOAA Research and Development High Performance Computing Program. The project described in this article was supported by the Inflation Reduction Act as well as the NOAA Software Engineering for Novel Architectures (SENA) project. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

References

Allen, M.R., Kettleborough, J.A., and Stainforth, D.A., 2002. Model error in weather and climate forecasting. In *ECMWF Predictability of Weather and Climate Seminar* (pp. 279-304). European Centre for Medium Range Weather Forecasts, Reading, UK, http://www.ecmwf.int/publications/library/do/references/list/209.

Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. and Gulcehre, C., 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Bauer, P., Thorpe, A. and Brunet, G., 2015. The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), pp. 47-55.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q., 2022. Pangu-weather: A 3D high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y. and Li, H., 2023. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, *6*(1), p. 190.

Franklin, J. L. 2010: Tropical cyclone forecast verification at the National Hurricane Center. *20th Conference on Probability and Statistics in the Atmospheric Sciences*. 6.2. https://ams.confex.com/ams/90annual/techprogram/paper_160383.htm

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., Xue, X., Hou, D., Peng, J., Nageswararao, M.M., and Tallapragada, V., 2022. GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Monthly Weather Review*, *150*(3), pp. 647-665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T.M., Whitaker, J.S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., Sinsky, E., Zhu, Y., Tallapragada, V., Guan, H., and Zhou, X., 2022. The reanalysis for the global ensemble forecast system, version 12. *Monthly Weather Review*, *150*(1), pp. 59-79. https://doi.org/10.1175/MWR-D-21-0023.1.

Hersbach, H., and coauthors, 2020: The ERA5 global reanalysis. Q.J.R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803

Hu, Y., Chen, L., Wang, Z., and SwinVRNN, H.L., A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation., 2023, 15, p. e2022MS003211. *DOI: https://doi.org/10.1029.*

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J.,

Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society* 77(3), pp. 437-472.

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

Kleist, D., Carley, J.R., Collard, A., Liu, E., Liu, S., Martin, C.R., Thomas, C., Treadon, R., and Vernieres, G., 2023. Current State of Data Assimilation Capabilities at NCEP's Environmental Modeling Center, NOAA Office Note 514, DOI: https://doi.org/10.25923/pjs0-4j42

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., and Merose, A., Hoyer S., Holland G., Vinyals O., Stott J., Pritzel A., Mohamed S., Battaglia P., 2023. Learning skillful medium-range global weather forecasting. *Science* 382, 1416-1421 (2023). DOI:10.1126/science.adi2336

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M.C., Lessig, C., Maier-Gerber, M., Magnusson, L., and Bouallègue, Z.B., 2024. AIFS-ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*.

Lopez-Gomez, I., McGovern, A., Agrawal, S., and Hickey, J., 2023. Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, *2*(1).

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Miles, A., Kirkham, J., Durant, M., Bourbeau, J., Onalan, T., Hamman, J., & Patel, Z. (2020). zarr-developers/zarr-python: v2. 4.0. Zenodo https://doi.org/10.5281/zenodo, 4069231.

Molteni, F., Buizza, R., Palmer, T.N., and Petroliagis, T., 1996. The ECMWF ensemble prediction system: methodology and validation. *Quarterly journal of the royal meteorological society*, *122*(529), pp. 73-119.

Marchok, T., 2021: Important factors in the tracking of tropical cyclones in operational models. *Journal of Applied Meteorology and Climatology*, 60(9), DOI: https://doi.org/10.1175/JAMC-D-20-0175.1

NOAA. NOAA global forecast system (GFS) data, 2024a. URL https://registry.opendata.aws/noaa-gfs-bdp-pds.

NOAA NOAA global ensemble forecast system (GEFS) data, 2024b. URL *https://registry.opendata.aws/ noaa-gefs.*

NOAA, 2024: The Global Ensemble Forecast System (version 13) Replay dataset. NOAA Open Data Dissemination Program. Subset used: [MONTH YEAR – MONTH YEAR], accessed [DAY MONTH YEAR], https://psl.noaa.gov/data/ufs_replay/

Palmer, T.N., Shutts, G.J., Hagedorn, R., Doblas-Reyes, F.J., Jung, T. and Leutbecher, M., 2005. Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, *33*(1), pp. 163-193.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K. and Hassanzadeh, P., 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M., 2023. GenCast: diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., and Chantry, M., 2024. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), p.e2023MS004019.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., and Prudden, R., 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), pp. 672-677.

Ritchie, H., Temperton, C., Simmons, A., Hortal, M., Davies, T., Dent, D., and Hamrud, M., 1995. Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model. *Monthly Weather Review*, *123*(2), pp. 489-514.

Saha, S., Moorthi, S., Pan, H.L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., and Liu, H., 2010. The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), pp. 1015-1058, https://doi.org/10.1175/2010BAMS3001.1

Sapunov, G. (2024). Deep Learning with JAX. Manning. ISBN 9781633438880

Scher, S., and Messori, G., 2019. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, *12*(7), pp. 2797-2809.

Shafran, P., T. L. Jensen, J. H. Gotway, B. Zhou, K. Nevins, Y. Lin, and G. DiMego, 2015: Web-based verification capability using NCEP's verification database and DTC's METviewer. *27th Conf. on Hurricanes and Tropical Meteorology*, Chicago, IL, Amer. Meteor. Soc., 14A.7, http://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273777.html.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.Y., Wong, W.K., and Woo, W.C., 2017. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.Y., Wang, W., and Powers, J.G., 2008. A description of the advanced research WRF version 3. *NCAR technical note*, *475*(125), pp. 10-5065.

Sønderby, C.K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N., 2020. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*

Steinkraus, D., Buck, I., and Simard, P.Y., 2005, August. Using GPUs for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (pp. 1115-1120). IEEE.

Weyn, J.A., Durran, D.R., and Caruana, R., 2019. Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), pp. 2680-2693.

Weyn, J.A., Durran, D.R., Caruana, R., and Cresswell-Clay, N., 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7), p.e2021MS002502.

Zhou, B., and Coauthors, 2015: An overview of grid-to-grid verification at Environmental Modeling Center (EMC) of NCEP. *27th Conf. on Hurricanes and Tropical Meteorology*, Chicago, IL, Amer. Meteor. Soc., 12A.4, http://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273363.html.

Supplementary Information Section

North America:

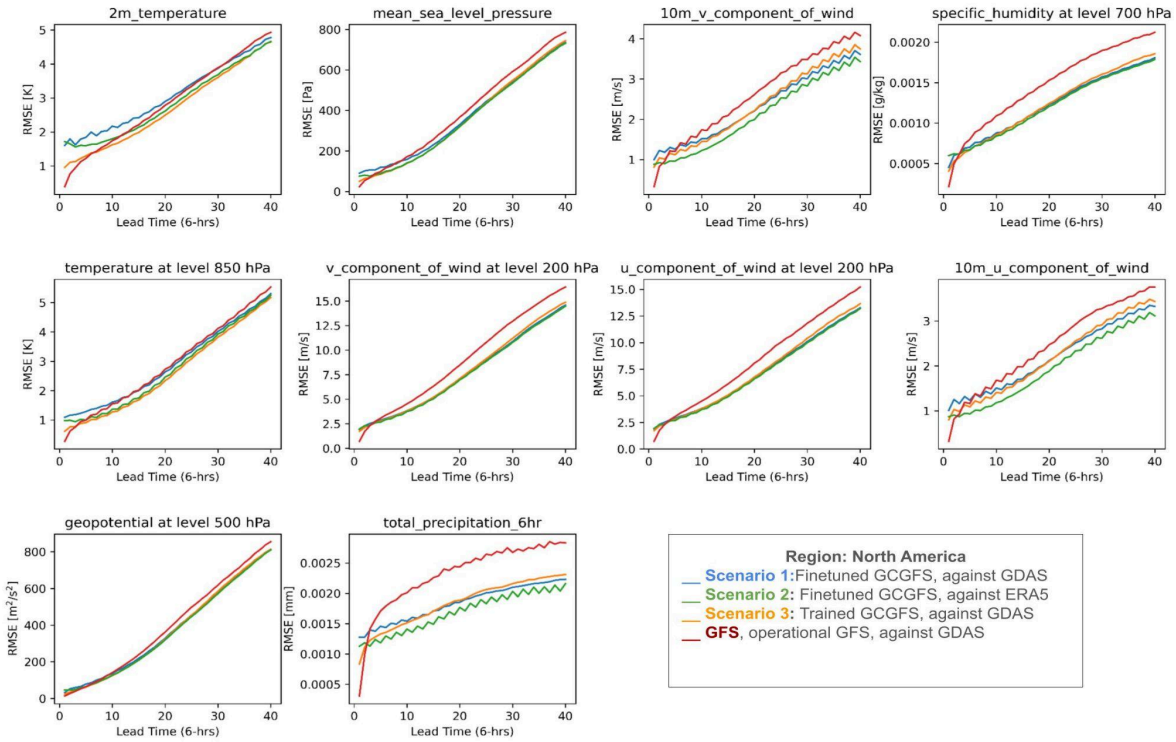

Figure 11: RMSEs of fine-tuned GCGFS (blue), trained GCGFS (orange), and fine-tuned GCGFS against ERA5 (green), and GFS (red) forecasts averaged over North America for the entire year 2023 (00z and 12z).

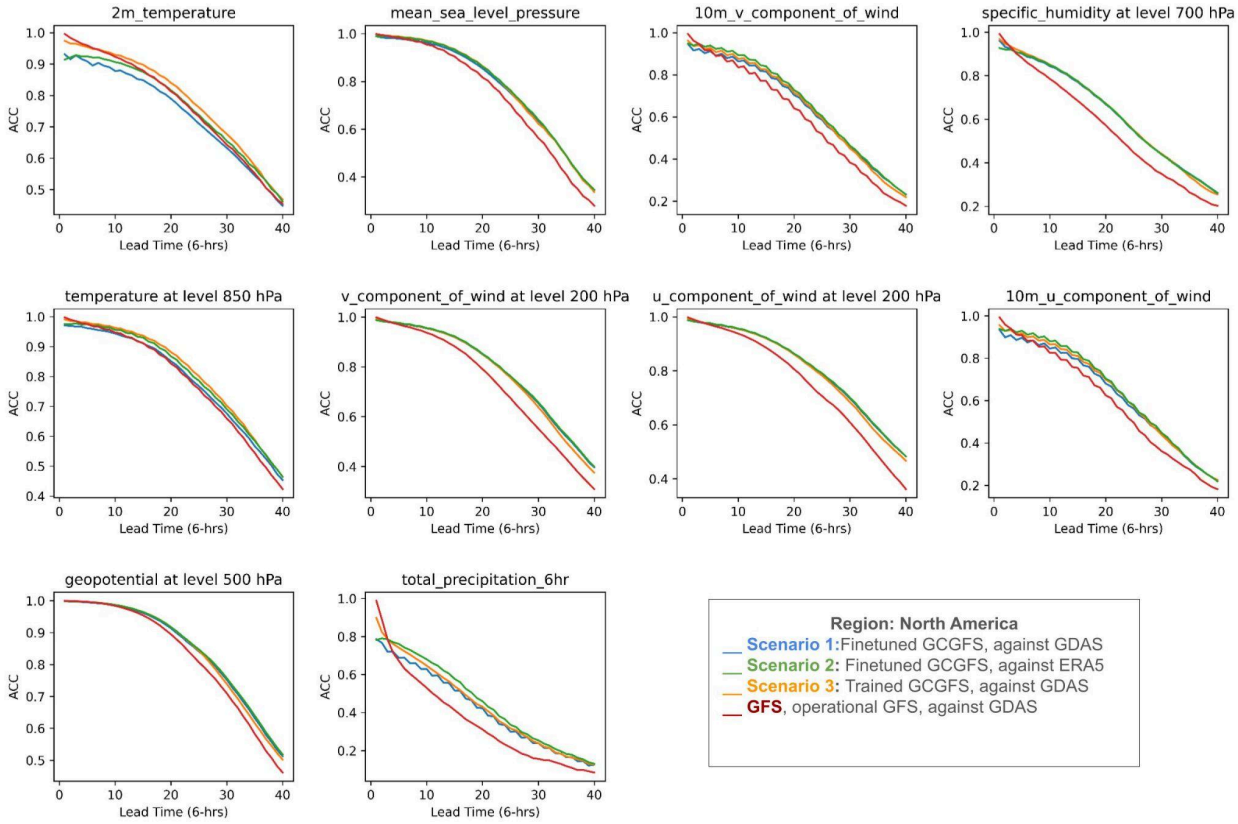

Figure 12, as Figure 11, but for ACC.

Northern Hemisphere:

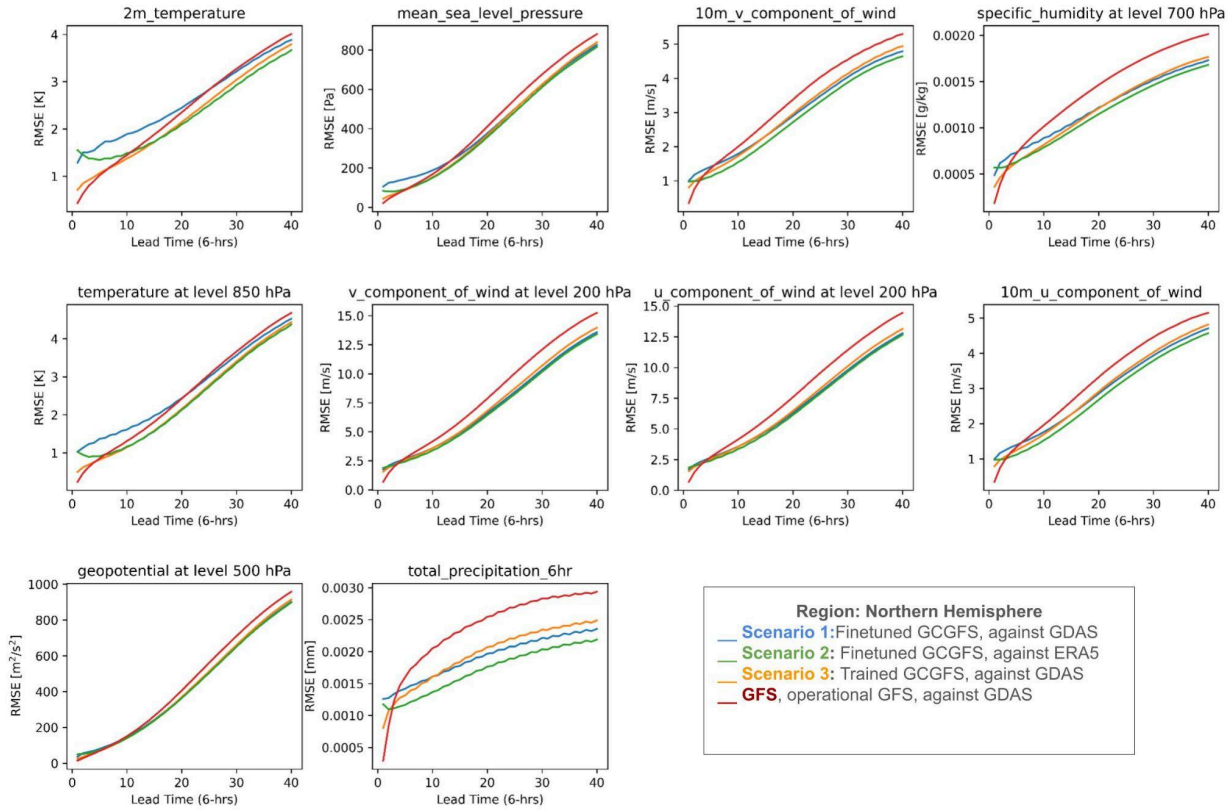

Figure 13: RMSEs of fine-tuned GCGFS (blue), trained GCGFS (orange), and fine-tuned GCGFS against ERA5 (green), and GFS (red) forecasts averaged over the Northern Hemisphere for the entire year 2023 (00z and 12z).

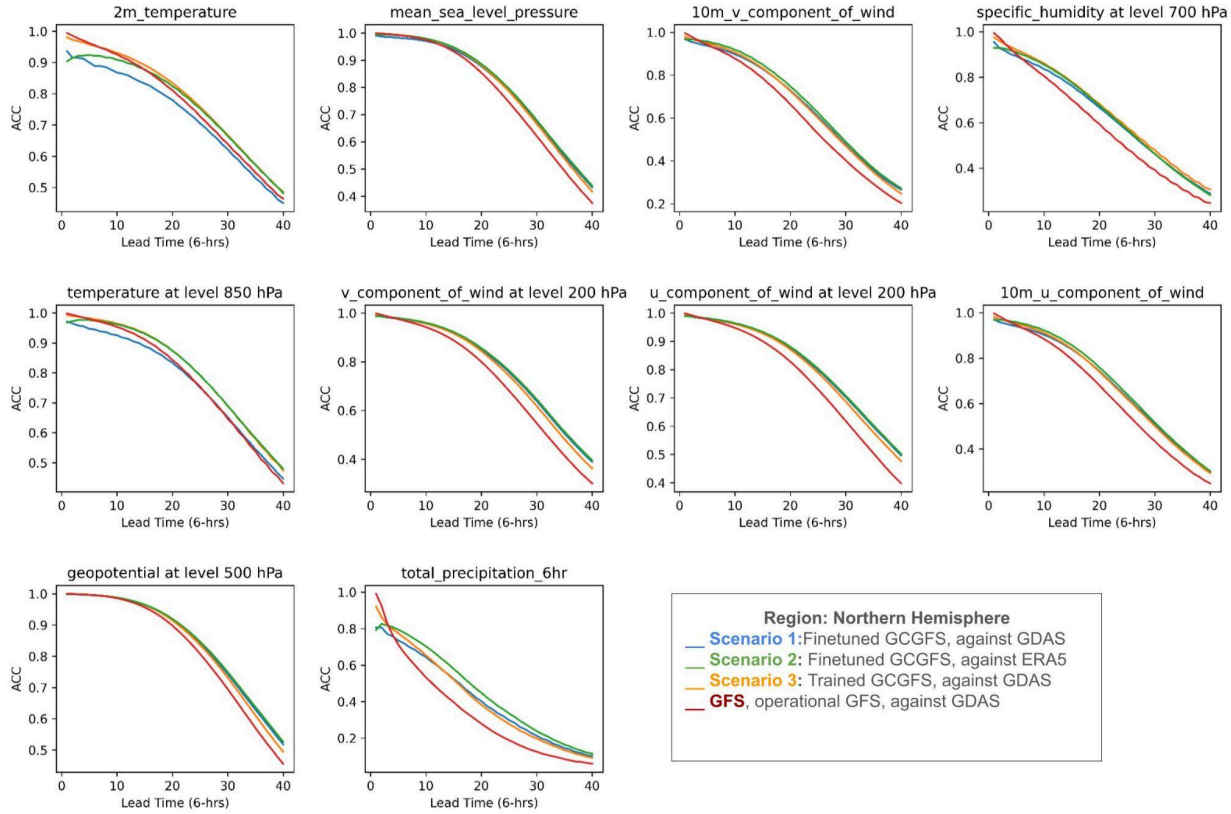

Figure 14, as in Figure 13, but for ACC.

Southern Hemisphere:

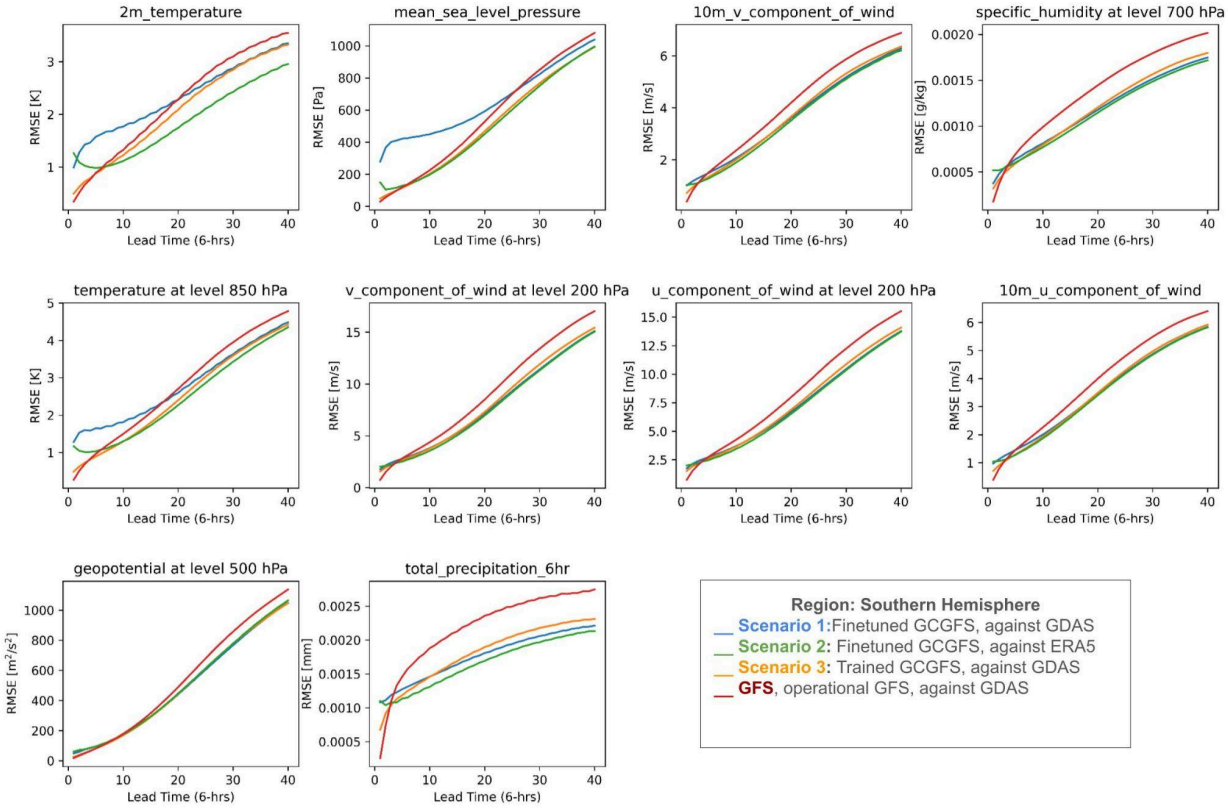

Figure 15: RMSEs of fine-tuned GCGFS (blue), trained GCGFS (orange), and fine-tuned GCGFS against ERA5 (green), and GFS (red) forecasts averaged over the Southern Hemisphere for the entire year 2023 (00z and 12z).

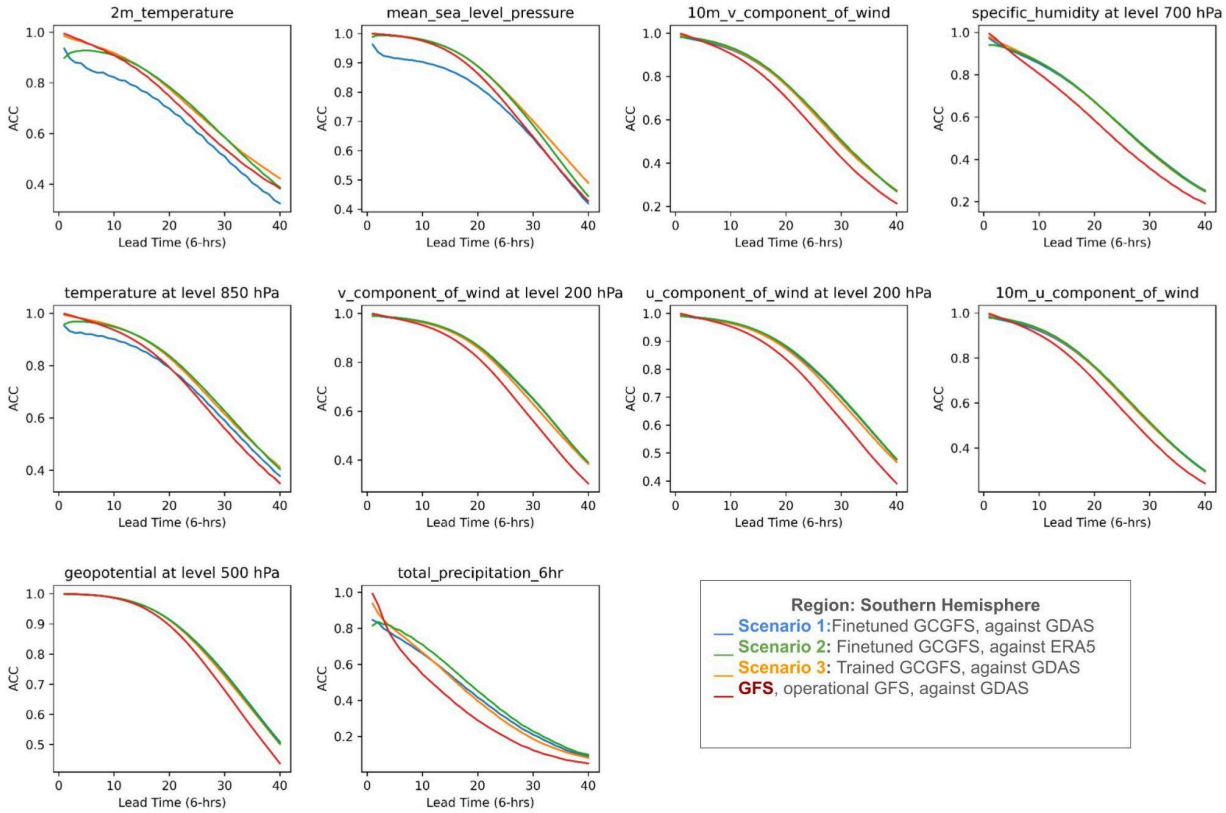

Figure 16. as in Figure 15, but for ACC.

Tropics:

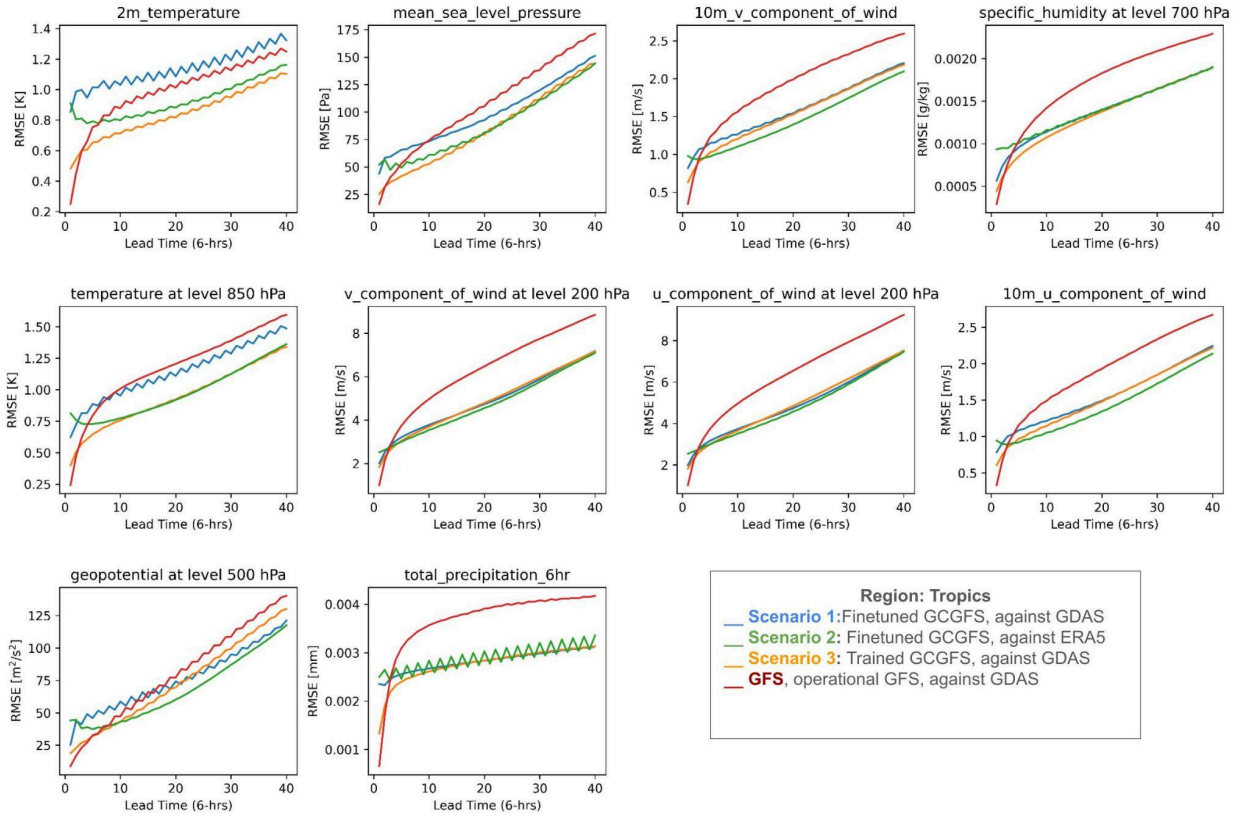

Figure 17: RMSEs of fine-tuned GCGFS (blue), trained GCGFS (orange), and fine-tuned GCGFS against ERA5 (green), and GFS (red) forecasts averaged over tropics for the entire year 2023 (00z and 12z).

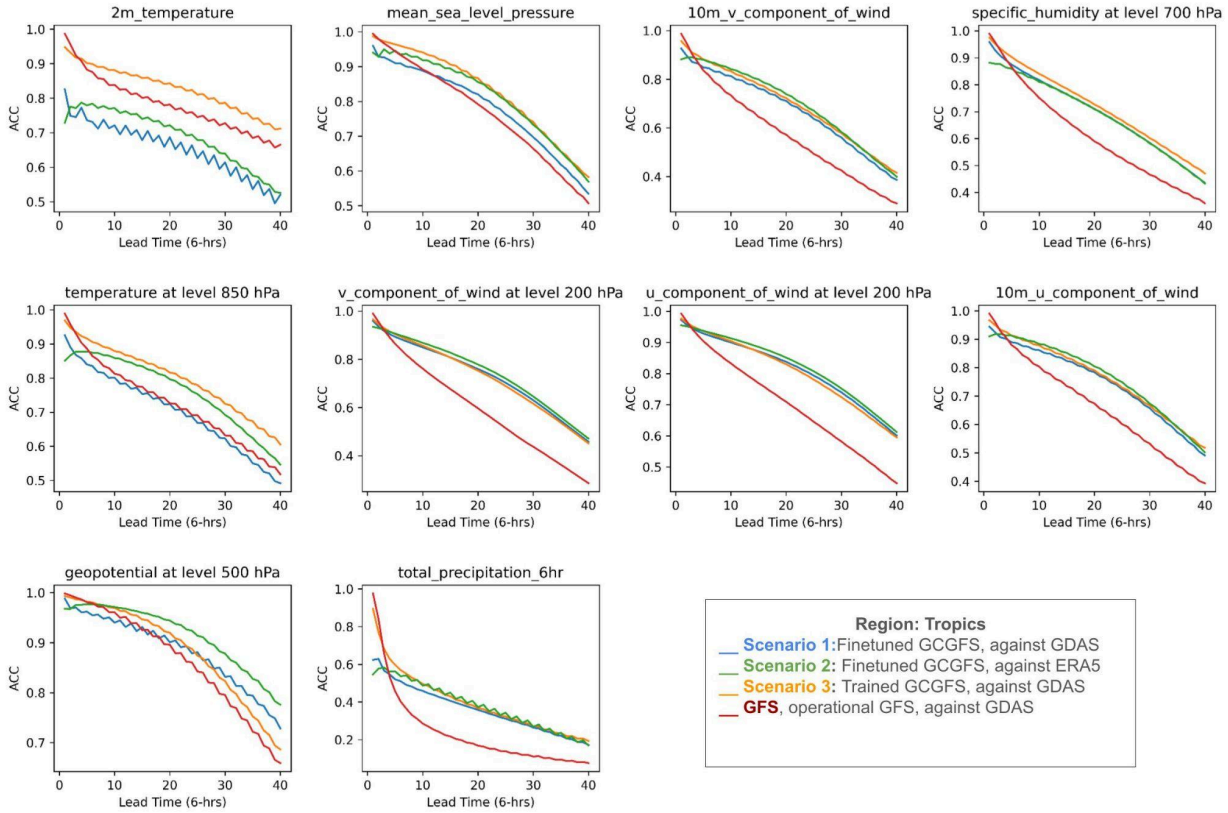

Figure 18. as in Figure 17, but for ACC.