Ensemble Prediction Approach for Improving Graph-Cast Forecast Accuracy

Pranavan Subendiran
Department of Computer Science and Engineering
University of Moratuwa
pranavans.21@uom.lk

Abstract

The "GraphCast" is a machine learning approach to predict the weather. It used Graph Neural Networks to predict the weather based on 39 years of historical data. The major limitation of this approach is that it produces deterministic forecasts, which can deviate significantly in long-term climate predictions. This study utilizes the ensembling techniques to overcome this limitation by incorporating the output of multiple models, with the addition of noise to their input. The outputs of each model instance will be combined and evaluated against the original model.

1 Introduction

GraphCast is a scalable graph-neural-network model that generates 10-day global weather forecasts in under one minute. Weather conditions are known to be uncertain, as a small change in environmental variables can significantly deviate the outcomes. To make effective decisions considering uncertainties in long-term weather conditions, probabilistic forecasts are required rather than deterministic forecasts. Traditional numerical weather forecast ensemble techniques address this limitation by tweaking the input and feeding it to multiple instances of models and combining those results to achieve probabilistic forecasting (initial-condition ensemble). But the traditional numerical forecasting models incur a high computational cost. This proposal aims to leverage GraphCast's efficiency to develop a compact initial-condition ensemble to enable probabilistic forecasts with minimal computation. This approach allows for improving the existing "GraphCast" without re-training or fine-tuning the existing model.

2 Related Work

2.1 Machine Learning-Based Weather Prediction Models

"GraphCast" leverages graph neural networks (GNNs) to predict global atmospheric dynamics up to 10 days ahead at a fine 0.25° resolution [1]. Using an encode-process-decode framework on a multi-mesh icosahedral grid, it efficiently communicates information across local and distant regions. The model encodes recent climate states and environmental data, processes them through 16 message-passing layers, and decodes predictions onto the original grid. Trained on ERA5 data (1979–2017) and tested on 2018–2021, GraphCast outperformed HRES forecasts in over 90% of targets, including more accurate tropical cyclone tracks up to five days ahead. Its limitation is producing only deterministic forecasts, making uncertainty representation challenging.

2.2 Ensemble and Perturbation Methods

Ensemble forecasting generates probabilistic forecasts by running multiple model instances with perturbed inputs, capturing both expected values and forecast uncertainty [2]. Perturbations can be designed from historical errors without retraining the model, ensuring spatially and physically consistent noise. Ensemble means help reduce systematic bias, enabling risk-aware, probabilistic forecasts from existing deterministic models.

3 Methodology

3.1 Overview

The primary objective of this study is to enhance the forecast accuracy of GraphCast without retraining or fine-tuning the model. Instead, this work explores ensemble-based aggregation and bias correction techniques as post-processing strategies. The underlying idea is that by combining multiple perturbed model outputs, it is possible to reduce random errors and systematic biases, thereby improving predictive robustness and accuracy.

3.2 Data Input

The experiments in this study utilize the ERA5 reanalysis dataset, renowned for providing high-quality atmospheric variables essential for weather prediction research. For this work, a subset of ERA5 was selected with a spatial resolution of 1 degree and 13 vertical pressure levels, focusing on a single time step to represent a single forecast instance. The variable of interest is the 2-meter temperature (2m_temperature), which serves as a key indicator of near-surface atmospheric conditions. This configuration was chosen to balance computational efficiency with sufficient atmospheric detail.

3.3 Approach

The overall process involved loading pre-trained models, running individual and ensemble forecasts, and applying aggregation strategies. The following ensemble-based post-processing techniques were implemented:

- Ensemble Mean: The average of a variable (e.g., 2m temperature) across all ensemble members. This method tends to reduce random noise and provides a smooth, stable forecast.
- Ensemble Median: The median value of the variable across all ensemble members, which mitigates the influence of extreme outliers compared to the mean.
- Trimmed Mean: A variant of the mean where a fraction of ensemble members is randomly omitted before averaging, reducing the impact of highly deviated predictions.

Additionally, small Gaussian noise was introduced to the model inputs for each ensemble member to create diversity in predictions. This simulates stochastic variability that might occur under slightly different initial conditions. The results were visualized for the 2-meter temperature field using matplotlib.

3.4 Implementation

GraphCast_small (Google DeepMind), which is a reduced version of the original GraphCast model, was chosen for its lower computational demands. Pre-trained checkpoints and sample data were obtained directly from the Google Cloud storage bucket provided by DeepMind. Experiments were conducted on a Google Colab (free tier) environment. Xarray was used for data handling. Matplotlib was used for visualizing the results.

4 Experiments and Results

4.1 Setup

The study compared the baseline GraphCast_small model with ensemble-based forecasts using mean, median, and trimmed mean aggregation. Due to resource constraints, a 10-member ensemble was run for a single prediction step, with independent Gaussian noise (10^{-6}) added to inputs. Ensemble predictions were aggregated and compared visually and quantitatively to the baseline to evaluate improvements in stability and accuracy.

4.2 Baselines

The baseline for evaluation was the original single-run prediction from the pre-trained GraphCast_small model. Comparisons were performed between the baseline outputs and ensemble-aggregated results. Visualizations of temperature distributions and difference maps were generated to illustrate the relative improvements or deviations produced by ensemble aggregation.

4.3 Evaluation Metrics

The Root Mean Square Error (RMSE) was employed as the primary metric for numerical evaluation. RMSE quantifies the deviation between the ensemble-predicted and baseline temperature fields, providing a clear measure of prediction accuracy and variance reduction.

4.4 Experimental process

Each experimental configuration consisted of an ensemble with 10 members, generated under identical conditions except for small random noise perturbations introduced into the input data. For every ensemble setup, the GraphCast_small model was used to produce forecasts for a single prediction time step. After obtaining the individual ensemble predictions, the aforementioned three aggregation techniques were applied to combine the outputs. Additional experiments were conducted by varying the magnitude of the added noise and the number of ensemble members to analyze their effects on forecast consistency. This process enabled a systematic evaluation of ensemble-based post-processing methods for improving prediction reliability without altering or retraining the underlying model.

4.5 Results

Applying ensemble predictions slightly improves the performance of the original GraphCast_small model. At worst, it maintains the same performance, and it does not lead to any degradation.

4.5.1 Quantitative Results

Table 4.1 presents the RMSE values for each approach across the selected variables. In all cases, specific humidity remains unchanged. However, both 2m temperature and 10m wind components show improved performance through the use of ensemble methods. Among the ensemble aggregation techniques, the Ensemble Mean and Ensemble Median perform at comparable levels, while the Trimmed Mean demonstrates a slight additional improvement over the other two methods.

Method	2m_temperature	specific_humidity	10m_u_component_of_wind	10m_v_component_of_wind
Single GraphCast_small	0.5693	0.0002	0.6312	0.6416
Ensemble Mean	0.5688	0.0002	0.6310	0.6410
Ensemble Median	0.5689	0.0002	0.6310	0.6410
Trimmed Mean	0.5689	0.0002	0.6309	0.6409

Table 4.1: RMSE for each variable comparing the baseline single model with 10-member ensemble forecasts

4.5.2 Comparing Plots

The predicted outputs from each approach were compared against the ground truth and visualized by plotting the targets, predictions, and their differences.
Figure 4.1 illustrates the target and predicted values from a single model, along with the differences between the original and predicted outputs obtained from a single model.

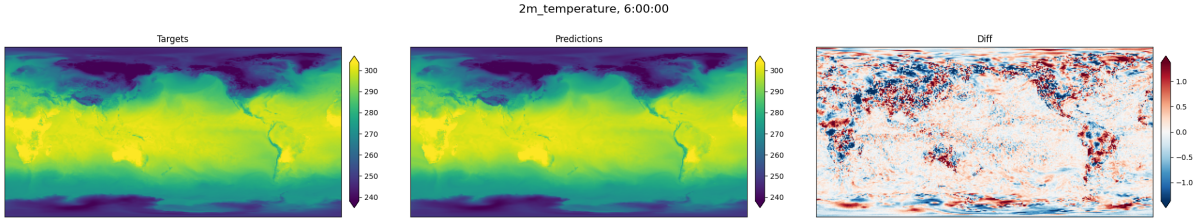

Figure 4.1: Single Model Prediction

Figure 4.2 shows the target values, ensemble mean predictions, and the deviation between the original and ensemble mean results.

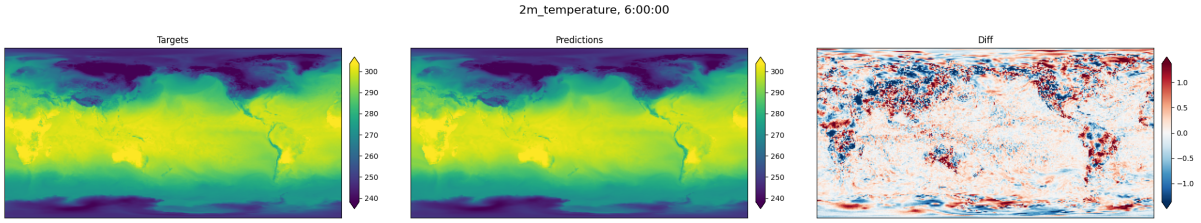

Figure 4.2: Ensemble Mean Prediction

5 Discussion

Results demonstrate that, without any retraining and fine-tuning, it is possible to marginally improve the predictive performance of the GraphCast_small model using ensemble-based post-processing strategies. For each variable of interest, the ensemble methods enhanced or maintained the baseline model's performance, which demonstrates their robustness and reliability.
For the 2m temperature and 10m wind components, ensemble aggregation techniques reduced the RMSE values from the base GraphCast_small outputs, indicating ensemble averaging effectively removes random prediction noise and improves forecast validity. Otherwise, the specific humidity variable remained unchanged by all methodologies, indicating its prediction uncertainty is low or less sensitive to ensemble averaging.
Out of the ensemble methods that were experimented with, Ensemble Mean and Ensemble Median worked very similarly to one another, but there was a small added improvement with the Trimmed Mean method. This minor improvement shows that the elimination of outlier predictions (as is done in trimmed mean averaging) can add further stability to the prediction and remove minor biases.
Overall, these findings confirm that ensemble aggregation methods can enhance the performance of the baseline GraphCast_small model. This confirms the potential of ensemble-based post-processing as a viable means of enhancing weather forecasting systems without retraining the models.

6 Conclusion

This study demonstrates that ensemble-based post-processing methods can improve the predictive accuracy and robustness of the GraphCast_small model without the need for retraining or fine-tuning. By introducing small perturbations to the input data and aggregating multiple model outputs through mean, median, and trimmed mean approaches, it was possible to slightly reduce random noise and systematic bias in the forecasts. The results showed a consistent reduction in RMSE for key atmospheric variables, validating the benefit of ensemble aggregation in stabilizing predictions. Among the methods explored, the trimmed mean achieved marginally superior results, highlighting the potential of outlier exclusion to improve ensemble stability. Overall, these findings confirm that ensemble-based strategies are a practical, less computationally intensive way to extend deterministic machine learning weather models like GraphCast_small toward more reliable probabilistic forecasting.

7 Future Works

The current study demonstrates the feasibility of enhancing GraphCast forecasts through ensemble-based post-processing. However, several avenues remain for future exploration to further improve ensemble reliability and predictive accuracy:

1. Multi-step Forecast Evaluation: The present experiments are limited to single-step forecasting. Future work should extend this analysis to multi-step and longer lead-time forecasts to evaluate how ensemble aggregation performs over extended temporal horizons, especially in long-term climate predictions.

2. Dynamic Noise Modeling: In this study, a fixed Gaussian noise magnitude of 10^{-6} was used to perturb the inputs. Future experiments could incorporate adaptive noise modeling techniques that leverage spatiotemporal correlations from historical forecast errors. Such dynamic perturbation schemes would yield more physically consistent and realistic ensemble diversity.

3. Larger Ensemble Configurations: Due to computational constraints, the current setup utilized only ten ensemble members. Increasing the ensemble size in future work could provide a more comprehensive representation of forecast uncertainty. It will also enable a systematic analysis of the trade-off between computational cost and accuracy improvement.

8 References

[1] R. Lam, A. Sanchez-Gonzalez, M. Willson, *et al.*, "Learning skillful medium-range global weather forecasting," *Science (New York, N.Y.)*, vol. 382, eadi2336, Nov. 2023. DOI: 10.1126/science.adi2336.

[2] M. Leutbecher and T. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008, Predicting weather, climate and extreme events, ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2007.02.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999107000812.