# PiggyCast: IMPROVING WEATHER PREDICTION ACCURACY THROUGH A STACKING-BASED ENSEMBLE AI APPROACH

Josiah K. Kimani[1] Oliver Angélil[2] Chris Fotso [2]

[1]African Institute for Mathematical Sciences (AIMS) [2]Ishango.ai

Climate Change AI
CCAI @ NeurIPS 2025
San Diego, California + Virtual
supported by: Mila

AIMS African Institute for Mathematical Sciences SOUTH AFRICA

## Abstract

Recently, Artificial Intelligence Weather Prediction (AIWP) models have outperformed classical Numerical Weather Prediction (NWP) Models in various weather prediction benchmarking criteria. Given the paradigm shift from numerical to machine learning models, such forecasts can be generated in seconds to minutes on a standard laptop. In this study, we propose a supervised machine learning ensemble strategy, known as **PiggyCast** (a portmanteau of the words **piggy**back and fore**cast**), trained on top of these forecast datasets (a method known as "stacking") to predict ERA5 variables, thereby exploiting the strengths of each base model and aiming to outperform forecasts from any base model alone. The improvement in PiggyCast's Root Mean Squared Error on Geopotential Height at 500 hPa, relative to the base models, was notable, with an increase in performance as forecast lead time increased.


Figure 1. PiggyCast Model: Traditional machine learning ensemble model (stacking) trained on top of forecasts of current state-of-the-art weather prediction models. Image generated with Gemini.

## Objectives

1. Analyse the **similarities and dissimilarities** in error patterns among **numerical, AI-based and hybrid weather prediction models** for **optimised model selection**.
2. Develop and assess an **ensemble machine learning model** through **stacking** forecasts from the selected weather prediction models for **enhanced predictive performance**.
3. Investigate the **effect of input features on the trained ensemble model** for **interpretability and explainability** of the forecasting process.

## Data Description

The data used in this study comprises the ERA5 reanalysis dataset and the forecasts of different AI, numerical and hybrid weather prediction models. The data are publicly accessible via the WeatherBench 2 framework [4], facilitating independent analysis, evaluation, and further research. We considered one atmospheric weather variable: **Geopotential Height at 500 hPa** for **9 lead times (48, 72, 96, 120, 144, 168, 192, 216 and 240 hours)** at longitude/latitude $64 \times 32$ $(5.625^\circ)$ spatial grid resolution and **12-hour** temporal resolution for the entire globe.

Table 1. Summary of various weather prediction models considered in this study

| Model | Resolution | Forecast Range | Architecture |
|---|---|---|---|
| IFS HRES | 0.1° | 10 days | Traditional NWP |
| Pangu-Weather | 0.25° | 7 days - 10 days (scalable) | 3D Vision Transformer |
| GraphCast | 0.25° | 10 days | Graph NN |
| NeuralGCM | $\sim 0.7°$ | 10 days | Hybrid (GCM + ML) |

## Methodology

1. **Model Interdependency (Objective 1): Multidimensional Scaling (MDS)** [2]
   - The stress function is defined as:

$$\text{Stress}(X) = \sqrt{\frac{\sum_{i<j}(d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2}{\sum_{i<j} d_{ij}^2}}, \tag{1}$$

   where $\sum_{i<j} d_{ij}^2$ is the sum of the squares of the upper elements of the dissimilarity matrix, which normalises the stress and $(\mathbf{x}_j, \mathbf{x}_j) \in \mathbb{R}^p$ the MDS coordinates and $p$ the dimension of the embedding space.

2. **PiggyCast with XGBoost Regressor(Objective 2):** XGBoost algorithm [1].
   - Let $f_{t+\tau}^{(i)}(\mathbf{x})$ denote the forecasted value at spatial location $\mathbf{x}$ and lead time $\tau$ hours by model $i$. The feature vector for each instance is defined as:

$$\mathbf{x}_{\text{input}} = \left[f_{t+\tau}^{(\texttt{GraphCast})}(\mathbf{x}), f_{t+\tau}^{(\texttt{Pangu})}(\mathbf{x}), f_{t+\tau}^{(\texttt{NeuralGCM})}(\mathbf{x}), f_{t+\tau}^{(\texttt{IFS HRES})}(\mathbf{x}), \text{lon}(\mathbf{x}), \text{lat}(\mathbf{x})\right], \tag{2}$$

   with the target variable: $y = f_{t+\tau}^{(\texttt{ERA5})}(\mathbf{x})$.
   - The loss function around $\hat{y}_i^{(t-1)}$ (prediction from the ensemble up to iteration $t-1$):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t), \tag{3}$$

   where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the first-order gradient, and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the second-order derivative (Hessian).
   - Evaluation Metrics **Area-weighted RMSE:** Let $M_1$ and $M_2$ represent ERA5 value and a model spatial forecast, respectively, then

$$\text{RMSE}(M_1, M_2) = \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot (M_{1,i} - M_{2,i})^2}{\sum_{i=1}^{n} w_i}}, \tag{4}$$

   where $w_i = w(\phi_i) = \cos\left(\frac{\pi \cdot \phi_i}{180}\right)$, is the area weight for grid point $i$ on latitude $\phi$ and $n$ is the number of grid points.
   **Model Performance Across Lead Times:** For each fold and model, we compute the RMSE, & average across folds:

$$\text{RMSE}_{\text{avg}}^{(m,\tau)} = \frac{1}{K}\sum_{k=1}^{K}\text{RMSE}_k^{(m,\tau)}, \tag{5}$$

   where $m$ is the model and $K = 10$ is the number of folds.

3. **Feature Attribution (Objective 3):** SHAP (**SHapley Additive exPlanations**) values [3].
   - Given a trained model $f$ and input $\mathbf{x}$, the SHAP value $\phi_j$ for feature $j$ represents the marginal contribution of that feature:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M}\phi_j, \tag{6}$$

   where $\phi_0$ is the base value (mean model output) and $M$ is the number of input features

## Results

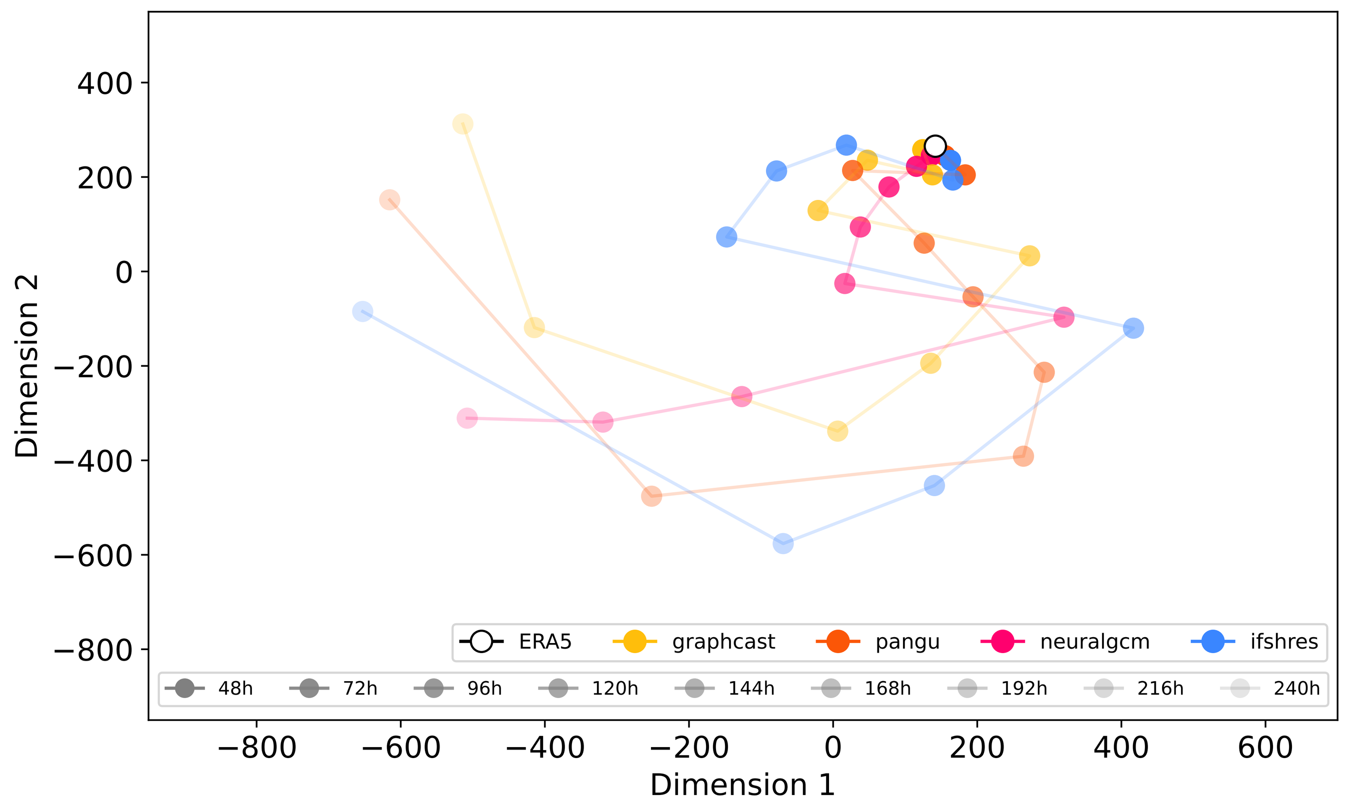


Figure 2. 2-D MDS plot of pairwise area-weighted RMSE over 48-240 hours lead times.

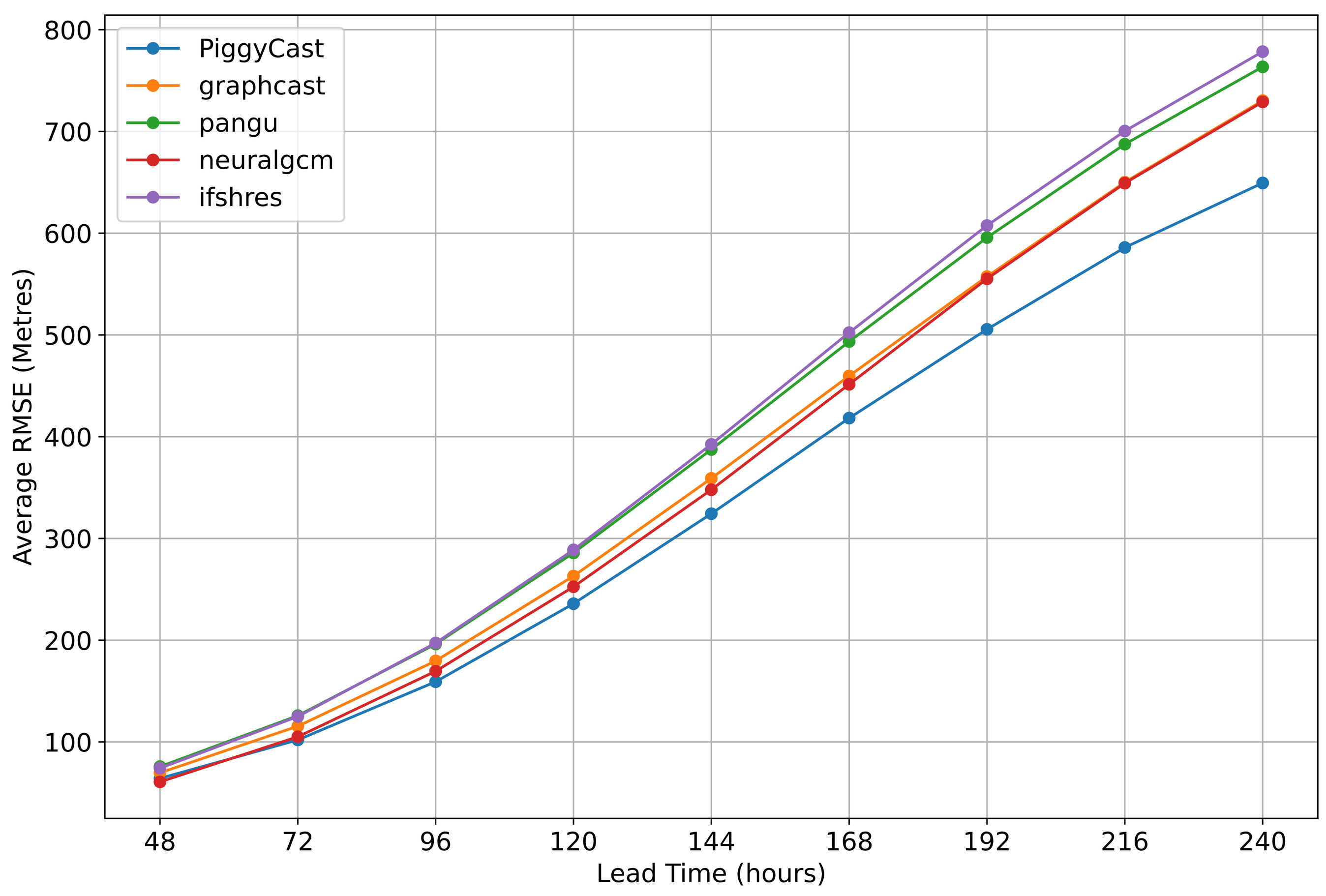


Figure 3. A plot of average area-weighted RMSE for all models, 48-240 hours lead times.

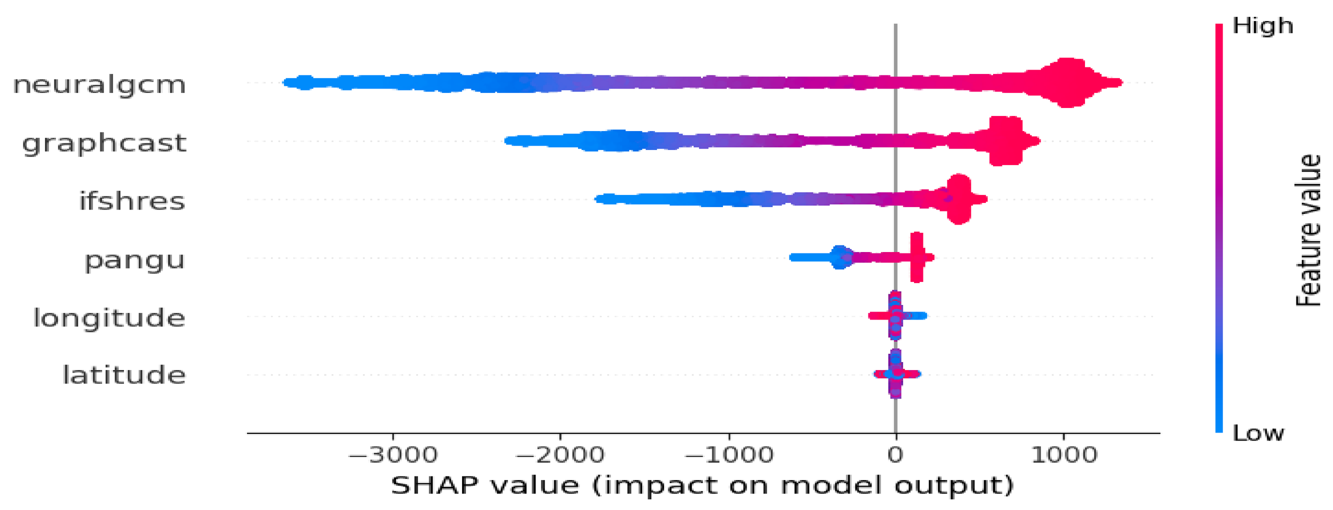


Figure 4. A beeswarm plot of SHAPley values of PiggyCast's features at 72-hour lead time.

## Conclusion

Given the low compute cost of making forecasts, and that each frontier AIWP model has its strengths and limitations (depending on the weather variable, region of the globe, and forecast lead time), **we argue that the future of the most skilful weather forecasts will likely come from machine learning stacking**, by the very nature that stacking typically yields performance better than any base model alone.

## Future Work

- Incorporation of **additional weather variables** and **weather prediction models**.
- **Testing the sensitivity of PiggyCast**, relative to the best base model, to the **spatial scale of a region**. Since in this study we optimise for weather globally, we suspect that when **targeting smaller spatial scales, PiggyCast may exhibit impressive results**.
- Extension of the framework to address **extreme weather events, nonstationary climate conditions** and **uncertainty quantification**.

## References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[2] Trevor Cox and Michael Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, 2 edition, 2000.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[4] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2024.

Source Code