Ensemble Prediction Approach for Improving Graph-Cast Forecast Accuracy

1st Pranavan Subendiran
Department of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
pranavans.21@uom.lk

2nd Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
rtuthaya@cse.mrt.ac.lk

[1]

***Abstract*—Machine learning (ML) has transformed weather forecasting by modeling complex atmospheric processes from historical data. GraphCast, developed by Google DeepMind, uses Graph Neural Networks (GNNs) to deliver global forecasts with high spatial resolution. However, its deterministic nature limits reliability under rapidly changing atmospheric conditions. This study introduces an ensemble-based post-processing framework to improve GraphCast's accuracy and robustness without retraining. Multiple forecasts are generated by adding Gaussian noise to inputs. Output of the ensemble is collected and aggregated using mean, median, and trimmed mean methods. A linear regression-based bias correction further refines predictions. Using the GraphCast_small model trained on ERA5 reanalysis data, performance is evaluated using Root Mean Square Error (RMSE) for key atmospheric variables such as 2-meter temperature and 10-meter wind speed components. Results show that ensemble aggregation consistently reduces RMSE, highlighting ensemble post-processing as a computationally efficient enhancement to deterministic ML weather forecasting.**

***Index Terms*—Ensemble, Weather Forecasting, GraphCast, ERA5, GraphCast_small**

I. Introduction

Traditional numerical weather prediction (NWP) models solve complex partial differential equations that govern atmospheric dynamics. However, these models are computationally expensive and often limited by the accuracy of their initial conditions. The evolution of machine learning (ML) has opened new pathways for efficient weather forecasting, offering improved accuracy and robustness to noise in initial atmospheric states compared to traditional NWP approaches.

Recent advances in deep learning have produced data-driven forecasting systems such as FourCastNet [3], Pangu-Weather [4], and ClimaX [5]. These models leverage high-capacity neural operators and transformer-based architectures to capture multi-scale atmospheric interactions without relying on explicit physical parameterizations. Among these, GraphCast [1] is one of the most advanced ML-based forecasting systems. It employs Graph Neural Networks (GNNs) to simulate the evolution of atmospheric states using gridded ERA5 reanalysis data, providing 10-day forecasts at a fine spatial resolution of $0.25°$. Due to the resource availability, conducting experiments with the GraphCast model was challenging. Therefore, the GraphCast_small model, which is a scaled-down version of the GraphCast model with a resolution of $1°$ instead of $0.25°$, and the same encode-process-decode architecture was used to conduct the ensemble experiments.

Weather conditions are inherently uncertain, so small perturbations in environmental variables can lead to significant deviations in predicted outcomes, especially over longer forecast horizons. To address such uncertainty, ensemble forecasting has been widely adopted in NWP to generate probabilistic forecasts [2], [6]. By introducing controlled perturbations to model inputs and combining multiple outputs, ensemble methods capture the spread of possible atmospheric evolutions, yielding more stable and reliable forecasts. Traditional NWP ensembles achieve this through repeated model runs with varied initial conditions, but such methods are computationally intensive at high resolutions.

This study aims to extend GraphCast into an ensemble forecasting framework without requiring model retraining. The proposed approach introduces ensemble perturbations and aggregation techniques, including the ensemble mean, median, and trimmed mean. By applying small Gaussian perturbations to the model's input space, the method simulates stochastic variability while maintaining computational efficiency. Furthermore, a bias correction step is applied to reduce systematic deviations between model predictions and ground-truth observations.

Finally, the magnitude of added Gaussian Noise and the number of members in the ensemble were varied to observe their impact on the produced results. The results are evaluated quantitatively using Root Mean Square Error (RMSE) metrics and qualitatively through difference maps.

II. Related Work

A. Machine Learning-Based Weather Prediction Models

Recent years have seen the emergence of powerful deep-learning architectures for global weather prediction. FourCastNet [3] employs Adaptive Fourier Neural Operators to model medium-range weather dynamics and has achieved skillful forecasts comparable to the ECMWF's IFS model. Pangu-Weather [4], developed by Huawei, utilizes a 3D transformer-

[1]Source code is available at: https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210491P-Climate-AI_Climate-Modeling

based architecture to deliver 3D high-resolution forecasts, demonstrating superior accuracy in medium-range predictions. Similarly, ClimaX [5] introduces a foundation model that unifies diverse meteorological tasks through transfer learning, providing a scalable framework for weather and climate modeling.

"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.

B. 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, ensuring spatially and physically consistent noise. Ensemble means help reduce systematic bias from existing deterministic models.

Recent studies have also explored neural network–based post-processing to refine ensemble forecasts. Rasp et.al [6] proposed neural network post-processing for bias and spread correction of ensemble weather forecasts, achieving significant improvements in calibration and sharpness. These approaches demonstrate the potential of hybrid statistical–machine learning ensembles to enhance deterministic models such as GraphCast.

III. Methodology

A. Overview

The primary objective of this study is to enhance the forecast accuracy of GraphCast_small without retraining or fine-tuning the model. 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 accuracy and robustness. Further, this study analyzes the prediction accuracy of the ensemble by changing the ensemble size and the magnitude of the noise added.

B. Data Configuration

The experiments in this study utilize the ERA5 reanalysis dataset, a widely recognized and high-quality source of atmospheric data produced by the European Center for Medium-Range Weather Forecasts (ECMWF). ERA5 provides comprehensive global coverage of essential atmospheric, land, and oceanic variables, making it a standard benchmark for weather prediction research.

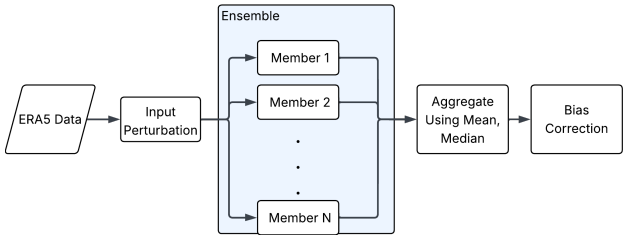

Fig. 1. Methodology Overview

For this study, a subset of ERA5 data was employed with a spatial resolution of $1^\circ \times 1^\circ$ and 13 vertical pressure levels. The experiments focus on a single time step, representing an individual forecast instance to evaluate the short-term predictive capability of the proposed framework. The variables of interest include:

- 2-meter temperature (2m_temperature): a key indicator of near-surface thermal conditions.
- 10-meter zonal wind component (10m_u_component_of_wind): representing the east-west wind flow.
- 10-meter meridional wind component (10m_v_component_of_wind): representing the north-south wind flow.

This configuration provides a balance between computational efficiency and atmospheric representativeness, allowing the experiments to capture essential surface-level dynamics while maintaining manageable data volume and processing requirements.

C. Model Configuration

The GraphCast_small model (from Google DeepMind), a lightweight variant of the original GraphCast, was selected to accommodate the computational limitations of the Google Colab environment. This model maintains the core GNN architecture but operates at lower resolution ($1^\circ \times 1^\circ$).

D. Ensemble-Based Approach

To introduce uncertainty into GraphCast_small's deterministic forecasts, ensemble members were generated by applying Gaussian perturbations to the model inputs. For each ensemble run, independent random noise with mean 0 and a specified standard deviation was added to all input fields. The perturbation magnitude controls ensemble diversity while preserving physically consistent inputs.

Each ensemble member produced an independent forecast using the pre-trained model. These forecasts were then aggregated using three statistical techniques:

- 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.

To correct small systematic deviations from observed data, a bias-correction step was applied to the ensemble mean predictions. A simple linear regression $y = a \times x + b$ was fitted between ensemble means (x) and target observations (y), and the resulting coefficients were used to generate bias-adjusted forecasts.

Furthermore, a sensitivity analysis was conducted to understand how different experimental settings influence the forecasting performance. This analysis examined the effect of two key factors:

1) The ensemble size.
2) The magnitude of the added Gaussian noise.

By systematically varying the number of ensemble members and the strength of the input perturbations, the study evaluated how these parameters impact the accuracy of the ensemble forecasts. The results from this analysis helped identify the optimal balance between computational cost and forecast reliability.

E. Experimental Setup

All experiments were conducted using the Google Colab environment to ensure accessibility and computational efficiency. The implementation utilized JAX and Haiku for model execution, Xarray for data handling, NumPy for numerical computation, and Matplotlib, along with Cartopy for data visualization and spatial mapping.

The GraphCast_small model and sample ERA5 data were obtained directly from the official DeepMind GraphCast cloud repository [7], [8]. Each experiment was executed under identical runtime conditions with fixed random seeds to ensure reproducibility and consistent noise perturbations across ensemble members.

IV. Experiments and Results

A. Experiment Design

The experiments were designed to comprehensively evaluate the effectiveness of ensemble-based post-processing in improving the predictive accuracy and robustness of the GraphCast model. The study was structured into ensemble aggregation analysis and sensitivity analysis, each targeting a specific aspect of the ensemble framework's performance.

In the first phase, the analysis of ensemble aggregation methods utilized an ensemble of 10 members. Each ensemble member received slightly perturbed input data, where Gaussian noise with a standard deviation of 10^{-7} was added to simulate uncertainty in the atmospheric conditions. This controlled perturbation ensured that the ensemble captured small variations in the model's input space without significantly altering the overall physical coherence of the data. Three aggregation techniques, namely ensemble mean, ensemble median, and trimmed mean, were applied to the ensemble outputs to assess their effectiveness in enhancing forecast stability, suppressing random variability, and mitigating the impact of outlier predictions, helping to produce smoother and more reliable atmospheric forecasts.

The second phase, the sensitivity analysis, systematically examined how variations in ensemble size and noise magnitude influence forecast performance. To assess the effect of ensemble size, ensembles containing 2, 5, 10, and 15 members were generated. Each configuration followed the same input perturbation procedure, and the forecasts were aggregated using the ensemble mean. By comparing the resulting outputs, the study aimed to identify an optimal ensemble size that balances computational efficiency with forecast accuracy and stability.

To further investigate the impact of noise magnitude, Gaussian perturbations with standard deviation values of 10^{-1}, 10^{-3}, 10^{-5}, 10^{-6}, 10^{-7}, and 10^{-8} were applied to the model inputs. For each noise level, an ensemble of 10 members was generated, and their predictions were combined using the ensemble mean. This step enabled a systematic evaluation of how different levels of input uncertainty influence the ensemble's predictive skill and consistency.

All experimental configurations were benchmarked against a single deterministic baseline forecast produced by the original GraphCast_small model. The comparison provided quantitative insights into the extent to which ensemble-based post-processing improves forecast performance relative to the deterministic approach. Overall, these experiments were carefully structured to uncover the relationships between ensemble diversity, noise intensity, and predictive reliability, offering a deeper understanding of how ensemble strategies can enhance machine-learning-based weather forecasting systems.

B. Baseline

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.

C. Evaluation Metrics

The Root Mean Square Error (RMSE) was employed as the primary metric to evaluate forecast accuracy and reliability. RMSE measures the average magnitude of prediction errors by comparing model forecasts with observed ground-truth values. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where y_i denotes the true observations, $\hat{y}_i$ represents the predicted values, and N is the total number of samples. In this

study, RMSE quantifies the deviation between ensemble-based forecasts and the baseline deterministic outputs produced by GraphCast_small. Lower RMSE values indicate higher forecast accuracy and reduced variability, making it a suitable measure to assess the effectiveness of the ensemble post-processing approach across different configurations.

D. Ensemble Aggregation Results

The ensemble aggregation experiments evaluated how different statistical methods affect the stability and reliability of forecasts produced by the GraphCast_small model. Outputs from 10 independently perturbed ensemble members were combined using the ensemble mean, median, and trimmed mean techniques. These aggregated predictions were compared against the single deterministic baseline forecast.

TABLE I
RMSE COMPARISON OF DIFFERENT ENSEMBLE AGGREGATION TECHNIQUES (ENSEMBLE MEAN, MEDIAN, TRIMMED MEAN)

Var	Orig.	Mean	Med.	Trim.
2m_temperature	0.568011	0.567846	0.567846	0.567866
10m_u_comp.wind	0.630723	0.630629	0.630644	0.630643
10m_v_comp.wind	0.640502	0.640381	0.640423	0.640406
specific_humidity	0.000225	0.000225	0.000225	0.000225
temperature	0.442397	0.442735	0.442612	0.442755

The ensemble aggregation methods consistently improved the smoothness and coherence of forecasted atmospheric fields, particularly for the 2-meter temperature and 10-meter wind components. As shown in Table I, all three aggregation methods slightly reduced the RMSE compared to the baseline model, indicating a reduction in random variability. The trimmed mean yielded the lowest RMSE values across all evaluated variables, suggesting that removing a small fraction of outlier members enhances ensemble robustness.

The visualization 2m_temperature from the corresponding table is shown in Figure 2

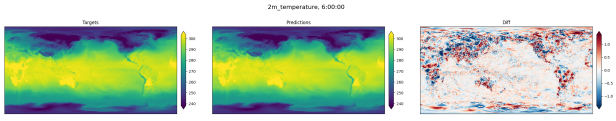

Fig. 2. Ensemble Mean and target predictions along with the difference

Although the numerical improvements may appear modest, they were consistent and systematic across all experiments. The ensemble-based forecasts exhibited smoother spatial patterns and reduced localized noise, demonstrating the effectiveness of aggregation as a post-processing enhancement for deterministic ML forecasts.

Overall, these results indicate that ensemble aggregation is a reliable and efficient post-processing strategy for mitigating unpredictable fluctuations in deterministic machine learning predictions, thereby enhancing their overall stability, reliability, and interpretability.

E. Bias Correction Results

A bias-correction step was performed on the ensemble-mean forecasts to address persistent systematic deviations from observed data. A simple linear regression model was fitted between the ensemble predictions and the ground-truth observations to estimate correction coefficients, expressed as:

$$y = a \times x + b$$

where y represents the corrected forecast, x the ensemble mean, and a and b the regression coefficients.

TABLE II
RMSE COMPARISON OF BASELINE, ENSEMBLE MEAN, AND BIAS-CORRECTED FORECASTS

Var.	Orig.	Ens. Mean	Bias Corr.
2m_temperature	0.567974	0.567814	0.566055
10m_u_comp.wind	0.630767	0.630644	0.630066
10m_v_comp.wind	0.640420	0.640355	0.640153
specific_humidity	0.000225	0.000225	0.000225
temperature	0.442355	0.442752	14.338505

After applying bias correction, minor but consistent improvements were observed in the RMSE scores as shown in Table II.

The variables 2m_temperature, 10m_u_comp.wind, and 10m_v_comp.wind show slight but consistent improvements after applying ensemble mean and bias correction. In contrast, the temperature variable exhibits a significant increase in error compared to the ensemble mean, which indicates bias correction may have introduced overadjustments for this specific field.

Though small in scale, the improvement trend was uniform across all variables other than the temperature variable. The bias-corrected ensemble forecasts showed better calibration and reduced systematic offsets, confirming that simple statistical adjustments can complement ensemble aggregation for higher forecast reliability.

F. Effect of Noise Magnitude

The sensitivity of the ensemble system to the magnitude of Gaussian perturbations was examined to understand how the strength of input noise influences predictive accuracy.

In ensemble forecasting, the amplitude of noise applied to the model inputs directly determines how widely the ensemble members diverge from one another. Very small perturbations tend to produce nearly identical forecasts, offering limited information about uncertainty. Conversely, excessively large perturbations can distort the physical consistency of the inputs, causing unrealistic deviations in the resulting predictions. Achieving an appropriate balance is therefore essential to preserve both diversity and reliability within the ensemble.

To explore this relationship, a sequence of experiments was carried out by varying the standard deviation of the Gaussian noise applied to the model inputs while maintaining a fixed ensemble size of 10 members.

For each configuration, the perturbed inputs were processed independently through the pre-trained GraphCast_small

model, and the resulting forecasts were aggregated using the ensemble-mean technique. The Root Mean Square Error (RMSE) was then calculated for each atmospheric variable to quantify how forecast accuracy varied with the level of perturbation. The results are shown in Table III.

TABLE III
RMSE COMPARISON OF FORECASTS UNDER DIFFERENT NOISE MAGNITUDES

Noise	2m_temp.	10m_u_comp.	10m_v_comp.
Orig. Pred.	0.567974	0.630767	0.640420
Noise 10^{-1}	3.632622	3.691192	2.915304
Noise 10^{-3}	3.392228	3.552326	2.715450
Noise 10^{-5}	2.452921	2.108152	1.830944
Noise 10^{-6}	0.590361	0.647309	0.655850
Noise 10^{-7}	0.567785	0.630632	0.640402
Noise 10^{-8}	0.567828	0.630739	0.640429

As shown in the Table III, when moving from higher to lower noise magnitudes, the RMSE values generally decrease, indicating improved forecast accuracy. However, beyond the noise level of 10^{-7}, a slight increase in RMSE is observed for certain variables, suggesting that 10^{-7} represents the optimal perturbation magnitude.

These results highlight the critical role of stochastic perturbation magnitude as a control parameter in ensemble-based post-processing. Appropriately scaled noise enables the ensemble to capture the inherent uncertainty of atmospheric systems more effectively, thereby enhancing both accuracy and stability. In practical terms, perturbation levels around 10^{-7} offer the most favorable balance between ensemble diversity and forecast fidelity, providing a simple yet effective mechanism for improving deterministic machine-learning weather prediction models.

G. Effect of Ensemble Size

To investigate how the number of ensemble members influences forecast performance, experiments were conducted with ensembles of 2, 5, 10, and 15 members, while keeping the noise magnitude fixed at 10^{-7}. Each ensemble was generated by applying independent Gaussian perturbations to the model inputs, and the forecasts were aggregated using the ensemble mean. The results were evaluated using RMSE to quantify forecast accuracy across key variables.

TABLE IV
RMSE COMPARISON ACROSS DIFFERENT ENSEMBLE SIZES

Var.	M=2	M=5	M=10	M=15
2m_temp.	0.567918	0.567859	0.567790	0.567802
10m_u_comp	0.630905	0.630703	0.630621	0.630688
10m_v_comp	0.640605	0.640384	0.640331	0.640317
sp_humidity	0.000225	0.000225	0.000225	0.000225
temperature	0.443039	0.442807	0.442754	0.442707

As shown in Table IV, increasing the number of ensemble members generally improved forecast accuracy. RMSE values for all variables decreased as the ensemble size increased from 2 to 10, reflecting enhanced stability and reduced random variability. Larger ensembles average out more stochastic noise, allowing the forecasts to better represent the underlying atmospheric state.

However, beyond 10 members, the improvement in accuracy diminishes. The RMSE values for 10- and 15-member ensembles were close, indicating that the forecasts had reached a point of convergence where adding more members contributed little additional benefit. This diminishing return occurs because once a sufficient level of ensemble diversity is achieved, new members tend to replicate existing patterns rather than introduce new information.

From a computational resource point of view, larger ensembles increase runtime and resource usage nearly linearly with the number of members. Therefore, identifying an efficient ensemble size is essential for balancing accuracy and efficiency. The results suggest that an ensemble of around 10 members achieves this balance-offering stable, high-quality forecasts without unnecessary computational overhead.

H. Summary of Findings

This study shows that ensemble-based post-processing is an effective and computationally efficient way to enhance the predictive accuracy of the GraphCast_small model. The optimal configuration for improving deterministic forecasts was identified through ensemble aggregation, bias correction, and sensitivity analysis.

Ensemble aggregation methods such as mean, median, and trimmed mean produced smoother forecasts than the baseline model. The trimmed mean consistently achieved the lowest RMSE, demonstrating robustness against outliers and improved stability. Despite small numerical differences, these methods systematically reduced random fluctuations in predictions.

Bias correction using simple linear regression further improved RMSE for near-surface variables like 2m_temperature and 10m_wind components. However, slight over-adjustment in temperature at pressure levels indicated that bias correction should be applied selectively based on variable sensitivity.

Sensitivity analyses revealed that reducing the Gaussian noise from 10^{-1} to 10^{-8} improved accuracy, with 10^{-7} identified as the optimal magnitude. Increasing ensemble size enhanced stability up to about 10 members, beyond which performance gains leveled off.

Overall, an ensemble of about 10 members with a Gaussian noise of 10^{-7} and trimmed mean aggregation achieved the best balance between accuracy, robustness, and efficiency, providing a simple yet powerful framework for improving GraphCast forecasts.

V. DISCUSSION

The experimental results confirm that ensemble-based post-processing can improve the reliability and interpretability of deterministic machine-learning forecasts such as GraphCast_small. The proposed framework, which combines stochastic perturbation, statistical aggregation, and simple bias

correction, demonstrates that meaningful gains in forecast accuracy can be achieved without retraining or modifying the underlying model.

A. Interpreting Ensemble Aggregation Performance

The ensemble aggregation experiments highlight that even small perturbations in the model's input space can generate sufficient diversity to stabilize the final predictions. Aggregation through mean, median, and trimmed mean consistently reduced RMSE compared to the single deterministic output, validating the effectiveness of ensemble averaging in suppressing random noise and improving spatial coherence. The trimmed mean, in particular, emerged as the most stable technique, reflecting its capacity to minimize the influence of extreme ensemble members and capture the central tendency of the forecast distribution.

B. Effectiveness of Bias Correction

The bias-correction step introduced an additional layer of refinement to the ensemble forecasts. By applying a simple linear regression between ensemble predictions and observed values, small but consistent improvements were observed for key surface variables. This indicates that even a minimal post-processing adjustment can correct residual systematic errors inherited from the model. However, the over-adjustment observed in the temperature variable suggests that bias correction must be variable-specific. Inappropriate or aggressive, or too simple statistical adjustments may amplify error in regions where the ensemble already exhibits balanced behavior. Future applications may benefit from adaptive or multivariate bias-correction schemes that account for inter-variable relationships and spatial dependencies.

C. Sensitivity to Ensemble Parameters

The sensitivity analyses provided key insights into the trade-off between ensemble diversity and computational efficiency. The noise-magnitude experiments established 10^{-7} as the optimal perturbation level for the GraphCast_small model, beyond which noise no longer contributed beneficial variability and instead introduced artificial distortions. This finding highlights that stochastic perturbations must be carefully scaled relative to the input sensitivity.

Similarly, experiments with varying ensemble sizes revealed a saturation point around 10 members, beyond which improvements in RMSE became negligible. This suggests that the ensemble had achieved sufficient sampling of the forecast uncertainty at this size. Increasing ensemble size further would result in diminishing accuracy gains while linearly increasing computational cost.

VI. Conclusion

This study demonstrated that ensemble-based post-processing can effectively enhance the accuracy and stability of the GraphCast small model without retraining. By introducing controlled Gaussian perturbations and applying statistical aggregation, deterministic forecasts were converted into more robust estimates.

Ensemble aggregation consistently reduced RMSE and produced smoother spatial fields, with the trimmed mean method offering the most stable performance. A simple linear bias correction further improved near-surface variables, confirming that lightweight statistical adjustments can refine ML-based forecasts.

Sensitivity analyses identified an ensemble size of around 10 members and a noise magnitude of 10^{-7} as optimal, balancing accuracy and computational cost. These results highlight that a well-tuned ensemble configuration provides a practical pathway to enhance deterministic machine-learning weather models, improving both reliability and interpretability.

VII. Future Work

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:

A. Multi-step Forecast Evaluation

The present experiments are limited to single-step forecasting due to the resource constraints. 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.

B. Dynamic Noise Modeling

In this study, a fixed Gaussian noise magnitude of 10^{-7} 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.

References

[1] R. Lam et al., "GraphCast: Learning skillful medium-range global weather forecasting," arXiv.org, Aug. 04, 2023. https://arxiv.org/abs/2212.12794?mc_cid=2622455cb4&mc_eid=51768751d5

[2] M. Leutbecher and T. N. Palmer, "Ensemble forecasting," Journal of Computational Physics, vol. 227, no. 7, pp. 3515–3539, Mar. 2008, doi: https://doi.org/10.1016/j.jcp.2007.02.014

[3] J. Pathak et al., "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators," arXiv:2202.11214 [physics], Feb. 2022, Available: https://arxiv.org/abs/2202.11214

[4] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast," arXiv:2211.02556 [physics], Nov. 2022, Available: https://arxiv.org/abs/2211.02556

[5] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "ClimaX: A foundation model for weather and climate," arXiv:2301.10343 [cs], Feb. 2023, Available: https://arxiv.org/abs/2301.10343

[6] S. Rasp and S. Lerch, "Neural Networks for Postprocessing Ensemble Weather Forecasts," Monthly Weather Review, vol. 146, no. 11, pp. 3885–3900, Oct. 2018, doi: https://doi.org/10.1175/mwr-d-18-0187.1.

[7] "Google Cloud console," Google.com, 2025. https://console.cloud.google.com/storage/browser/dm_graphcast (accessed Oct. 14, 2025).

[8] google-deepmind, "graphcast/graphcast at main · google-deepmind/graphcast," GitHub, 2023. https://github.com/google-deepmind/graphcast/tree/main/graphcast (accessed Oct. 14, 2025).