PiggyCast: Improving Weather Prediction Accuracy through a Stacking-Based Ensemble AI Approach

Josiah Kiarie Kimani (josiah@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Oliver Angélil
Ishango.ai, Switzerland
and
Chris Toumping Fotso, PhD student
Ishango.ai, Germany

13 June 2025

Submitted in partial fulfilment of AI for Science Master's degree at AIMS South Africa

AIMS | African Institute for Mathematical Sciences SOUTH AFRICA

Abstract

Recently, AI Weather Prediction (AIWP) models have outperformed classical Numerical 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. Forecast datasets from frontier AIWP models for the year 2020 have been made publicly available on the WeatherBench 2 website, facilitating independent analysis, evaluation, and further research. In this study, we introduce a traditional machine learning model 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. We coin our model **'PiggyCast'**, as we effectively piggyback off the work done by leading AI research teams with expertise and compute budgets for model training that are hard to match in an MSc thesis. The improvement in PiggyCast's Root Mean Squared Error on Geopotential Height at 500 hPa pressure, relative to the base models, was notable, with an increase in performance as forecast lead time increased. 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.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Josiah Kiarie Kimani, 13 June 2025

Abstract in Swahili

Hivi karibuni, miundo ya Kutabiri Hali ya Hewa kwa kutumia Akili Unde (AIWP) imeweza kuzipiku mbinu za kawaida za kihesabu katika vipimo mbalimbali vya tathmini ya utabiri wa hali ya hewa. Kwa mabadiliko haya makubwa kutoka kwa mifumo ya kihesabu kwenda kwenye ujifunzaji wa mashine, sasa utabiri unaweza kuzalishwa kwa sekunde hadi dakika kwa kutumia kompyuta ya kawaida. Seti za data za utabiri kutoka kwa miundo ya kisasa ya AIWP kwa mwaka wa 2020 zimewekwa wazi kwa umma kupitia tovuti ya WeatherBench 2, jambo ambalo limewezesha uchambuzi wa kujitegemea, tathmini na utafiti zaidi. Katika utafiti huu, tunawasilisha mfano wa ujifunzaji wa mashine wa jadi uliofunzwa juu ya seti hizi za data za utabiri (inayojulikana kama "stacking") ili kutabiri ERA5 — kwa kutumia uwezo wa kila mfano wa msingi, kwa lengo la kuzidi ubora wa utabiri wa kila mfano mmoja mmoja. Tunaupa jina mfano wetu **"PiggyCast"**, kwa kuwa tunategemea kazi iliyofanywa na timu kubwa za utafiti wa AI zenye utaalamu na uwezo mkubwa wa kompyuta ambao si rahisi kufikiwa ndani ya tasnifu ya shahada ya uzamili. PiggyCast ilionesha maboresho ya maana kwenye makosa ya mizizi ya wastani wa mraba (RMSE) kwa urefu wa geopotential kwenye shinikizo la 500 hPa, ikilinganishwa na miundo ya msingi, na utendaji wake uliendelea kuboreka kadri muda wa utabiri unavyoongezeka. Kwa kuzingatia kuwa gharama ya kompyuta ya kufanya utabiri ni ndogo, na kila mfano wa AIWP una nguvu na mapungufu yake (kulingana na kipimo cha hali ya hewa, eneo la dunia, na muda wa utabiri), tunapendekeza kwamba mustakabali wa utabiri wa hali ya hewa wenye umahiri zaidi utaegemea mbinu ya stacking, kwa kuwa kawaida stacking huleta matokeo bora zaidi kuliko mfano wowote mmoja peke yake.

Contents

1. Introduction

This chapter introduces our research by highlighting the background of this study. The objectives that guide this research are outlined along with a brief explanation of the significance of this research. Finally, the outline of the research is summarised.

1.1 Background

Accurate weather prediction is vital for effective disaster readiness, resource management, and societal resilience. Traditionally, weather and climate forecasts have depended on dynamical, physics-based Numerical Weather Prediction (NWP) models, which explicitly model weather processes by solving the governing equations of fluid dynamics and thermodynamics (Scher, 2020; Rasp et al., 2024). While these models have achieved remarkable skill, their computational complexity and sensitivity to initial conditions present challenges, especially for high-resolution and long-range forecasts (Kieu, 2024).

Recent advances in statistical and machine learning (ML) techniques have positioned them as powerful alternatives or complements to traditional NWP models. These methods, often described under the broader umbrella of Artificial Intelligence Weather Prediction (AIWP) models, can efficiently learn complex relationships from large datasets and correct systematic biases in dynamical model outputs while remaining computationally efficient in generating forecasts (Kochkov et al., 2023; Ben Bouallègue et al., 2023). However, purely data-driven models may lack physical interpretability and struggle to generalise beyond the training data, particularly under changing climate conditions and extreme weather events (Rasp et al., 2024).

To harness the strengths of both paradigms, hybrid forecasting systems have gained popularity. Hybrid models deliberately integrate predictions from dynamical (physics-based) and data-driven (artificial intelligence or statistical) models, aiming to enhance forecast skill across a range of hydroclimatic and meteorological variables and events, such as temperature, rainfall, streamflow, and extreme weather (Kochkov et al., 2023).

At the cusp of rapid progress in Artificial Intelligence (AI), the increasing availability of high-quality weather and climate data, and advances in computational resources, benchmarking efforts, such as WeatherBench and its successors [1], have provided standardised frameworks for evaluating and comparing the skill of AI, NWP, and hybrid models on common datasets and metrics (Rasp et al., 2024, 2020).

This study builds on these developments by systematically analysing the error structures of leading AI-based, hybrid, and traditional NWP models to quantify their similarities and differences. Additionally, it recommends an ensemble machine learning model trained on top of the forecasts of these models to demonstrate that such a combined approach can surpass the predictive performance of any single base model.

1.2 Objectives of the Study

The objectives guiding this study are to:

1. Analyse the similarities and dissimilarities in error patterns among numerical, AI-based and hybrid weather prediction models for optimised model selection.

[1] Weatherbench website https://sites.research.google/weatherbench/

2. Develop and assess an ensemble machine learning model through stacking forecasts from numerical, AI-based and hybrid 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.

1.3 Significance of the Study

Understanding the strengths and limitations of different weather forecasting paradigms is essential for advancing operational weather prediction. By elucidating the error characteristics of AI, hybrid, and NWP models, this work informs the development of next-generation weather forecasting systems. The proposed stacking approach, applied to state-of-the-art (SOTA) weather prediction models with interpretability in mind, offers potential for improving forecast accuracy and reliability, optimising model selection for specific applications, and guiding future research in both meteorology and data science.

1.4 Research Outline

The rest of this report is organised as follows:

- Chapter 2 explores the literature on related work to our research.
- Chapter 3 highlights the data sources and methodology employed in this study.
- Chapter 4 outlines the study's findings and discusses their significance in the context of the current research.
- Chapter 5 summarises the conclusions and suggests directions for potential future work.

2. Literature Review

This chapter covers the recent literature on weather forecasting, highlighting the evolution of weather forecasting from NWP systems to AI-based models and, finally, to hybrid forecasting models. The chapter also includes a summary of the benchmarking frameworks in weather forecasting, current gaps, and potential research directions.

2.1 Numerical Weather Prediction: Foundations and Limitations

NWP has been the cornerstone of operational meteorology for decades, leveraging the fundamental equations of atmospheric physics to simulate the evolution of weather systems (Scher, 2020). The success of the ECMWF Integrated Forecast System High Resolution (IFS HRES) and other NWP models is evident in their ability to capture large-scale atmospheric dynamics and provide reliable forecasts across a range of spatial and temporal scales (Magnusson et al., 2024). However, these models are computationally intensive and sensitive to initial conditions, which fundamentally limits their skill, particularly at high resolutions and over longer lead times (Krishnamurthy, 2019; Kochkov et al., 2023). The atmosphere's chaotic nature, as described in the pioneering work of Lorenz (1963), means that small uncertainties in initial states can rapidly amplify, constraining practical predictability to about two weeks for most variables.

Despite advances in data assimilation, model resolution, and ensemble forecasting, systematic biases and underrepresentation of certain phenomena, such as extreme weather events and localised convection, persist in NWP systems (Magnusson et al., 2024; Kochkov et al., 2023). These challenges have motivated the exploration of alternative and complementary approaches, particularly those based on data-driven and hybrid methodologies.

2.2 Rise of Data-Driven and AI-Based Weather Models

The proliferation of large-scale atmospheric datasets and advances in machine learning (ML) have catalysed a paradigm shift in weather forecasting. Data-driven models, especially those built on deep learning architectures, have demonstrated the ability to learn complex, nonlinear relationships from historical weather data, offering rapid inference and competitive skill compared to traditional NWP (Rasp et al., 2020, 2024; Bi et al., 2022; Lam et al., 2023). Notable examples include GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2022), FuXi (Chen et al., 2023) and NeuralGCM (Kochkov et al., 2023), which have achieved SOTA performance on benchmark frameworks such as WeatherBench and WeatherBench 2.

These AI-based models excel in medium-range forecasting ($1 - 14$ days) (Lam et al., 2023), often matching or surpassing NWP skill for variables such as 850 hPa temperature (T850) and 500 hPa geopotential height (Z500) (Rasp et al., 2020; Magnusson et al., 2024). Their computational efficiency enables rapid forecast generation, which is particularly valuable for operational settings and ensemble prediction. However, purely data-driven approaches face notable limitations: they may lack physical interpretability, struggle to generalise beyond the training data, and underperform for rare or extreme events and under nonstationary climate conditions (Kochkov et al., 2023; Magnusson et al., 2024; Rasp et al., 2024).

2.3 Hybrid Forecasting: Bridging Physics and Data

To address the respective limitations of NWP and AI-based models, hybrid forecasting systems have gained prominence. These approaches combine the strengths of physics-based simulation and data-driven learning, aiming to enhance forecast skill, robustness, and interpretability across meteorological and hydroclimatic variables (Ben Bouallègue et al., 2023). Hybrid models can take various forms, including post-processing corrections (where ML models adjust NWP outputs), coupled architectures, and serial or parallel integration of dynamical and statistical components.

Recent studies have demonstrated that hybrid systems can outperform either paradigm alone, particularly for bias correction, downscaling, and probabilistic forecasting (Ben Bouallègue et al., 2023; Magnusson et al., 2024). NeuralGCM, for instance, integrates a differentiable dynamical core with machine-learned physics parameterisations, achieving both stable long-term climate simulations and competitive short-term forecast skill (Kochkov et al., 2023). The blending of AI and NWP is also evident in operational workflows, where ensemble post-processing and machine learning-based calibration are now standard practice in many weather centres (ECMWF, 2025; Met Office, 2025; Lerch et al., 2024).

2.4 Benchmarking and Model Evaluation

The rapid evolution of AI and hybrid weather models has underscored the need for standardised evaluation frameworks. Initiatives such as WeatherBench and WeatherBench 2 provide open-access datasets, common metrics, and rigorous protocols for comparing the skill of diverse forecasting systems (Rasp et al., 2020, 2024). These benchmarks facilitate reproducibility and transparency, enabling the research community to systematically assess progress and identify persistent challenges.

Evaluation metrics typically encompass anomaly correlation coefficient (ACC), root mean squared error (RMSE), and spectral fidelity for key atmospheric variables (e.g., 500 hPa geopotential height, surface temperature, precipitation). Recent work has also highlighted the importance of explainability and interpretability, with tools such as SHAPley Additive exPlanations (SHAP) providing insights into feature contributions and model decision-making (Lundberg and Lee, 2017; Silva et al., 2022). Analysing the strengths and limitations of model performance across different lead times, regions, and event types provides valuable insights for advancing model design and operational adoption.

2.5 Current Research Gaps and Directions

While AI and hybrid models have achieved remarkable advances, several open questions remain. Generalisation to out-of-sample conditions, including climate change scenarios and extreme events, is a key challenge (Kochkov et al., 2023; Rasp et al., 2024). Ensuring physical consistency, interpretability, and trustworthiness of forecasts is essential for operational and societal uptake. Furthermore, integrating uncertainty quantification and probabilistic forecasting remains an active area of research, especially as ensemble and hybrid approaches become more prevalent (Ben Bouallègue et al., 2023; Magnusson et al., 2024).

Additionally, while WeatherBench 2 provides protocols for benchmarking individual models, there is no framework for assessing cross-model ensemble performance across different variables. There is limited exploration of stacking frontier AIWP models. Gu et al. (2022) applied stacking to rainfall predictions in Taihu Basin, China, by using four base models (extreme gradient boosting (XGBoost), artificial neural networks (ANN), k-nearest neighbours (KNN) and support vector regression (SVR)) aggregated by a weighting algorithm. This approach outperformed the individual base models, providing a proof of concept for narrower contexts in flood control and water resource management projects (Gu et al., 2022).

In this study, our goal is to build upon this concept by training an ensemble machine learning model through stacking predictions of the frontier models in weather prediction. By leveraging their mutual complementarity, this approach could achieve unprecedented accuracy and robustness, thereby highlighting an opportunity to improve operational meteorology.

3. Data and Methodology

This chapter describes the data, its source and use in this study. In addition, the chapter outlines both the unsupervised and supervised learning techniques implemented.

3.1 Data Sources and Descriptions

The data used in this study comprises the ERA5 global atmospheric reanalysis dataset and the forecasts of different AI, numerical and hybrid weather prediction models. Both the ERA5 reanalysis data and forecasts are publicly accessible via the WeatherBench 2 (WB2) framework (Rasp et al., 2024).

WB2 is a benchmarking framework established to evaluate and compare data-driven and NWP models for global, medium-range weather forecasting (1-14 days). WB2 is developed collaboratively by Google DeepMind and the European Centre for Medium-Range Weather Forecasts (ECMWF) (ECMWF, 2023) and sets a reproducible standard for assessing the upcoming weather models (Rasp et al., 2024). It is an update of the original framework suggested by Rasp et al. (2020) aimed at accelerating the advancement of data-driven weather prediction models.

Figure 3.1 shows a sample of the WB2 scorecard of different weather prediction models for some variables such as pressure, temperature, humidity and wind vector. IFS HRES forecasts are evaluated against the IFS analysis, while all the other models are evaluated against ERA5 reanalysis. The values shown are absolute RMSE and the percentage improvement over IFS HRES visualised by the colouring (Rasp et al., 2024).

Through WB2, ERA5 reanalysis data and the models' forecasts are not only publicly available but also cloud-based (Rasp et al., 2024), optimised, and ready-to-use re-gridded at various resolutions for model training and evaluation. The datasets are accessible in a Google Cloud bucket in Zarr format (Google Cloud and ECMWF, 2023; WeatherBench 2 Contributors, 2024).The flexible and cloud-compatible Zarr file format supports N-dimensional data arrays. A user can chunk and compress (Research and ECMWF, 2023) the data based on their needs without loading the entire dataset to memory. Chunking is made possible by the use of Zarr stores, where Data are reorganised to utilise the High-Resolution Rapid Refresh (HRRR) model output (Gowan et al., 2022).

In contrast to the traditional Gridded Binary Second Edition (GRIB2) file format that most numerical weather prediction centres use (Gowan et al., 2022), the Zarr format shows improvements in data processing speed and access, with recorded speeds of 40 times faster for certain applications (Gowan et al., 2022). In addition, the Zarr format has been recommended for its flexibility and use in operational settings and machine learning workflows, specifically in the domain of global atmospheric sciences (Gowan, 2021). Increase in precipitation accuracy (Gowan, 2021) has been evident by the use of Zarr format with the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement Mission (IMERG) (Huffman et al., 2020) through different methods for bias correction.

Against this backdrop, WB2 leverages on Zarr format for handling large-scale geospatial and weather data in Google Cloud and HashiCorp Cloud Platform (HPC) environments (Research and ECMWF, 2023; Pandya and Guha Thakurta, 2022). It is also efficient in operational settings and scaling since the Zarr format supports parallelism. Data are split into single chunks, typically in latitude and longitude as well as time chunks, enabling efficient parallel read/write and computation (Research and ECMWF, 2023).

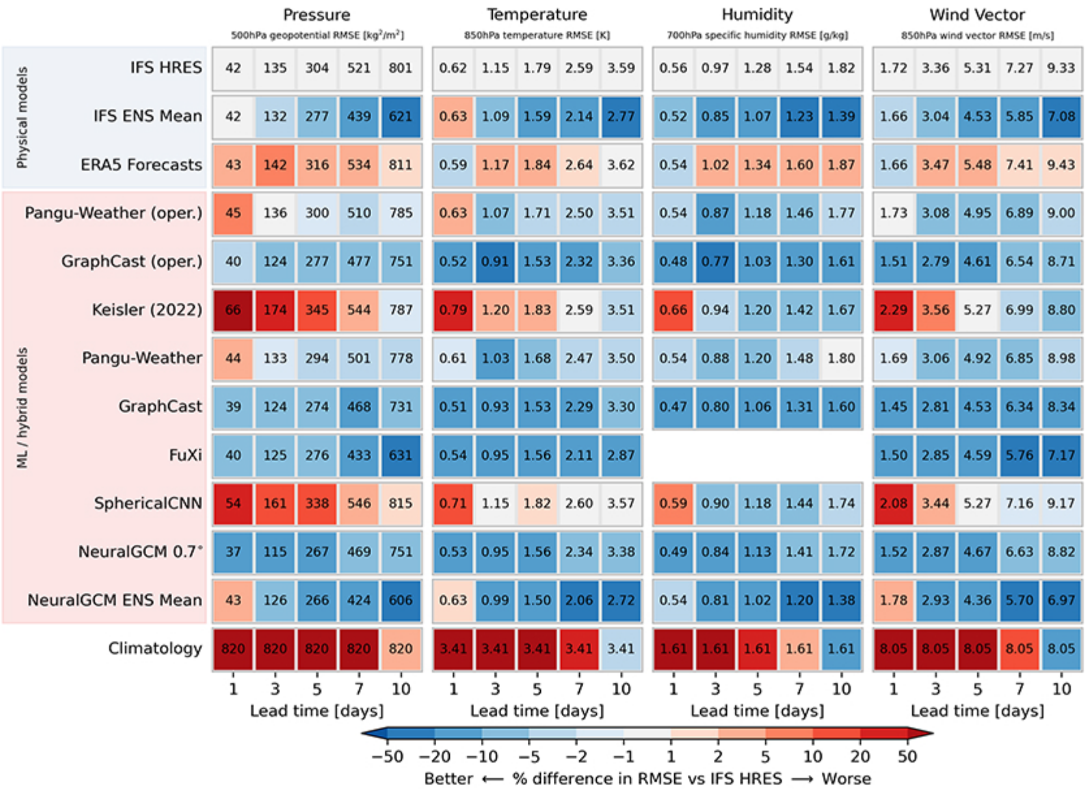

Figure 3.1: Sample of WB2 scorecard of different weather prediction models compared to IFS HRES for some variables taken from Rasp et al. (2024) Figure 1.

WB2 provides the datasets at various spatial resolutions ($0.25°$, $1.5°$ and $5.625°$), with filenames indicating the longitude-latitude grid size (1440×721, 240×121 and 64×32) (WeatherBench 2 Contributors, 2024). First-order conservative regridding was performed on all datasets, where weighting is proportional to the region where grid cells overlap on the desired and original grids (Rasp et al., 2024). Notably, 1440×721 and 240×121 datasets denoted with *'with_poles'* contained the poles ($-90°$ and $90°$ latitudes) while 64×32 files did not contain the poles, ensuring equal spacing of the grid points (WeatherBench 2 Contributors, 2024).

For bias-free model evaluation and efficiency on WB2, all model forecasts and ground truth datasets are in Zarr format (WeatherBench 2 Contributors, 2024). Overall, both local and cloud-based workflows are supported by the Zarr format, which has guaranteed compatibility with Python libraries such as `xarray` and `zarr` (WeatherBench 2 Contributors, 2024).

The loading of the data set from WB2 is very direct and is easy with the `xarray` Python package. The following code snippet is an example of loading 2020 forecasts for the GraphCast operational model:

```
# Install necessary packages
!pip install zarr xarray gcsfs

# Authenticate user for Google Cloud Storage
from google.colab import auth
```

```
auth.authenticate_user()

# import xarray package
import xarray as xr

# Load dataset
forecasts = xr.open_zarr('gs://weatherbench2/datasets/
                graphcast_hres_init/2020/date_range_2019-11-16_2021-02-01_12_
                hours-4x32_equiangular_conservative.zarr')
```

.

3.1.1 Weather Variable Under Study

In this study, we considered one atmospheric weather variable: Geopotential Height at 500 hPa pressure level for 9 lead times (48, 72, 96, 120, 144, 168, 192, 216 and 240 hours) of the models' forecasts compared to the ERA5 reanalysis for the entire globe.

Geopotential height is a fundamental atmospheric variable in weather forecasting that defines the height above mean sea level at which a specific atmospheric pressure is found. It is normalised using constant acceleration due to gravity (Wikipedia contributors, 2025; Weather Atlas, 2023). In contrast to simply measuring geometric height, which is the physical distance above sea level, geopotential height accounts for the work required to lift a unit mass against gravity from sea level to a specific point in the atmosphere (Stull, 2017; Omta and Larsen, 2018). The mathematical formulation of geopotential height based on fundamental physics in gravitational potential energy and geopotential can be found in the Appendix Section A.

Because of its physical, practical, and historical significance in meteorology, geopotential height at 500 hPa is crucial for assessing weather prediction models, including numerical and AI-based models (Rasp et al., 2020; Zhou et al., 2007). Below is a detailed explanation validating its importance in this study and weather forecasting.

1. Mid-Tropospheric Benchmark

As the geopotential height at 500 hPa pressure level (≈ 5.5 km altitude) lies in the mid-troposphere, it captures the dominant large-scale dynamics such as Rossby waves, jet streams and trough-ridge systems which play an important function in shaping weather patterns over continental and global scales (Holton, 2004; Vallis, 2017; Zhou et al., 2007). Kieu (2024) and Rasp et al. (2020) indicate that errors at this level propagate to surface weather conditions like temperature, pressure, and precipitation, making it a sensitive indicator of model performance.

2. Level of Non-Divergence

At the 500 hPa pressure level, divergence or convergence in the horizontal wind field is closely linked to vertical motion in the atmosphere. This level often acts as a pivot for upward or downward air movements. Geopotential height patterns at 500 hPa help identify these regions of divergence and convergence, making it a crucial layer for understanding and predicting the development and movement of weather systems (American Meteorological Society, 2022; Alobaidy et al., 2022; Bluestein, 1992).

3. Established Operational Metric

Operational centres like ECMWF and the National Centres for Environmental Prediction (NCEP) (NOAA

NCEP, 2025) have historically used geopotential height at 500 hPa pressure level as a standard benchmark in both operational and research communities (Kasahara and Washington, 1985; Rasp et al., 2020; Sun et al., 2023).

4. Error Saturation and Predictability

Geopotential height at 500 hPa pressure level errors grow systematically with lead time, eventually saturating at approximately 2 weeks, thus helping quantify forecast skill and atmospheric predictability (Lorenz, 1982; Zhou et al., 2007; Kieu, 2024).

5. Spectral Sensitivity

Moreover, geopotential height at 500 hPa pressure level fields are rich in spectral content and useful in evaluating model error across spatial scales, from synoptic systems (troughs, ridges) to smaller features (Sun et al., 2023; Dueben et al., 2021; Judd and Smith, 2008; Boer, 1984). Rasp et al. (2020) greatly employs 500 hPa geopotential height power spectra to quantify spectral fidelity as a key metric for diagnosing model biases in WB2.

3.1.2 ERA5 Reanalysis Dataset

As part of the Copernicus Climate Change Service (C3S), the ECMWF developed the fifth-generation global atmospheric reanalysis dataset, known as ERA5 (Hersbach et al., 2020). It covers the period from January 1940 to 5 days behind the present time, updated daily with hourly estimates of a range of global upper-surface, surface, and oceanic variables spanning 137 (1 hPa - 1000 hPa) pressure levels regridded at $0.25°$ (1440×721) longitude/latitude spatial grid resolution (Rasp et al., 2024).

In this study, the ERA5 reanalysis is used as the ground truth for the global atmospheric occurrence. These data are made available in WB2 in Zarr format, with the temporal coverage of 1959 to 2023. The data constitutes 13 pressure levels downsampled to 6 hours temporal resolution (Rasp et al., 2024). Rasp et al. (2024) choice for 13 pressure levels is to balance between model complexity and computational efficiency. Originally, WB2 downloaded the dataset from Copernicus Climate Data Store (CDS) (Copernicus Climate Change Service (C3S), 2025).

The CDS is an open-source, cloud-based platform managed by the C3S, implemented and maintained by the ECMWF under the mandate of the European Union (EU) (Copernicus Climate Change Service (C3S), 2025). It is a one-stop catalogue for accessing a variety of high-quality climate datasets covering observations from satellites and in situ sources, historical climate records, global and regional reanalyses, seasonal forecasts and climate projections (Buontempo et al., 2022). The CDS facilitates both simple visualisations and processing of large data volumes, supporting diverse user needs to address climate change challenges.

Specifically, this study used ERA5 at longitude/latitude 64×32 $(5.625°)$ spatial grid resolution using an equiangular grid (Rasp et al., 2024) with conservative remapping, 12-hour temporal resolution, atmospheric geopotential height variable at the 500 hPa of pressure and 2020 period for simplicity and computational efficiency. However, the approach used in this study can still be applied to other available resolutions and variables. The ERA5 dataset utilised in this investigation is summarised in Table 3.1.

3.1.3 Forecasts of Weather Prediction Models

As part of the overall model training and evaluation, this study analysed 4 model forecast datasets from the following WB2 models:

Table 3.1: Details of the WeatherBench 2 ERA5 ground-truth dataset used

Attribute	**Value**
Dataset Name	1959-2023_01_10-6h-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/era5/
Data Source	ERA5 reanalysis (ECMWF)
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (ground truth/observation)

1. Integrated Forecast System High-Resolution (IFS HRES) (ECMWF, 2023)
2. Pangu-Weather (Operational) (Bi et al., 2022)
3. GraphCast (Operational) (Lam et al., 2023)
4. Neural General Circulation Model (NeuralGCM, Deterministic) (Kochkov et al., 2023)

Refer to Table 3.2 for a summary of these models.

Table 3.2: Summary of various weather forecasting models considered in this study

Model	**Resolution**	**Forecast Range**	**Architecture**	**Strength**
IFS HRES	0.1°	10 days	Traditional NWP	▪ Highest-resolution ▪ Widely trusted ▪ Operational
Pangu-Weather	0.25°	7 days - 10 days (scalable)	3D Vision Transformer	▪ Fast ▪ High accuracy
GraphCast	0.25°	10 days	Graph NN	▪ Fast ▪ Operational ▪ High skill ▪ Deterministic
NeuralGCM	$\sim$ 0.7°	10 days	Hybrid (GCM + ML)	▪ Physics-based ▪ Learnable components

The selection criteria of these models were: availability of their forecasts on WB2 and their RMSE skill (≥ 100) against ECMWF HRES on geopotential height at the 500 hPa of pressure (WeatherBench 2

Contributors, 2024). This performance is concisely shown in Figure 3.2 courtesy of Rasp (2024).

The forecasts for FuXi (Chen et al., 2023) and Keisler (2022) (Keisler, 2022) models were considered but not used. The Fuxi model forecasts dataset was dropped because, after exploratory data analysis, we established that the dataset was truncated at day '2020-12-16T00:00:00.000000000'; hence, it was insufficient for the forecast period in the study.

Similarly, the Keisler model forecasts dataset was dropped because, after exploratory data analysis, we noticed missing values in the forecasts provided by WB2 (32,768 values for 16 timesteps over a 64×32 grid resolution). This issue was reported to the WB2 team as a GitHub issue as directed by WeatherBench 2 Contributors (2024).

For a more detailed overview of FuXi and Keisler models and their forecast datasets, refer to the Appendix Section B. Other models were left out since they had not yet been made available on WB2 in time for this study.

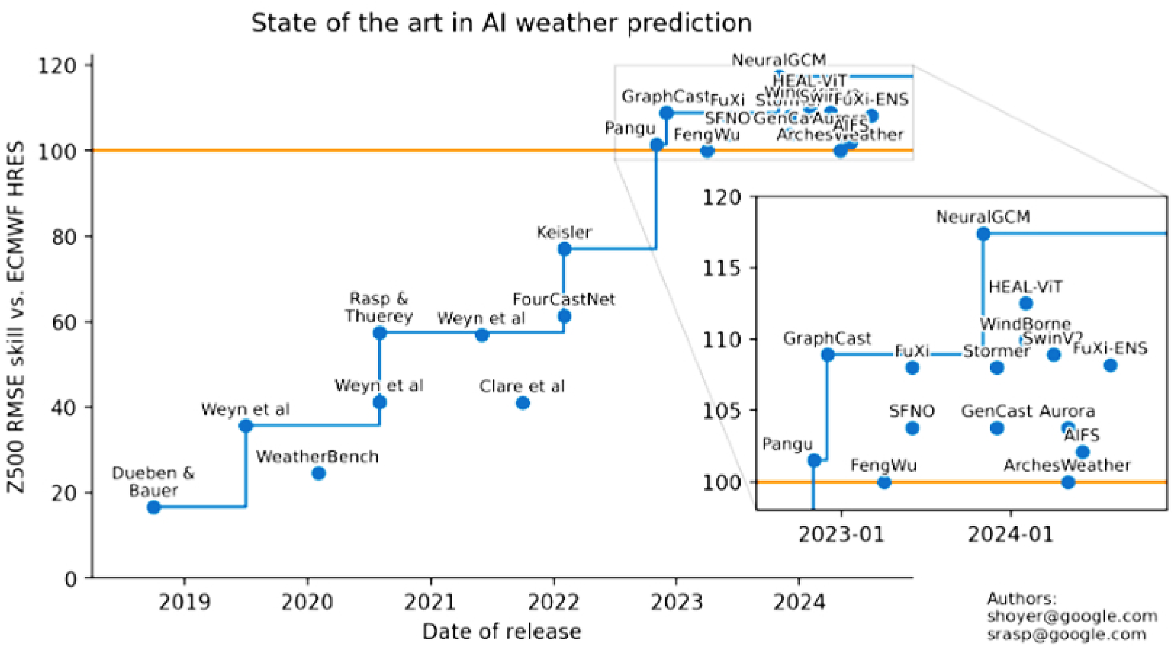

Figure 3.2: AIWP models' RMSE skill vs. ECMWF HRES for geopotential height at 500 hPa pressure level against the model's release date. The models considered in this study had an RMSE skill ≥ 100. Image Credit to Rasp (2024).

1. Integrated Forecast System High-Resolution (IFS HRES)

IFS HRES is a model developed by ECMWF and used as a gold-standard benchmark in weather forecasting (Rasp et al., 2024). IFS HRES performance on precipitation and rare extremes beyond 7 days is still unbeaten yet (Leon, 2023). It is a physics-based numerical model with 0.1° (3600×1801) longitude/latitude spatial grid resolution and 137 pressure levels.

Since it runs simulations of complex numerical hydrostatic primitive equations with physical parameterisations, it is computationally expensive and, therefore, requires hours on supercomputers (ECMWF, 2023). Rackow et al. (2025) as well as ECMWF (2023) confirms that IFS HRES still has systematic biases and truncation errors largely due to, but not limited to, time step sensitivity (Tompkins, 2004), the model's numerical representation in finite difference schemes and well-known stratospheric biases (Lawrence et al., 2022).

In this study, we specifically considered the IFS HRES dataset (Rasp et al., 2024) with longitude/latitude 64×32 $(5.625°)$ spatial resolution grid with equiangular conservative remapping, 12-hourly temporal resolution, upper-air atmospheric geopotential variable at 500 hPa pressure level and 2020 forecast period in conformity with the ERA5 ground truth dataset defined in Section 3.1.2. The Table (3.3) summarises the details of the IFS HRES forecasts dataset used in this study.

Table 3.3: Details of the WeatherBench 2 IFS HRES forecasts dataset used

Attribute	**Value**
Dataset Name	2016-2022-0012-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/hres/
Data Source	IFS HRES WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

2. Pangu-Weather (Operational)

Pangu-Weather model is an AIWP model built on a 3D vision transformer architecture developed by a research team in Huawei Cloud (Bi et al., 2022). The model employs three-dimensional (3D) deep neural networks with Earth-specific priors, 3D Earth-Specific transformer (3DEST) architecture (Bi et al., 2022), with the main goal of expediting an AI-based approach for precise medium-range global weather forecasting.

By formulating height as a separate dimension, this method enables the model to represent the interrelationships between atmospheric states at various pressure levels. Furthermore, Pangu-Weather presents a hierarchical temporal aggregation strategy that reduces accumulation errors in medium-range forecasting by training a number of models with varying forecast lead times (1, 3, 6, and 24 hours) and testing by autoregressively chaining the predictions together using the fewest steps from these models to reach the desired lead time (Bi et al., 2022).

The model's input variables include five atmospheric variables (geopotential height (Z), specific humidity (Q), temperature (T), eastward (U) and northward (V) components of wind speed, and four surface variables (2-m temperature (2T), 10U and 10V components of 10-m wind speed, and mean sea-level pressure (MSL)) at $0.25°$ longitude/latitude (1440×721) spatial grid resolution and 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1,000 hPa). Across seven days of forecasts, Pangu-Weather outperformed the operational IFS in deterministic forecast outcomes for all of these investigated variables (Bi et al., 2022), and it is scalable to ten days of forecasts.

Similarly, in this study, we considered the Pangu-Weather operational forecasts for the period of 2020 at the spatial resolution of $5.625°$ longitude/latitude (64×32) with equiangular conservative remapping, 12-hourly temporal resolution and upper-air atmospheric geopotential variable at 500 hPa. These forecasts were run by WB2 in a quasi-operational setup initialised with IFS HRES analysis data (WeatherBench 2 Contributors, 2024) since Bi et al. (2022) asserted that Pangu-weather had been initialised using ERA5 reanalysis data, which is not available in real-time for operational forecasts. The Table (3.4) summarises the details of the Pangu-Weather forecasts dataset used in this study.

Table 3.4: Details of the WeatherBench 2 Pangu-Weather (Operational) forecasts dataset used

Attribute	**Value**
Dataset Name	2020_0012_64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/pangu_hres_init/
Data Source	Pangu-Weather (Operational) WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

3. GraphCast (Operational)

GraphCast is an operational AIWP model developed by researchers at Google DeepMind and Google Research built on a Graph Neural Networks (GNNs) architecture (Lam et al., 2023). GraphCast was aimed at advancing accurate, cheaper and efficient weather forecasting, ultimately demonstrating better weather forecasting skill than IFS HRES on 90.3% of $1,380$ evaluated variables and 13 pressure levels across 10-day forecasts. It is efficient because it uses high-performance computing, which allows it to generate a 10-day prediction on a single Google Cloud TPU v4 device in less than a minute (Lam et al., 2023).

The GNNs architecture inspired by previous work (Pfaff et al., 2021; Rasp et al., 2020; Rasp and Thuerey, 2021), comprise of an encoder which maps the input state of the weather from a latitude/longitude grid to an intermediate space using a multi-mesh graph representation derived from icosahedral meshes, a processor which updates the feature space on the multi-mesh through learned message-passing computations and finally a decoder which maps the processed features from the icosahedral meshes back to the original latitude/longitude grid representation.

In addition, GraphCast also supports severe event prediction, such as tropical cyclones, where it was significantly better than IFS HRES for lead times between 18 hours and 4.75 days (Lam et al., 2023). Vertically integrated water vapour transport (IVT) prediction improved from 25% at short lead time to 10% at longer horizons when compared to IFS HRES. Extreme heat and cold precision-recall curves for GraphCast were generally above IFS HRES for 5-day and 10-day lead times (Lam et al., 2023). However, WB2's case study on Hurricane Laura (Pasch et al., 2021) noted that although GraphCast had a good track prediction and reasonable cyclone form, it failed to correctly predict the intensity of Hurricane Laura's wind speed and pressure, hence predicting the hurricane's landfall west of the actual location (Rasp et al., 2024). GraphCast does not explicitly model uncertainty like ensemble prediction systems, limiting its value for applications requiring probabilistic information about forecasts (Lam et al., 2023).

At $0.25°$ longitude/latitude (1440×721) spatial grid resolution and 37 pressure levels (1, 2, 3, 5, 7, 10, 20, 30, 50, 70, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975 and 1,000 hPa), GraphCast's input variables comprise 6 atmospheric variables: geopotential (Z), temperature (T), specific humidity (Q), vertical wind component (W), northward wind component (V) and eastward wind component (U) and 5 surface variables: mean sea-level pressure (MSL), 10-metre northward wind component (10V), 10-metre eastward wind component (10U), 2-meter temperature (2T), and total precipitation (TP) (Lam et al., 2023).

In this study, we considered GraphCast Operational forecasts for 2020 at $5.625°$ longitude/latitude (64×32) spatial grid resolution with equiangular conservative remapping, 12-hourly temporal resolution and upper-air atmospheric geopotential variable at 500 hPa. This was done in tandem with the other models and the target ERA5 reanalysis ground truth during training and evaluation. The Table (3.5) summarises the details of the GraphCast forecasts dataset used in this study.

Table 3.5: Details of the WeatherBench 2 GraphCast (Operational) forecasts dataset used

Attribute	**Value**
Dataset Name	date_range_2019-11-16_2021-02-01_12_hours-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/graphcast_hres_init/2020/
Data Source	GraphCast (Operational) WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

4. Neural General Circulation Model (NeuralGCM Deterministic)

NeuralGCM is the first fully-differentiable hybrid model, that is, a General Circulation Model (GCM) with machine learning components, developed by researchers from Google DeepMind, Google Research, ECMWF and Earth, Atmospheric and Planetary Sciences at the Massachusetts Institute of Technology (Kochkov et al., 2023).

NeuralGCM contrasts with traditional GCMs, which are physics-based simulators and pure machine learning models trained solely on reanalysis data. The model architecture consists of two key components: a differentiable dynamical core and a learned physics module. The differentiable dynamical core solves the discretised governing dynamical equations for large-scale fluid motion and thermodynamics under gravity and the Coriolis force, while the learned physics module models physical processes not accounted for by the dynamical core and computational errors using a fully connected neural network (Kochkov et al., 2023).

The dynamical core evolves seven prognostic variables: specific liquid cloud water content (q_{cl}), vorticity (ζ), specific cloud ice (q_{ci}), divergence (δ), logarithm of the surface pressure (log p_s), specific humidity (Q), and temperature (T), while the learned physics module takes these variables in the atmospheric column in addition to total incident solar radiation, sea ice concentration and sea surface temperature (Kochkov et al., 2023).

NeuralGCM utilises encoder and decoder modules to map ERA5 data at pressure levels to the model state on sigma coordinates (terrain-following) and then map the model state on sigma coordinates back to pressure levels, respectively. It was trained and evaluated at horizontal resolutions with grid spacings of $2.8°$ (128×64), $1.4°$ (256×128) and $0.7°$ (512×256) (Kochkov et al., 2023).

While NeuralGCM attains SOTA accuracy for deterministic forecasts at short lead times comparable to GraphCast and outperforming IFS HRES, its ensemble performance is slightly worse than ECMWF Ensemble at very early lead times (Kochkov et al., 2023).

Similarly, in this study, we considered the NeuralGCM model forecasts for 2020 at $5.625°$ longitude/latitude (64×32) spatial grid resolutions with equiangular conservative remapping, 12-hourly temporal resolution and upper-air atmospheric geopotential variable at 500 hPa. The Table (3.6) summarises the details of the NeuralGCM forecasts dataset used in this study.

Table 3.6: Details of the WeatherBench 2 NeuralGCM (deterministic) forecasts dataset used.

Attribute	**Value**
Dataset Name	2020-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/neuralgcm_deterministic/
Data Source	NeuralGCM WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

3.2 Model Interdependency with Unsupervised Learning

3.2.1 Multidimensional Scaling (MDS)

An unsupervised machine learning method called Multidimensional Scaling (MDS) embeds high-dimensional data into a lower-dimensional Euclidean space to display its structure (Borg and Groenen, 2005; Kruskal and Wish, 1978; Cox and Cox, 2000). In this study, we use *metric*, also called *absolute*, MDS to investigate the similarity structure of different weather prediction models based on their forecast errors.

Motivation

MDS aims to represent each model as a point in a low-dimensional space such that the Euclidean distances between these points closely approximate the dissimilarities (in our case, RMSE values) between their forecasts. This provides an interpretable n-D embedding space (n is the number of dimensions) where models with similar error patterns appear close together, and dissimilar models appear farther apart.

Distance Matrix through Pairwise RMSE Computation

To quantify the dissimilarity between different forecasting models, we compute the pairwise RMSE between their predicted 500 hPa geopotential height fields. The comparison is performed against the

ERA5 reanalysis dataset, treating it as the observational reference.

Given that the Earth's surface area varies with latitude, a uniform treatment of grid points would bias global statistics toward high-latitude regions, which occupy less physical space despite being equally represented in a latitude–longitude grid. To address this, we apply area weighting proportional to the cosine of the latitude, following standard geophysical practice (Hastings and Dunbar, 1999; Wilks, 2011).

The area weight for a grid point on latitude ϕ is given by:

$$w(\phi) = \cos\left(\frac{\pi \cdot \phi}{180}\right). \tag{3.2.1}$$

Note that this area weight value is added as a column to our data for every grid point and applied throughout our study.

Let M_1 and M_2 represent two spatial forecast fields (e.g., from two different models or a model and ERA5), then the area-weighted RMSE is computed as:

$$\text{RMSE}(M_1, M_2) = \sqrt{\frac{\sum_{i=1}^{n} w_i \cdot (M_{1,i} - M_{2,i})^2}{\sum_{i=1}^{n} w_i}}, \tag{3.2.2}$$

where $w_i = w(\phi_i)$, is the area weight for grid point i on latitude ϕ and n is the number of grid points.

To compare a set of k such forecast fields, we computed the area-weighted RMSE for every pair, resulting in a symmetric $k \times k$ dissimilarity matrix $D = [d_{ij}]$, where:

$$d_{ij} = RMSE(M_i, M_j). \tag{3.2.3}$$

This matrix captures the relative dissimilarities among forecast fields and serves as the precomputed distance matrix input to metric MDS.

Metric MDS via SMACOF

Unlike Classical MDS, which relies on eigendecomposition of a transformed distance matrix, metric MDS employs an iterative algorithm, **Scaling by MAjorising a COmplicated Function** (SMACOF), to minimise a loss function known as *stress* (De Leeuw and Heiser, 1977; Borg and Groenen, 2005).

Given the dissimilarity matrix $D = [d_{ij}]$, MDS seeks coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^p$, p being the dimension of the embedding space, such that the pairwise Euclidean distances $\|\mathbf{x}_i - \mathbf{x}_j\|$ in the embedding space approximate the original variations d_{ij}.

The stress function is defined as:

$$\text{Stress}(X) = \sqrt{\frac{\sum_{i<j}(d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2}{\sum_{i<j} d_{ij}^2}}, \tag{3.2.4}$$

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. The SMACOF algorithm iteratively updates the positions $X \in \mathbb{R}^{k \times p}$ to minimise this stress.

Implementation

We used the `scikit-learn` implementation of metric MDS (Pedregosa et al., 2011) to compute a two-dimensional embedding using the SMACOF algorithm. For improved robustness, four random initialisations of the embedding space were performed, each optimised for up to 300 iterations to minimise the stress (3.2.4). The tolerance for stable convergence was set by epsilon (1e-6), controlling the relative change in stress. All initialisations were run in parallel to improve performance. The pairwise area-weighted RMSE matrix was then passed as a precomputed dissimilarity matrix.

3.2.2 Hierarchical Clustering Using Dendrograms

To further investigate the structural similarities and dissimilarities among forecasting models, and their relative deviation from the ERA5 reanalysis and each other, we applied hierarchical agglomerative clustering to the pairwise area-weighted RMSE dissimilarity matrix. The hierarchical agglomerative clustering begins by treating each model as an individual cluster, then at each iteration, the two clusters with the smallest dissimilarity (linkage distance) are merged. This clustering continues until all models and ERA5 are grouped into a single cluster, yielding a nested hierarchy of clusters that can be visualised as a dendrogram.

The linkage distance matrix is computed through the **Unweighted Pair Group Method with Arithmetic Mean (UPGMA)**, also referred to as the `average` linkage criterion (Müllner, 2011; Jain et al., 1999). This method merges clusters based on the average pairwise distance between all members of the two clusters.

The distance between clusters A and B using the average (UPGMA) method is defined as:

$$d(A, B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} D_{ij}, \tag{3.2.5}$$

where $|A|$ and $|B|$ denote the number of elements in clusters A and B, and D_{ij} is the RMSE between models i and j. This method promotes balanced clustering by averaging all inter-model distances.

The resulting dendrogram visually interprets the forecast models' performance relative to each other and ERA5. Clusters that merge at lower heights indicate models with more similar error characteristics. This approach complements the MDS analysis by offering a hierarchical view of model similarity and dissimilarity.

Implementation

We implemented the clustering using the `scipy.cluster.hierarchy` module (The SciPy community, 2025). Given the symmetric pairwise area-weighted RMSE matrix D, we first converted it to its condensed (vector-form) representation using `scipy.spatial.distance.squareform`, which flattens the upper triangle of the matrix without the diagonal into a 1D array as required by the `linkage` function (The SciPy community, 2025).

3.3 PiggyCast with Supervised Learning

To enhance forecast accuracy beyond individual numerical/AI weather prediction models, we propose a supervised learning ensemble strategy termed **PiggyCast** (a portmanteau of the words **piggy**back and fore**cast**).

Piggyback here means "to use something that already exists or has already been done successfully to do something else quickly or effectively" (Cambridge University Press, 2025).

This approach uses gradient-boosted decision trees, implemented via the XGBoost algorithm (Chen and Guestrin, 2016), to learn a mapping from ensemble forecasts and geographic coordinates to observed geopotential height at 500 hPa.

3.3.1 XGBoost Regressor: Mathematical Formulation

XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting machines, introduced by Chen and Guestrin (2016). The model constructs an ensemble of regression trees in an additive manner, where each new tree is trained to correct the residual errors of the previous trees. A key innovation of XGBoost lies in its regularised objective function and the use of second-order Taylor expansion, which contribute significantly to both its predictive performance and computational efficiency.

Regularised Objective Function

The learning objective in XGBoost consists of a differentiable loss function that measures the model's fit and a regularisation term that penalises model complexity:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.3.1}$$

where $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i^{(t)})^2$ denotes the loss function (squared error for regression), and each f_k is a regression tree from the function space $\mathcal{F}$. The regularisation term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{3.3.2}$$

where T is the number of leaves in the tree, w is the vector of leaf weights, γ penalises the number of leaves, and λ controls the L2 regularisation on the leaf weights. This formulation explicitly discourages overly complex models, helping to reduce overfitting.

Additive Training with Second-Order Approximation

XGBoost builds the model in a stage-wise manner. At each iteration t, a new function f_t is added to minimise the objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{3.3.3}$$

where $\hat{y}_i^{(t-1)}$ is the prediction from the ensemble up to iteration $t-1$. To make optimisation tractable, XGBoost uses a second-order Taylor expansion of the loss function around $\hat{y}_i^{(t-1)}$:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{3.3.4}$$

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). This use of both gradient and curvature information allows XGBoost to perform more accurate updates than methods relying on first-order approximations alone.

Optimal Leaf Weights and Tree Score

Let I_j be the set of instances assigned to leaf j in tree f_t. The optimal weight w_j for this leaf is derived by minimising the regularised objective:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}. \tag{3.3.5}$$

Substituting the optimal weights back into the objective function gives the total loss reduction (also called the gain) for the entire tree:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{3.3.6}$$

This formulation quantifies the benefit of a given tree structure and allows the model to score different tree configurations efficiently.

Split Finding and Gain Function

To find the best split, XGBoost evaluates the gain in the objective function when a node is split into left and right children. The gain is computed as:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \tag{3.3.7}$$

where G_L, H_L and G_R, H_R are the sums of gradients and Hessians for the left and right subsets, respectively. A positive gain indicates an improvement in the model, and the split with the highest gain is selected.

Through the combined use of regularisation, second-order optimisation, and greedy structure learning, XGBoost achieves high predictive accuracy and robustness against overfitting (Chen and Guestrin, 2016). These properties make it a strong choice for regression tasks, particularly when interpretability and computational efficiency are required (Islam et al., 2024).

3.3.2 Input Features and Target Variable

Let $f_{t+\tau}^{(i)}(\mathbf{x})$ denote the forecasted geopotential height at spatial location $\mathbf{x}$ and lead time τ hours by model $i \in \{\texttt{GraphCast}, \texttt{Pangu}, \texttt{NeuralGCM}, \texttt{IFS HRES}\}$. The feature vector for each instance is defined as:

$$\mathbf{x}_{\text{input}} = \left[f_{t+\tau}^{(\texttt{GraphCast})}(\mathbf{x}), f_{t+\tau}^{(\texttt{Pangu})}(\mathbf{x}), f_{t+\tau}^{(\texttt{NeuralGCM})}(\mathbf{x}), f_{t+\tau}^{(\texttt{IFS HRES})}(\mathbf{x}), \text{lon}(\mathbf{x}), \text{lat}(\mathbf{x}) \right], \tag{3.3.8}$$

with the target variable:

$$y = f_{t+\tau}^{(\texttt{ERA5})}(\mathbf{x}), \tag{3.3.9}$$

where $f_{t+\tau}^{(\texttt{ERA5})}$ is the reanalysis ground truth from ERA5, and $\text{lon}(\mathbf{x})$, $\text{lat}(\mathbf{x})$ denote the spatial coordinates.

3.3.3 Time Series Cross-Validation

Bergmeir and Benítez (2018) assert that time series data violates the standard assumption of independently and identically distributed (i.i.d.) samples due to its inherent temporal dependencies. As such, conventional cross-validation approaches—like random shuffling—are unsuitable because they introduce look-ahead bias and data leakage. Instead, time series-aware cross-validation techniques must be used, which honour the chronological order of observations.

We employ a rolling-origin evaluation strategy known as `TimeSeriesSplit` (Hyndman and Athanasopoulos, 2021; scikit-learn developers, 2025), which partitions the data into a sequence of non-overlapping training and test sets, with the test set always following the training set in time. Formally, for a given univariate time series $\{y_t\}_{t=1}^{T}$, and K folds, the k-th training set is defined as:

$$\mathcal{D}_{\text{train}}^{(k)} = \{y_1, y_2, \dots, y_{t_k}\}, \quad \mathcal{D}_{\text{test}}^{(k)} = \{y_{t_k+g+1}, \dots, y_{t_k+g+h}\} \tag{3.3.10}$$

where g is the size of the gap between training and test sets to reduce temporal autocorrelation, and h is the horizon or length of the test window (Cerqueira et al., 2020). This approach avoids information leakage while maintaining the temporal structure of the data, ensuring the test window always follows the training window, with a temporal buffer to mitigate short-term autocorrelations and hence a robust assessment of model generalisation to future data (Roberts et al., 2017).

In our implementation, we use 10-fold cross-validation `TimeSeriesSplit`, with each fold comprising:

- **Training set:** $64 \times 32 \times 2 \times 60$ time steps (60 days of training)
- **Test set:** $64 \times 32 \times 2 \times 30$ time steps (30 days of testing)
- **Gap:** $64 \times 32 \times 2 \times 5$ time steps (5 days as autocorrelation gap)

where $64 \times 32 \times 2$ are the longitude and latitude geopotential height values evaluated twice in a day (noon and midnight).

3.3.4 Area-Weighted RMSE Evaluation

Due to Earth's curvature, the areal density of grid points is non-uniform across latitudes. We addressed this using the area-weighted RMSE (Equation 3.2.2), ensuring fair contribution from each region (Hastings and Dunbar, 1999).

3.3.5 PiggyCast Implementation

To implement PiggyCast, we trained an XGBoost regression model to learn a weighted ensemble of NeuralGCM, GraphCast, Pangu and IFS HRES weather forecasting models as presented in equation 3.3.8. The model was trained using a time-aware cross-validation strategy (`TimeSeriesSplit`) to prevent information leakage and to reflect the temporal nature of the data as shown in equation 3.3.10. Each training fold was fit using XGBRegressor with GPU acceleration (`device="cuda"`), optimising the regularised objective function via second-order gradient boosting shown in equation 3.3.4.

Note that no hyperparameter tuning was done - we chose to stick with the XGBoost defaults since they are known to be a reasonable starting point (XGBoostDevelopers, 2025). Moreover, Perez et al. (2019) in their work on comparative analysis of XGBoost assert that using default parameters is a good starting point, but tuning can yield even better results.

Importantly, we incorporated spatial heterogeneity by applying area-based sample weights during training and evaluation, ensuring that larger grid areas had proportionally greater influence on the loss function. The predictive performance was evaluated using an area-weighted RMSE metric.

3.3.6 Model Performance Across Lead Times

We evaluate performance over nine lead times $\tau \in \{48, 72, 96, 120, 144, 168, 192, 216, 240\}$ hours. For each fold and model, we compute the area-weighted RMSE, then average across folds:

$$\mathsf{RMSE}_{\mathsf{avg}}^{(m,\tau)} = \frac{1}{K}\sum_{k=1}^{K} \mathsf{RMSE}_{k}^{(m,\tau)}, \tag{3.3.11}$$

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

3.3.7 Improvement Over IFS HRES

We quantify the improvement of PiggyCast and other models relative to the IFS HRES baseline using percentage reduction in average RMSE:

$$\%\mathsf{Improvement}^{(m,\tau)} = \left(\frac{\mathsf{RMSE}_{\tau}^{(\mathsf{IFS})} - \mathsf{RMSE}_{\tau}^{(m)}}{\mathsf{RMSE}_{\tau}^{(\mathsf{IFS})}}\right) \times 100. \tag{3.3.12}$$

This is computed per fold and per lead time to assess robustness and generalisation.

3.3.8 Model Interpretability via SHAP

To gain insight into the model's decision-making process, we consider SHAP (**SHapley Additive exPlanations**) values Lundberg and Lee (2017). Given a trained model f and input $\mathbf{x}$, the SHAP value ϕ_j for feature j represents the marginal contribution of that feature:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j, \tag{3.3.13}$$

where ϕ_0 is the base value (mean model output) and M is the number of input features. These values satisfy consistency and local accuracy, enabling clear attribution of feature influence across different folds and leadtimes (Molnar, 2024).

In this study, we use SHAP values to analyse which input forecasts or geographic features contribute most to error reduction. Feature attributions were computed using the `TreeExplainer` module from the `SHAP` library (Lundberg et al., 2024). These SHAP values provided both local and global insights into how different base model forecasts and geographic features contributed to PiggyCast's output.

4. Results and Discussion

This chapter outlines the results of the unsupervised and supervised machine learning techniques implemented in this study. Performance of the PiggyCast model, for geopotential height at 500 hPa, against the other weather prediction models is outlined. Finally, a discussion section of these results concludes the chapter.

4.1 Model Interdependency Results

4.1.1 Multidimensional Scaling (MDS)

Using MDS we seek to represent each model as a point in the low-dimensional space such that models with similar error patterns appear close together while dissimilar ones appear farther apart. As a distance matrix of the high-dimensional data is required as input in MDS, we calculated the pairwise area-weighted RMSE of the models' forecasts and ERA5 reanalysis for Geopotential Height at the 500 hPa pressure level.

The Figure 4.1 shows the pairwise area-weighted RMSE lower triangle matrix of models' forecasts and ERA5 dataset at 72 hours lead time. Similarly, the models are also compared against each other.

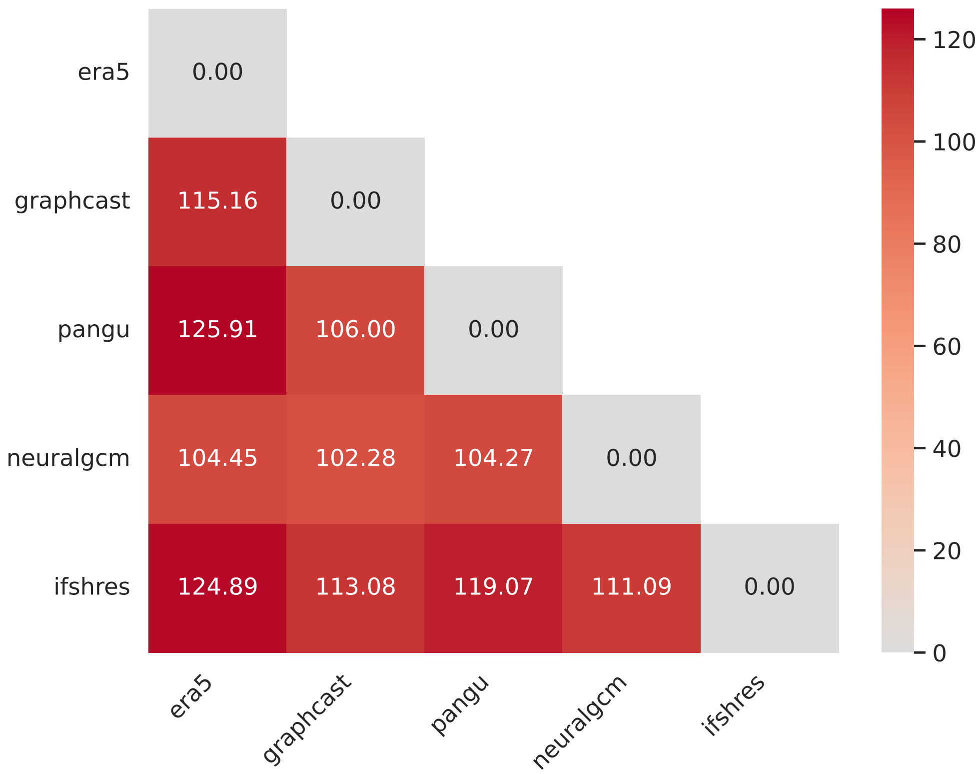

Figure 4.1: Plot of pairwise area-weighted RMSE (metres) matrix of AIWP models at 72-hour lead time 2020 forecasts.

NeuralGCM leads with the lowest RMSE of 104.45 m, followed by GraphCast with 115.16 m, IFS HRES with 124.89 m and Pangu at 125.91 m when compared to ERA5. As a sanity check, these RMSE values were similar to those reported on WB2 (Rasp et al., 2024), further validating our methodology and results.

Consequently, with the already calculated area-weighted pairwise matrix, we run the metric MDS using the SMACOF algorithm to achieve a 2-dimensional space of Euclidean distances between the model points.

Figure 4.2 shows the 2-dimensional MDS space for pairwise area-weighted RMSE for the 72-hour lead time.

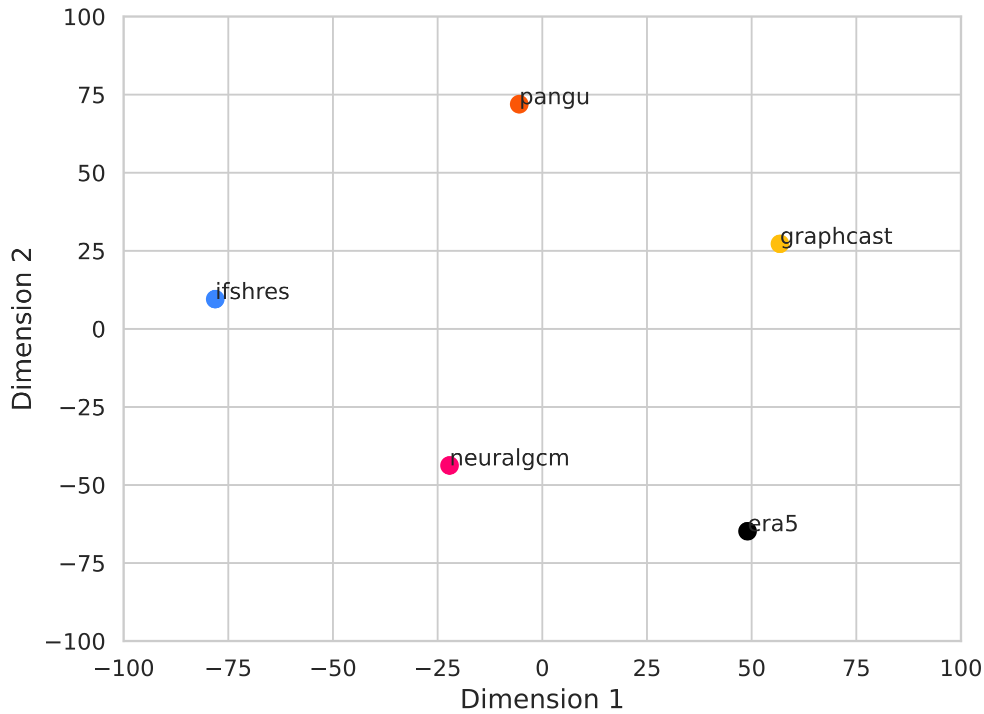

Figure 4.2: 2-D MDS plot of pairwise area-weighted RMSE for 72-hour lead time.

NeuralGCM and GraphCast are closest to ERA5, while Pangu and IFS HRES are the farthest, indicating the forecasts are most and least similar to the reanalysis, respectively. This likely highlights better performance of AI-based models compared to numerical weather prediction models over a 72-hour lead time.

We also plot the MDS evolution over the 48-240 hours lead time as shown in Figure 4.3 for the four models (models' forecast points for the respective lead times) and the ERA5 reanalysis point. From the 2-D MDS space, we can see that the models' forecast points are very close to ERA5 reanalysis at the start of the lead times, but diverge as the lead time increases, as expected. The shape of the evolution of models across the lead times is similar to tracing the letter 'S'. NeuralGCM and GraphCast remain the closest to ERA5, while Pangu and IFS HRES are farthest across lead times.

4.1.2 Hierarchical Clustering using Dendrograms

To further visualise the models' interdependency and performance of against ERA5 reanalysis, we plot a dendrogram of the hierarchical clustering tree, at 72-hour lead time as shown in Figure 4.4. From the plot, we observe that GraphCast and NeuralGCM have the earliest merge, showing high similarity in

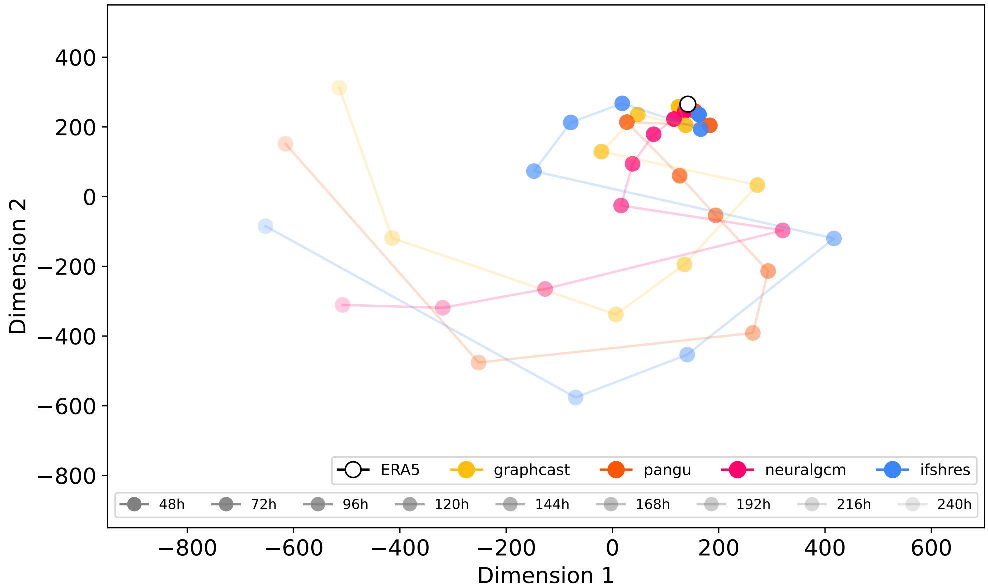

Figure 4.3: 2-D MDS plot of pairwise area-weighted RMSE over 48-240 hours lead times.

their forecast errors. This cluster is followed by a merge with Pangu, and finally, IFS HRES. The early merge of the models and the late merge with ERA5 shows that the models are closer to each other than they are to ERA5.

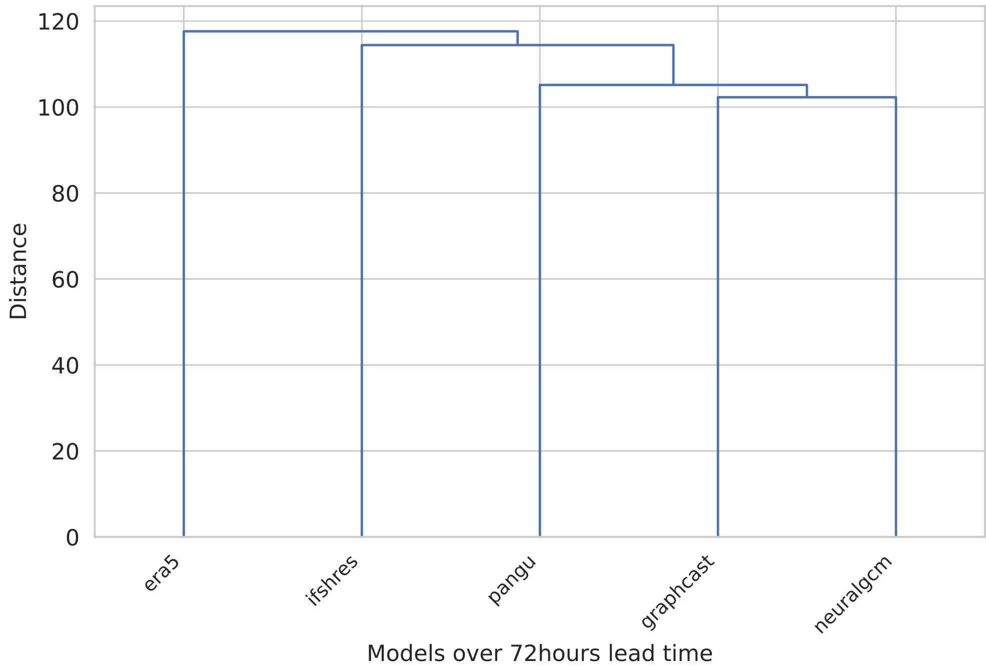

Figure 4.4: Hierarchical clustering dendrogram of pairwise area-weighted RMSE of Models (72-hour lead time).

Similarly, we plot the dendrogram of the hierarchical clustering tree of models and ERA5 across the

48-240 hours lead time as displayed in Figure 4.5. We get three clusters of orange, green and blue for clear distinction of the different cluster thresholds to observe. The orange cluster is comprised of early merges of models and ERA5 from 48-168 hours lead times. Specifically, we observe early merges of NeuralGCM and ERA5, followed closely by a merge composed of GraphCast and Pangu and finally IFS HRES at 48 hours lead time. This similar grouping is identified for a 72-hour lead time. The structure changes from 96-168 hours lead times, where NeuralGCM and GraphCast merge first, followed by Pangu, then IFS HRES. Here, IFS HRES is clustered on the left while the other models are clustered on the right. This ends the orange cluster.

At a 192-hour lead time, a green cluster on the right is observed. This cluster is made up of a merge of NeuralGCM and GraphCast, then followed by Pangu. Finally, IFS HRES is clustered higher above on the left, introducing the final blue cluster.

Similarly, for 216 and 240 hours lead times, on the right side of the cluster, NeuralGCM and GraphCast merge first, followed by Pangu, while on the left, IFS HRES is distinctly clustered alone.

From this dendrogram, we observe a distinct difference in performance (ascending order) among numerical weather prediction models (IFS HRES), pure AI-based models (GraphCast and Pangu) and Hybrid models (NeuralGCM).

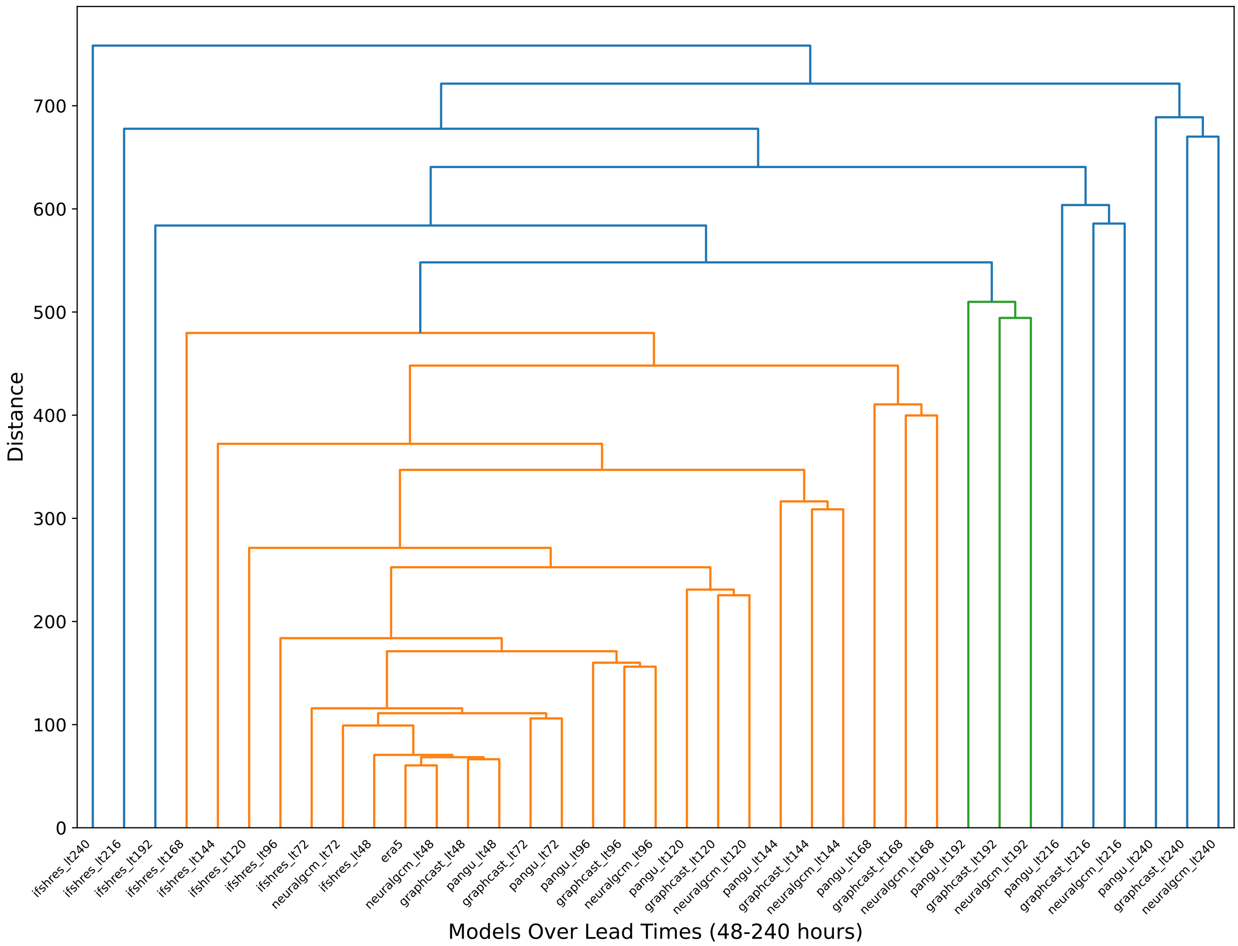

Figure 4.5: Hierarchical clustering dendrogram of pairwise area-weighted RMSE of models (48-240 hours lead time).

4.2 PiggyCast Results

Cross-validation with Time Series Split

The data split for training, autocorrelation gap and testing for the 10 folds over the 2020 forecast period at 72-hour lead time is shown in Figure 4.6. This implementation ensures that the inherent temporal dependencies in weather data are maintained, unlike the conventional cross-validation approaches like random shuffling.

Additionally, introducing a gap between the training and testing periods simulates the data latency present in real-world forecasting systems, where recent observations may not yet be assimilated or available at the time a forecast is issued.

This train, gap and test configuration is maintained for 10 folds across all 48-240 hours lead times.

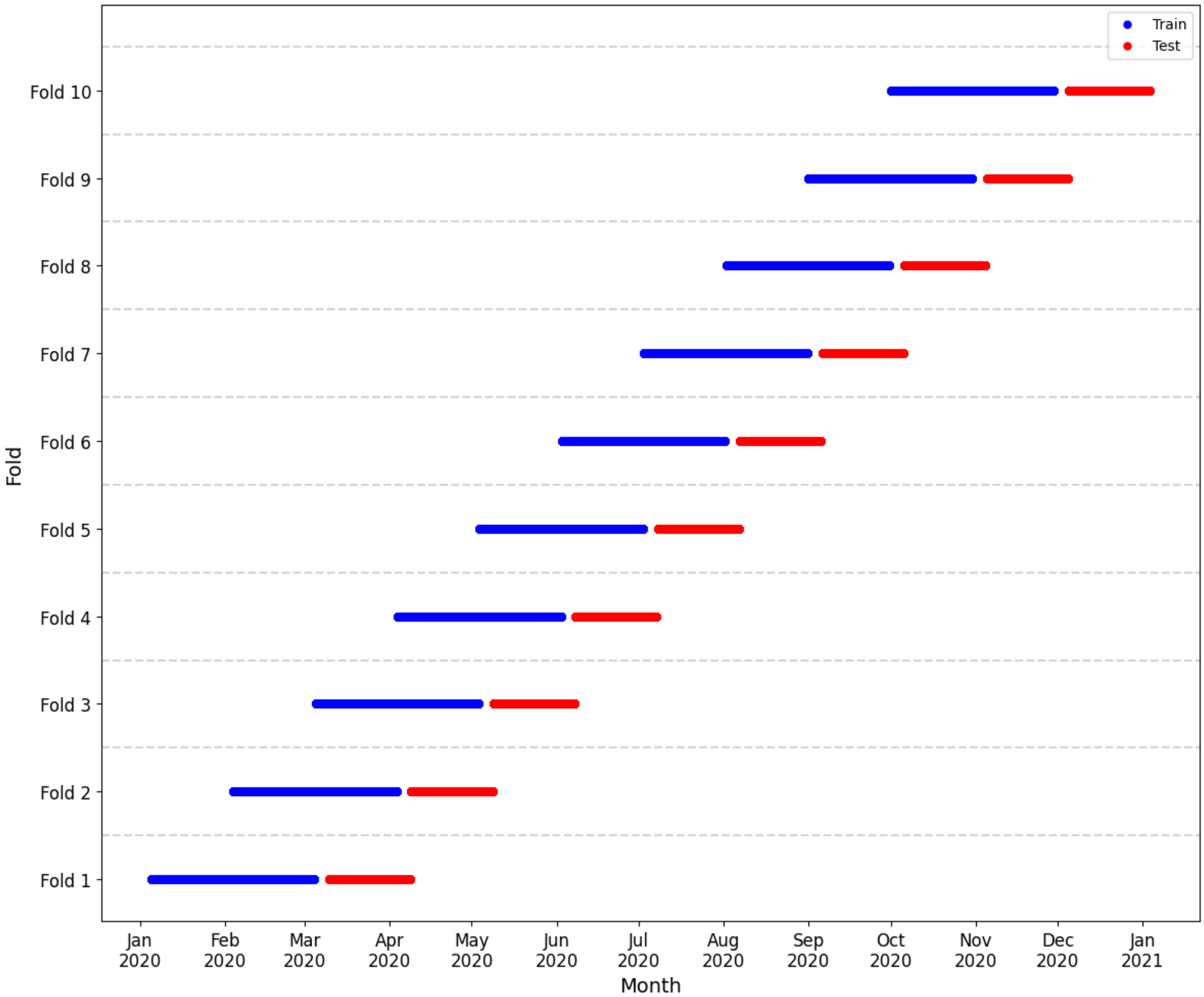

Figure 4.6: A plot of train, autocorrelation gap and test window sizes per fold for 72-hour lead time across January 2020 – January 2021.

XGBoost Regressor

We fit the XGBoost regressor using T4 Nvidia GPUs for accelerated training and testing for every fold, model and lead times. Here we track the performance of the XGBoost ensemble model using our coined model name **PiggyCast** against the performance of the other four models (NeuralGCM, GraphCast, Pangu and IFS HRES).

PiggyCast's Evaluation: Area-Weighted RMSE

The models are evaluated per fold using the weighted RMSE (Equation 3.2.2) for fair temporal grid area weighting. This weighted RMSE per fold and model for a 48-hour lead time is shown in Figure 4.7. Here, PiggyCast's performance is poor compared to the other models, specifically NeuralGCM. On average (across folds), NeuralGCM performs better with $\mathbf{60.76}$ m RMSE than all the other models, followed by PiggyCast ($\mathbf{64.18}$ m), GraphCast ($\mathbf{69.64}$ m), IFS HRES ($\mathbf{74.19}$ m) and finally Pangu ($\mathbf{75.76}$ m) on this short lead time.

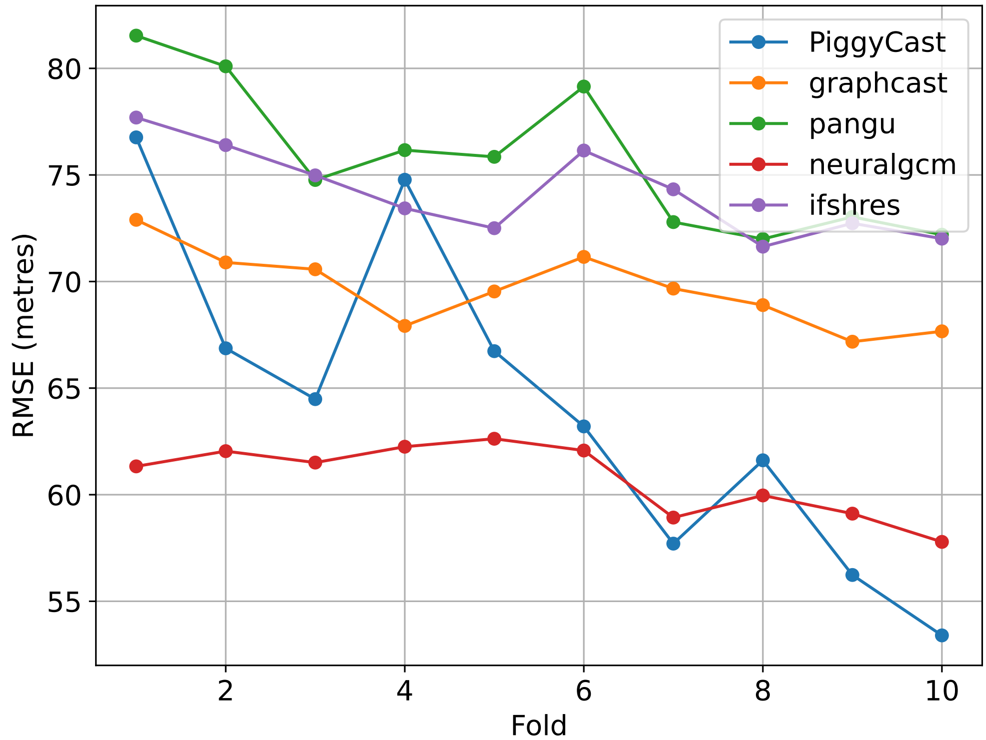

Figure 4.7: A plot of area-weighted RMSE per fold and model for a 48-hour lead time.

At a 72-hour lead time, PiggyCast performance is noted to improve and provides the best predictions compared to the other models, as shown in Figure 4.8. It is only at fold one where NeuralGCM is better than PiggyCast, while at fold four, NeuralGCM matches the performance of PiggyCast. On average (across folds), PiggyCast's RMSE leads with $\mathbf{101.99}$ m, followed by NeuralGCM ($\mathbf{105.11}$ m), GraphCast ($\mathbf{115.48}$ m), IFS HRES ($\mathbf{124.99}$ m) and finally Pangu ($\mathbf{125.84}$ m).

Similarly, PiggyCast's performance continues to dominate the other models until the last 240-hour lead time, as shown in Figure 4.9. On average (across folds), PiggyCast's RMSE leads with $\mathbf{649.38}$ m, followed by NeuralGCM ($\mathbf{729.08}$ m), GraphCast ($\mathbf{730.32}$ m), Pangu ($\mathbf{763.50}$ m) and finally IFS HRES ($\mathbf{778.44}$ m).

To evaluate the models' performance across 48-240 lead times, we calculate and plot the average RMSE across folds for each model per time lead as shown in the Figure 4.10. PiggyCast's dominance is observed

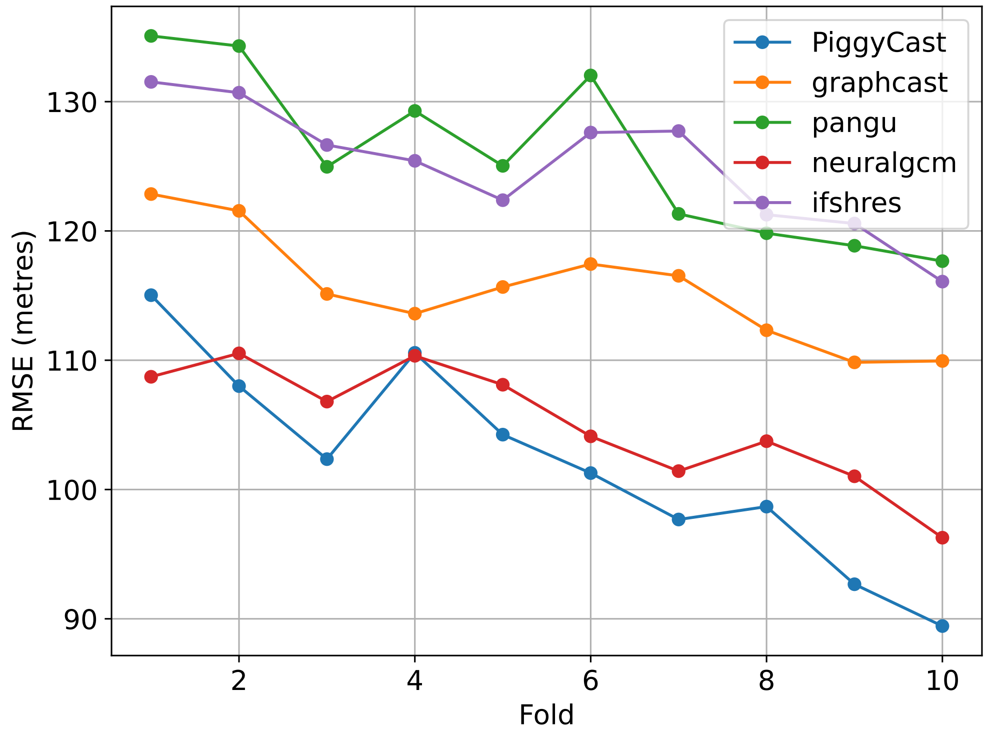

Figure 4.8: A plot of area-weighted RMSE per fold and model for 72-hour lead time.

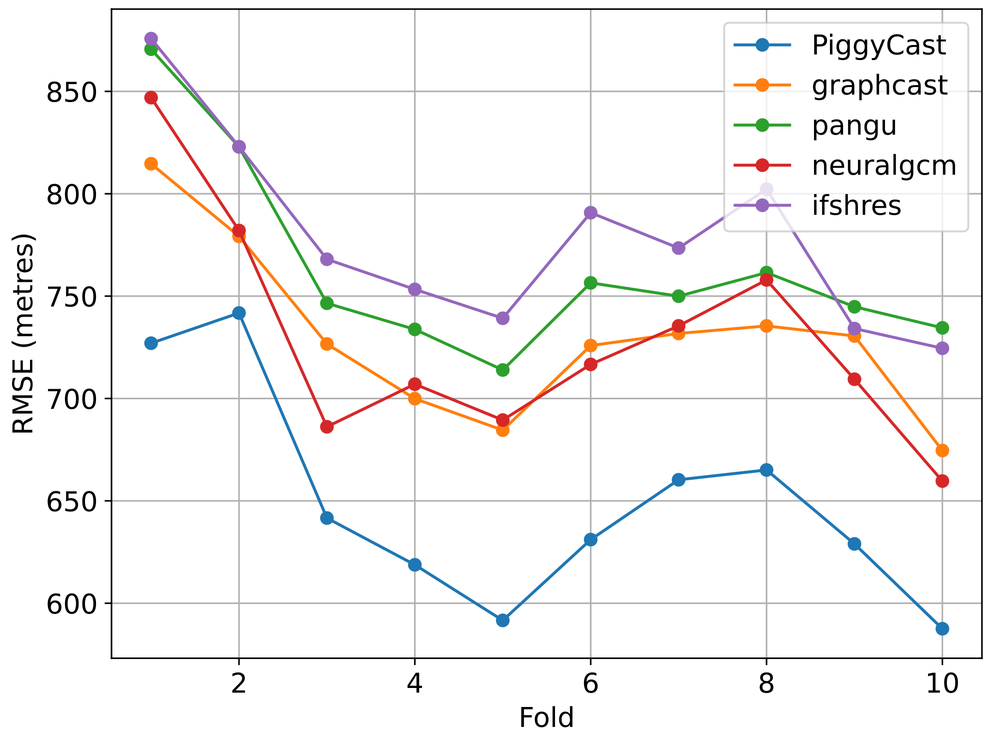

Figure 4.9: A plot of area-weighted RMSE per fold and model for 240-hour lead time.

as the lead time increases. NeuralGCM starts better but converges with GraphCast from 192-240 hours lead times. Similarly, IFS HRES and Pangu performance are similar but distinguishable from 120-240 hours lead times. A summary of the models' performance across all the lead times can be referenced in the Appendix C Table C.1.

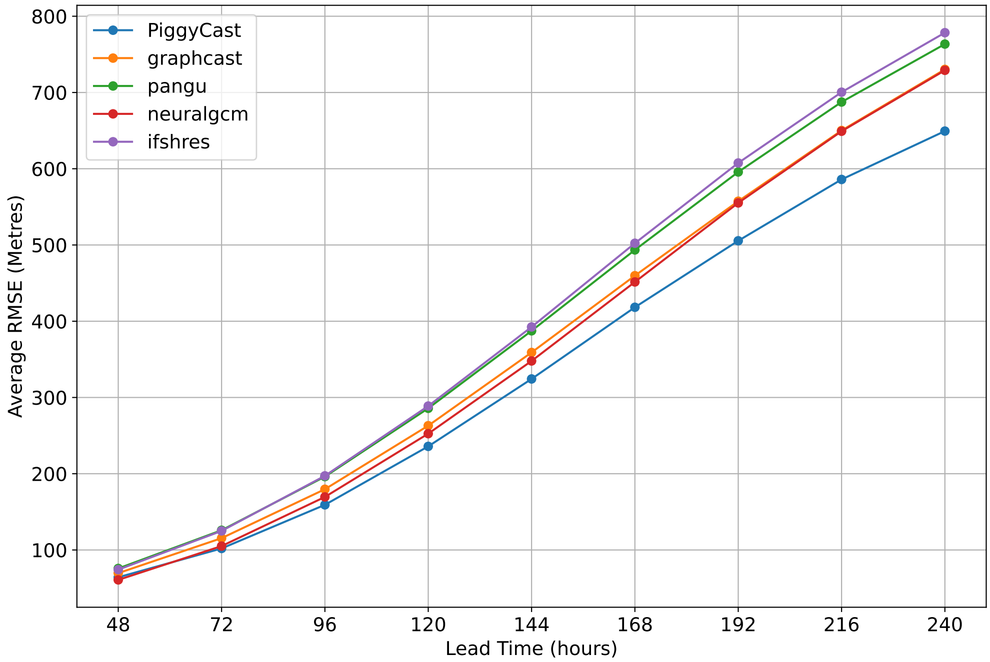

Figure 4.10: A plot of average area-weighted RMSE for all models, 48-240 hours lead times.

PiggyCast's Evaluation: Area-Weighted RMSE Percentage Improvement over IFS HRES

We also assessed the models' percentage RMSE improvement over IFS HRES across 48-240 hours lead time as shown in Figure 4.11. This helps paint a picture of the improvement of PiggyCast, as well as the base models, over IFS HRES, which is the gold-standard benchmark (Rasp et al., 2024) in WB2.

We observe that it is only at the 48-hour lead time that PiggyCast's percentage RMSE improvement ($\mathbf{13.49\%}$) comes second to any base model (here NeuralGCM leads with $\mathbf{18.10\%}$). PiggyCast's percentage RMSE improvement is highest at 96-hour lead time with $\mathbf{19.30\%}$ while lowest at 48-hour lead time with $\mathbf{13.49\%}$.

Interestingly, PiggyCast's percentage improvement remains fairly above $\mathbf{16\%}$ as the lead time increases to 240 hours, while the base models drastically decrease with an increase in lead time. At a 240-hour lead time, PiggyCast's percentage improvement is $\mathbf{16.58\%}$ while NeuralGCM $\mathbf{6.34\%}$, GraphCast $\mathbf{6.18\%}$, and lastly, Pangu $\mathbf{1.92\%}$. The summary table of percentage RMSE improvement over IFS HRES for all lead times can be found in Appendix C Table C.2.

Model Explainability using SHAPley values

We attempt to gain insight into PiggyCast's decision-making process and explain the contribution of features of the model by using SHAPley (SHAP) values. SHAP values help attribute the marginal contribution of every feature towards the prediction at local and global levels. The Figure 4.12 visualises the local contribution of each feature by either decreasing (negative SHAP values in blue) or increasing (positive SHAP values in red) the base value of each prediction for a 72-hour lead time. From the

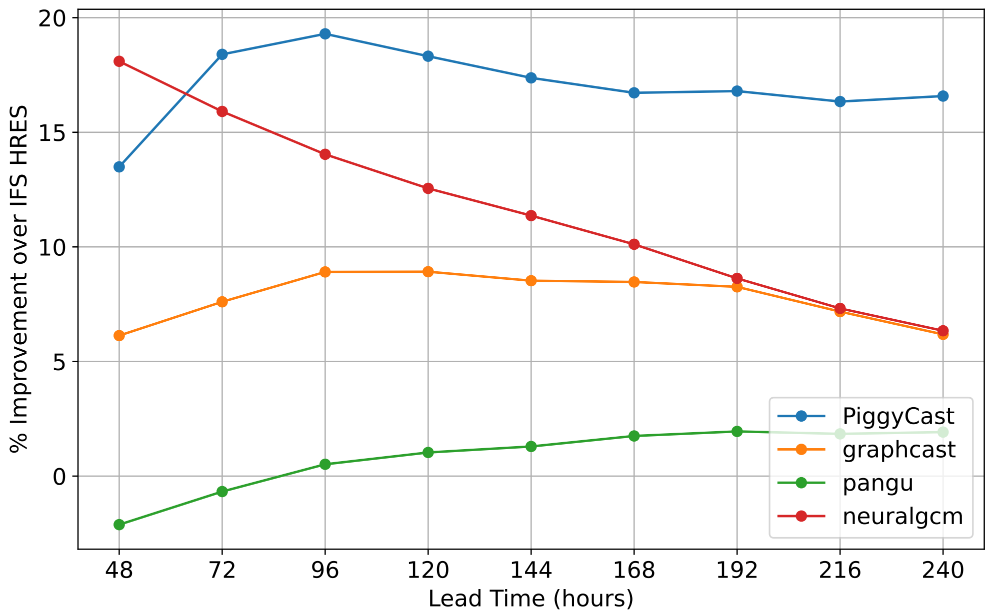

Figure 4.11: A plot of average area-weighted RMSE percentage improvement of models over IFS HRES over 48-240 hours lead time.

beeswarm plot, we observe that NeuralGCM, GraphCast and IFS HRES greatly influence PiggyCast's decision making at each prediction as seen with long and densely populated tails on either side of the zero divide. Pangu, longitude and latitude contribute the least to the prediction of the model, as seen with the short tails that are centred at zero.

For global feature contribution to PiggyCast's prediction, the Figure 4.13 helps ascertain the overall global contribution of each feature by using the mean absolute SHAP values at 72-hour lead time. From this bar plot, NeuralGCM leads with $\mathbf{1441.26}$ followed by GraphCast $\mathbf{959.16}$, IFS HRES $\mathbf{563.75}$, Pangu $\mathbf{221.79}$, longitude $\mathbf{7.88}$ and latitude $\mathbf{7.81}$.

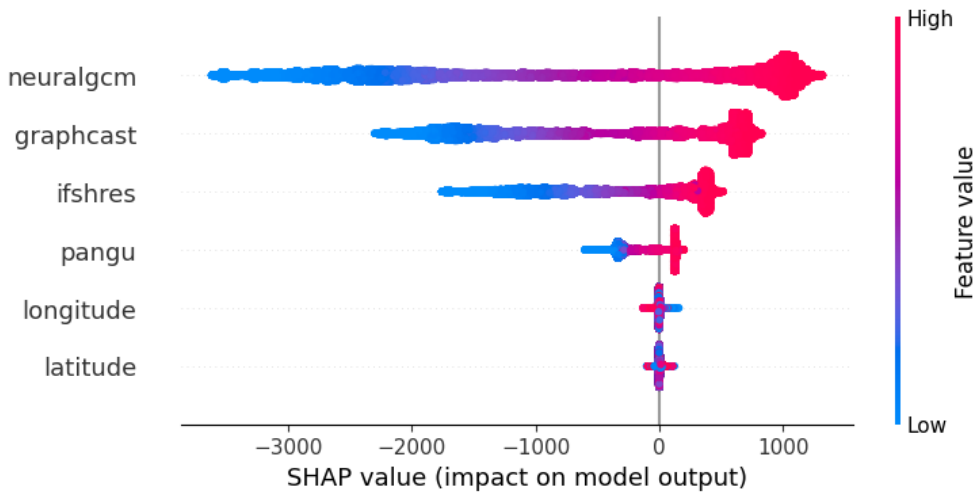

Figure 4.12: A beeswarm plot of SHAPley values of PiggyCast's features at 72-hour lead time.

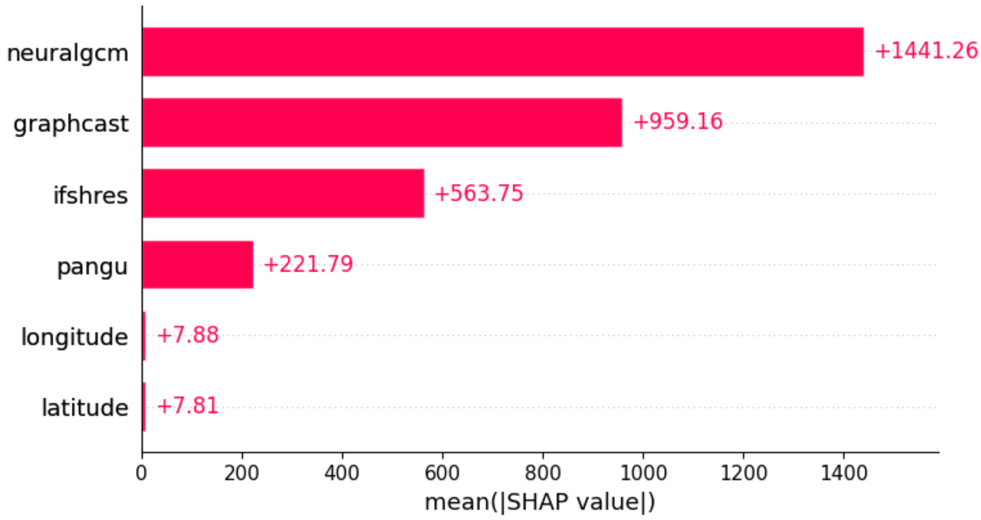

Figure 4.13: A bar plot of mean absolute SHAPley values of PiggyCast's features at 72-hour lead time.

Additionally, it is only at a 192-hour lead time do we notice GraphCast outcompete NeuralGCM in marginal contribution with a mean absolute SHAPley value of $\mathbf{1075.74}$ while NeuralGCM with $\mathbf{899.96}$. The contribution of spatial features (latitude and longitude) increases with the increase in lead time, with latitude contributing more than longitude. These additional plots can be found in the Appendix C.

4.3 Discussion of Results

The results obtained from the MDS and hierarchical clustering dendrogram analyses reveal critical insights into the performance and interrelationships of AI-based, numerical, and hybrid weather prediction models. These findings are consistent with, and in some cases extend, previous work such as Rasp (2024), reinforcing confidence in our methodology and results.

From the pairwise area-weighted RMSE matrix (Figure 4.1) and subsequent MDS visualisation (Figure 4.2), NeuralGCM is shown to be most similar to the ERA5 reanalysis dataset, reflecting its superior accuracy over all other models at the 72-hour lead time. The proximity of GraphCast and Pangu to NeuralGCM suggests shared forecasting characteristics among AI-based models, but the noticeable separation from IFS HRES indicates structural differences, likely due to IFS HRES's reliance on traditional NWP frameworks.

Our hierarchical clustering dendrogram analysis (Figures 4.4 and 4.5) reinforces these findings, with early merges between NeuralGCM, GraphCast, and Pangu, indicating high similarity in their error patterns, and a late merge for IFS HRES, highlighting its distinctiveness. This pattern is consistent with the growing body of work suggesting that hybrid and AI-based models are beginning to rival, and in some cases surpass, traditional NWP systems in operational skill, especially in the medium range (Rasp, 2024; Ben Bouallègue et al., 2023).

In addition to these comparative baselines, our proposed model, **PiggyCast**, demonstrates superior performance across 8 lead times (72, 96, 120, 144, 168, 192, 216, 240) and is only second to NeuralGCM at 48-hour lead time for 500 hPa geopotential height forecasts (see Figure 4.10).

Specifically, at 72-hour lead time (Figure 4.8), on average (across folds), PiggyCast's RMSE leads with

101.99 m, followed by NeuralGCM (**105.11** m), GraphCast (**115.48** m), IFS HRES (**124.99** m) and finally Pangu (**125.84** m). The RMSE values for our base models are similar to those reported on WB2 (Rasp et al., 2024), further validating our evaluation and results.

PiggyCast's performance can be attributed not only to our use of the XGBoost Regressor coupled with a timeseries-aware cross-validation (rolling-origin) evaluation strategy but also to the high-performing and diverse base models we piggyback off. Our design of PiggyCast, with interpretability in mind, offers actionable insights into which features most influence our prediction by employing SHAP values.

As observed through the beeswarm and bar plot of SHAP values at 72-hour lead time (Figures 4.12, 4.13), NeuralGCM, GraphCast, IFS HRES and Pangu contribute significantly to the model prediction while longitude and latitude contribute the least. However, with the increase in lead times, the contribution of longitude and latitude increases. This reflects the role of spatial anchoring over long lead times due to error propagation of atmospheric variables over time. This observation aligns with broader findings in AI and NWP literature, which emphasise the need for spatially aware model architectures (Magnusson et al., 2024; Silva et al., 2022; Wang et al., 2022).

5. Conclusion and Future Work

This final chapter summarises the work done and the future work beyond the scope of this master's.

5.1 Conclusion

This study set out to advance the understanding and practical application of both supervised and unsupervised machine learning techniques in operational weather forecasting and meteorology. Drawing on high-quality forecast datasets publicly available on the WB2 benchmarking framework (Rasp et al., 2024), the study employed these machine learning approaches to analyse the interdependency of NWP, AI-based and hybrid weather prediction models, develop an ensemble model through stacking of forecasts of these base models and finally perform feature attribution to the forecasting process of the trained ensemble machine learning model for interpretability and explainability.

The data foundation of this research was built on using ERA5 reanalysis and forecast datasets of IFS HRES (ECMWF, 2023), NeuralGCM (Kochkov et al., 2023), GraphCast (Lam et al., 2023) and Pangu-weather (Bi et al., 2022) weather prediction models. Geopotential height at 500 hPa pressure was the variable of study, which is critical in evaluating weather prediction models due to its physical, practical and historical significance in meteorology (Rasp et al., 2020; Zhou et al., 2007).

In the unsupervised learning phase, the MDS dimensionality reduction technique was employed to visualise the RMSE patterns in the forecast datasets and ERA5 reanalysis in a 2-dimensional MDS space. Hierarchical clustering using dendrograms further aided the visualisation and interpretation of error relationships of the models compared to ERA5 and to each other.

A key novel contribution of this thesis to operational weather forecasting was the supervised learning component; an ensemble machine learning model, named **PiggyCast**, was developed through stacking the forecasts of the above base models and the spatial location of each forecast. The improvement in PiggyCast's RMSE relative to the base models was notable, with an increase in performance across nine different lead times.

To aid in interpreting and explaining PiggyCast's forecasting process, feature attribution was done using SHAP values. This analysis provided transparency into the ensemble model's decision-making process, highlighting which input forecasts and spatial features contributed most significantly to error reduction at different lead times.

5.2 Future Work

The findings of this research highlight the value of integrating diverse modelling paradigms and employing advanced machine learning techniques for weather prediction. Nonetheless, several avenues remain open for future exploration. These include:

- Incorporation of additional weather variables and weather forecasting 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.

Acknowledgements

I would like to thank AIMS and its sponsors for their support of this study, as well as my supervisor, Dr. Oliver Angélil, and co-supervisor, Chris Toumping Fotso, from Ishango.ai, for their unwavering expertise and guidance through the course of this research. To my beloved mother, who did not witness the completion of this study and master's, I dedicate this work. To my lovely fiancé and brothers, thank you for your motivation and drive to believe in the beauty of my dreams. God has always been with me throughout this project. He has done exceedingly well and beyond what I prayed for.

References

Alobaidy, A. H., Al-Saadi, A. S., and Al-Ani, A. F. Evaluation of geopotential height at 500 hpa with rainfall events: A case study of iraq. *Al-Mustansiriyah Journal of Science*, 33(4):1–8, 2022. doi: 10.23851/mjs.v33i4.1161. URL https://mjs.uomustansiriyah.edu.iq/index.php/MJS/article/view/1161.

American Meteorological Society. Level of nondivergence. https://glossary.ametsoc.org/wiki/Level_of_nondivergence, 2022. Accessed 2025-05-21.

Andrews, D. G. *An Introduction to Atmospheric Physics*. Cambridge University Press, 2nd edition, 2010.

Ben Bouallègue, Z., Magnusson, L., Rodwell, M. J., Rasp, S., and Dueben, P. D. Hybrid forecasting: blending climate predictions with ai models. *Hydrology and Earth System Sciences*, 27(9):1865–1882, 2023. doi: 10.5194/hess-27-1865-2023. URL https://hess.copernicus.org/articles/27/1865/2023/.

Bergmeir, C. and Benítez, J. M. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Communications in Statistics—Simulation and Computation*, 47(5):1329–1343, 2018.

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

Bluestein, H. B. *Synoptic-Dynamic Meteorology in Midlatitudes, Volume I: Principles of Kinematics and Dynamics*. Oxford University Press, 1992.

Boer, G. J. Second-order statistics from the ECMWF ensemble prediction system. *Tellus A: Dynamic Meteorology and Oceanography*, 36(3):239–260, 1984. doi: 10.3402/tellusa.v36i3.11520.

Borg, I. and Groenen, P. J. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media, 2005.

Buontempo, C., Burgess, S. N., Dee, D., Pinty, B., Thépaut, J.-N., Rixen, M., Almond, S., Armstrong, D., Brookshaw, A., Alos, A. L., et al. The copernicus climate change service: climate science in action. *Bulletin of the American Meteorological Society*, 103(12):E2669–E2687, 2022.

Cambridge University Press. piggyback. https://dictionary.cambridge.org/dictionary/english/piggyback, 2025. Definition: "to use something that already exists or has already been done successfully to do something else quickly or effectively." Accessed 2025-05-23.

Cerqueira, V., Torgo, L., and Mozetič, I. Evaluating time series forecasting models: An empirical comparison on performance estimation methods. *Machine Learning*, 109:1997–2028, 2020. doi: 10.1007/s10994-020-05910-7.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *arXiv preprint arXiv:2306.12873*, 2023.

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

Copernicus Climate Change Service (C3S). Copernicus climate data store (cds). https://cds.climate.copernicus.eu/, 2025. Access to climate datasets, tools, and applications provided by the Copernicus Climate Change Service. Accessed 2025-05-13.

Cox, T. and Cox, M. *Multidimensional Scaling*. Chapman and Hall/CRC, 2 edition, 2000. ISBN 9781584880943. doi: 10.1201/9780367801700.

De Leeuw, J. and Heiser, W. J. Applications of the theory of majorization to multidimensional scaling: The smacof algorithm. *Psychometrika*, 42(1):85–93, 1977.

Dueben, P. D., Rasp, S., Fuhrer, O., Churazov, D. S., and Koldunov, N. V. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 14:2149–2167, 2021. doi: 10.5194/gmd-14-2149-2021.

ECMWF. Ifs documentation, 2023. https://www.ecmwf.int/en/publications/ifs-documentation.

ECMWF. Ecmwf's ai forecasts become operational. https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational, 2025. Accessed 2025-05-31.

Google Cloud and ECMWF. Weatherbench 2 dataset. https://console.cloud.google.com/storage/browser/weatherbench2, 2023. Public dataset containing weather forecast benchmarks and ERA5 reanalysis data. Includes multiple resolutions (0.25° to 1.5°). License details available in each subdirectory.

Gowan, T. A. *Data Analytics Applied to Satellite-Derived Precipitation Estimates and High-Resolution Model Output*. PhD thesis, The University of Utah, 2021.

Gowan, T. A., Horel, J. D., Jacques, A. A., and Kovac, A. Using cloud computing to analyze model output archived in zarr format. *Journal of Atmospheric and Oceanic Technology*, 39(4):449–462, 2022.

Gu, J., Liu, S., Zhou, Z., Chalov, S. R., and Zhuang, Q. A stacking ensemble learning model for monthly rainfall prediction in the taihu basin, china. *Water*, 14(3):492, 2022.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. ISBN 9780387848846.

Hastings, D. A. and Dunbar, P. Area-weighted statistics for climate data. *International Journal of Climatology*, 19(12):1335–1349, 1999.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020. URL https://doi.org/10.1002/qj.3803.

Holton, J. R. *An Introduction to Dynamic Meteorology*. Elsevier Academic Press, Amsterdam, 4th edition, 2004.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K.-L., Joyce, R. J., Kidd, C., Nelkin, E. J., Sorooshian, S., Stocker, E. F., Tan, J., et al. Integrated multi-satellite retrievals for the global precipitation measurement (gpm) mission (imerg). *Satellite precipitation measurement: Volume 1*, pages 343–353, 2020.

Hyndman, R. J. and Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 3 edition, 2021. URL https://otexts.com/fpp3/.

Islam, M. R., Ahmed, M. U., and Begum, S. ixgb: Improving the interpretability of xgboost using decision rules and counterfactuals. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, page 124740, 2024. doi: 10.5220/0012474000003636. URL https://www.scitepress.org/Papers/2024/124740/124740.pdf.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 2013. ISBN 9781461471370.

Judd, K. and Smith, L. A. Forecast verification of 500-hpa geopotential height anomalies by spectral decomposition. *Monthly Weather Review*, 136(10):3841–3857, 2008.

Kasahara, A. and Washington, W. M. Evaluation of the 500 mb height forecasts produced by the nmc medium-range forecast model. *Monthly Weather Review*, 113(6):1065–1079, 1985.

Keisler, R. Forecasting global weather with graph neural networks, 2022. URL https://arxiv.org/abs/2202.07575.

Kieu, C. Predictability of global ai weather models. *arXiv preprint arXiv:2410.03266*, 2024.

Kochkov, D. et al. Neuralgcm: A hybrid physics-ml general circulation model. *arXiv preprint arXiv:2305.08891*, 2023.

Krishnamurthy, V. Predictability of weather and climate. *Proceedings of the National Academy of Sciences of the United States of America*, 116(38):18617–18625, 2019. doi: 10.1073/pnas.1907917116. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6774281/.

Kruskal, J. B. and Wish, M. *Multidimensional Scaling*. Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills, CA, 1978. ISBN 9780803909403.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Lawrence, Z. D., Abalos, M., Ayarzagüena, B., Barriopedro, D., Butler, A. H., Calvo, N., de la Cámara, A., Charlton-Perez, A., Domeisen, D. I., Dunn-Sigouin, E., et al. Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems. *Weather and Climate Dynamics Discussions*, 2022:1–37, 2022.

Leon, J. Scale-dependent verification of precipitation and cloudiness at ecmwf. *Newsletter no*, 174: 18–22, 2023.

Lerch, S., Mayer, M. J., Demaeyer, J., et al. Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. *AI for the Earth Systems*, 3(1):1–19, 2024. doi: 10.1175/AIES-D-23-0070.1.

Li, F.-F., Johnson, J., and Yeung, S. Cs231n: Convolutional neural networks for visual recognition, lecture 11: Generative models and dimensionality reduction. https://cs231n.stanford.edu/slides/2019/cs231n_2019_lecture11.pdf, 2019. Stanford University course notes.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12509–12519, 2022. doi: 10.48550/arXiv.2111.09883. URL https://arxiv.org/abs/2111.09883.

Lorenz, E. N. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lorenz, E. N. Atmospheric predictability experiments with a large numerical model. *Tellus A*, 34(6): 505–513, 1982.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Lundberg, S. M., Lee, S.-I., and Contributors, S. Shap python package documentation. https://shap.readthedocs.io/en/latest/generated/shap.TreeExplainer.html, 2024. Accessed 2025-06-01.

Magnusson, L., Ben Bouallègue, Z., Rasp, S., Dueben, P. D., and Rodwell, M. J. The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based prediction systems. *Bulletin of the American Meteorological Society*, 105(6):E817–E835, 2024. doi: 10.1175/BAMS-D-23-0162.1. URL https://journals.ametsoc.org/view/journals/bams/105/6/BAMS-D-23-0162.1.xml.

Met Office. Artificial intelligence for numerical weather prediction. https://www.metoffice.gov.uk/research/approach/collaboration/artificial-intelligence-for-numerical-weather-prediction, 2025. Accessed 2025-05-31.

Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently Published, 2 edition, 2024. URL https://christophm.github.io/interpretable-ml-book/shap.html. Chapter 18: SHAP values, including theoretical foundations and practical applications.

Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

NOAA NCEP. National centers for environmental prediction. https://www.weather.gov/ncep/, 2025. Accessed: 2025-06-13.

Omta, A. W. and Larsen, D. The geoscience libretexts library: An interactive learning platform for instructors and students. In *AGU Fall Meeting Abstracts*, volume 2018, pages ED51I–0735, 2018.

Pandya, S. and Guha Thakurta, R. Hands-on infrastructure as code with hashicorp terraform. In *Introduction to Infrastructure as Code: A Brief Guide to the Future of DevOps*, pages 99–133. Springer, 2022.

Pasch, R. J., Berg, R., Roberts, D. P., and Papin, P. P. Tropical cyclone report: Hurricane laura (al132020), 20–29 august 2020. Technical Report AL132020, National Hurricane Center, 2021. URL https://www.nhc.noaa.gov/data/tcr/AL132020_Laura.pdf. Accessed 2025-05-16.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. Scikit-learn: Machine learning in Python – mds. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html, 2011.

Perez, J. D. S., Osaba, E., Molina, D., et al. A comparative analysis of xgboost. *arXiv preprint arXiv:1911.01914*, 2019. URL https://arxiv.org/pdf/1911.01914.pdf.

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.03409.

Rackow, T., Pedruzo-Bagazgoitia, X., Becker, T., Milinski, S., Sandu, I., Aguridan, R., Bechtold, P., Beyer, S., Bidlot, J., Boussetta, S., Deconinck, W., Diamantakis, M., Dueben, P., Dutra, E., Forbes, R., Ghosh, R., Goessling, H. F., Hadade, I., Hegewald, J., Jung, T., Keeley, S., Kluft, L., Koldunov, N., Koldunov, A., Kölling, T., Kousal, J., Kühnlein, C., Maciel, P., Mogensen, K., Quintino, T., Polichtchouk, I., Reuter, B., Sármány, D., Scholz, P., Sidorenko, D., Streffing, J., Sützl, B., Takasuka, D., Tietsche, S., Valentini, M., Vannière, B., Wedi, N., Zampieri, L., and Ziemen, F. Multi-year simulations at kilometre scale with the integrated forecasting system coupled to fesom2.5 and nemov3.4. *Geoscientific Model Development*, 18(1):33–69, 2025. doi: 10.5194/gmd-18-33-2025. URL https://gmd.copernicus.org/articles/18/33/2025/.

Rasp, S. Ai-weather sota vs time. figshare. Dataset, 2024. URL https://doi.org/10.6084/m9.figshare.28083515.v1. Public spreadsheet tracking state-of-the-art AI weather prediction models over time. Accessed 2025-05-19.

Rasp, S. and Thuerey, N. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405, 2021. doi: 10.1029/2020MS002405.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2024. URL https://arxiv.org/abs/2308.15560.

Research, G. and ECMWF. Weatherbench 2: Cloud-optimized zarr datasets. https://console.cloud.google.com/storage/browser/weatherbench2, 2023. Cloud-optimized Zarr format datasets for global weather benchmarking at multiple resolutions.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. Cross-validation strategies for spatial and spatiotemporal data. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Scher, S. *Artificial intelligence in weather and climate prediction*. PhD thesis, Stockholm University, 2020. URL https://www.diva-portal.org/smash/get/diva2:1425352/FULLTEXT01.pdf.

scikit-learn developers. *Time Series Split - scikit-learn 1.6 documentation*, 2025. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.

Shih, A., Sadigh, D., and Ermon, S. Training and inference on any-order autoregressive models the right way. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2205.13554. URL https://arxiv.org/abs/2205.13554.

Silva, T., Gentine, P., Reichstein, M., Koster, R., Dirmeyer, P., Qu, X., Tuttle, S., et al. Explainable machine learning for lightning prediction in an earth system model. *npj Climate and Atmospheric Science*, 5(1):1–10, 2022. doi: 10.1038/s41612-022-00293-2.

Soviany, P., Ionescu, R. T., Rota, P., and Sebe, N. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1426–1468, 2022. doi: 10.1007/s11263-021-01561-5. URL https://arxiv.org/abs/2101.10382.

Stull, R. B. 1.07: Atmospheric structure. in practical meteorology: An algebra-based survey of atmospheric science. https://geo.libretexts.org/Bookshelves/Meteorology_and_Climate_Science/Practical_Meteorology_(Stull)/01:_Atmospheric_Basics/1.07:_Atmospheric_Structure, 2017. LibreTexts. Accessed 2025-05-21.

Sun, S., Li, L., Zhao, B., Ma, M., Zhang, J., Liu, Y., and Zhang, Y. Multiscale feature analysis of forecast errors of 500hpa geopotential height for the CMA-GFS model. *Atmospheric Science Letters*, 24(10):e1174, 2023. doi: 10.1002/asl.1174. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.1174.

The SciPy community. *Hierarchical clustering (scipy.cluster.hierarchy)*, 2025. https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html.

Tompkins, A. *Moist physical processes in the IFS: Progress and Plans*. ECMWF, 2004.

Vallis, G. K. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, 2nd edition, 2017.

Wallace, J. M. and Hobbs, P. V. *Atmospheric Science: An Introductory Survey*. Academic Press, 2nd edition, 2006.

Wang, X., Li, M., Chen, Y., Yang, X., et al. Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems. *Frontiers in Environmental Science*, 10:1057081, 2022. doi: 10.3389/fenvs.2022.1057081.

Watt, T., Dueben, P. D., Rasp, S., Ben Bouallègue, Z., Magnusson, L., Rodwell, M. J., and Weyn, J. A. Do data-driven models beat numerical models in forecasting weather and extremes? *Geoscientific Model Development*, 17(22):7915–7937, 2024. doi: 10.5194/gmd-17-7915-2024. URL https://gmd.copernicus.org/articles/17/7915/2024/.

Watt, T., Rasp, S., Dueben, P. D., Ben Bouallègue, Z., Magnusson, L., Rodwell, M. J., and Weyn, J. A. An extension of the weatherbench 2 to binary hydroclimatic forecasts. *EGUsphere [preprint]*, 2025. doi: 10.5194/egusphere-2025-3. URL https://egusphere.copernicus.org/preprints/2025/egusphere-2025-3/.

Weather Atlas. Geopotential height | weather atlas. https://www.weather-atlas.com/g/geopotential-height, 2023. Accessed 2025-05-21.

WeatherBench 2 Contributors. *WeatherBench 2 Evaluation Quickstart*, 2024. URL https://weatherbench2.readthedocs.io/en/latest/evaluation.html. Online documentation. Accessed 2025-05-13.

Wikipedia contributors. Geopotential height. https://en.wikipedia.org/wiki/Geopotential_height, 2025. Accessed 2025-05-21.

Wilks, D. S. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 3rd edition, 2011.

XGBoostDevelopers. *XGBoost Documentation*, 2025. URL https://xgboost.readthedocs.io/en/stable/parameter.html. Accessed 2025-06-01.

Zhou, T., Zhang, X., Zheng, X., and Frederiksen, C. S. Statistical prediction of seasonal mean southern hemisphere 500-hpa geopotential height anomalies. *Journal of Climate*, 20(12):2812–2828, 2007. doi: 10.1175/JCLI4180.1.

AppendixA. Geopotential Height

A.1 Mathematical Formulation

Gravitational Potential Energy

Gravitational potential energy U of a mass m at a height z in Earth's gravitational field is defined as:

$$U = mgz \tag{A.1.1}$$

where:

- g is the local acceleration due to gravity ($\approx 9.81\,\mathrm{m/s^2}$ near the surface),
- z is the geometric height above sea level.

This is the classical definition of potential energy in a gravitational field (Andrews, 2010).

Geopotential

The **geopotential** $\Phi(z)$ is defined as the gravitational potential energy per unit mass:

$$\Phi(z) = \frac{U}{m} = gz \tag{A.1.2}$$

This represents the amount of work required to raise a unit mass from sea level to a height z against gravity (Wallace and Hobbs, 2006).

Geopotential Height

Because gravity varies, albeit slightly, with latitude and altitude, meteorology uses the **geopotential height** Z instead of geometric height, defined as:

$$Z = \frac{1}{g_0} \int_0^z g(z)\, dz' \tag{A.1.3}$$

where:

- $g(z)$ is the gravitational acceleration as a function of height,
- g_0 is the standard gravity ($9.80665\,\mathrm{m/s^2}$),
- Z is the geopotential height, expressed in meters.

If gravity is assumed constant (a good approximation for most applications), then:

$$Z \approx \frac{g}{g_0} z \tag{A.1.4}$$

and if $g \approx g_0$, then:

$$Z \approx z \tag{A.1.5}$$

This concept of geopotential height is fundamental in atmospheric science and meteorology, especially for analysing synoptic-scale features such as pressure systems and jet streams (Holton, 2004; Vallis, 2017).

AppendixB. Additional Forecast Models

B.1 Keisler

Keisler (Keisler, 2022) is an AI weather prediction model that is based on the Graph Neural Networks (GNNs) architecture developed by Ryan Keisler. The GNN architecture here was composed of an Encoder, which mapped the latitude and longitude grid to an icosahedron grid, a Processor, performing the message-passing on the icosahedron grid, and a Decoder, which mapped back the icosahedron grid to the latitude and longitude grid. Through autoregression, the forecasts were made for approximately 6 days while still maintaining numerical stability despite not infusing any physics, such as conservation of momentum or stability training with additional noise (Keisler, 2022).

With a resolution of $1°$ longitude/latitude (360×181) on 13 pressure levels (50 hPa, 100 hPa, 150 hPa, 200 hPa, 250 hPa, 300 hPa, 400 hPa, 500 hPa, 600 hPa, 700 hPa, 850 hPa, 925 hPa and 1,000 hPa), the model is trained to output 6 physical variables: geopotential height (Z), temperature (T), specific humidity (Q), vertical wind component (W), and northward (V) and eastward (U) wind components. Similarly, Keisler's input variables comprise 6 atmospheric variables: geopotential height (Z), temperature (T), specific humidity (Q), vertical wind component (W), northward wind component (V) and eastward wind component (U) (Keisler, 2022).

Keisler model demonstrated skill comparable to operational NWP models, such as IFS HRES, on some deterministic upper-level metrics (Rasp et al., 2024), prompting the need for an updated benchmark of the original weather benchmark (Rasp et al., 2020). However, the Keisler model has a coarse resolution and is currently not operational (Keisler, 2022).

Keisler's forecasts are available on WB2 with evaluation done for the year 2020 using 00 and 12 Coordinated Universal Time (UTC) initialisation times Keisler (2022). In this study, we considered the Keisler forecasts for the 2020 time period at $5.625°$ longitude/latitude (64×32) spatial resolution with equiangular conservative remapping, 12-hourly temporal resolution and upper-air atmospheric geopotential variable at 500 hPa pressure level (same as the other models' forecasts and ERA5 reanalysis). The Table (B.1) summarises the details of the Keisler (2022) forecasts dataset used in this study.

Table B.1: Details of the WeatherBench 2 Keisler (2022) forecasts dataset used

Attribute	Value
Dataset Name	2020-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/keisler/
Data Source	Keisler (2022) WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

The Keisler model forecasts dataset was dropped in this study because, after exploratory data analysis, we noticed that there were missing values in the forecasts provided by WB2 (32,768 values for 16 timesteps over 64×32 grid resolution). A case in point was for 48 hours lead time, the missing data was in early April (2020-04-04 and 2020-04-05), early May (2020-05-02 and 2020-05-03) and late September (2020-09-23 to 2020-09-26). Refer to Figure B.1 for the missing values for 48 48-hour lead time.

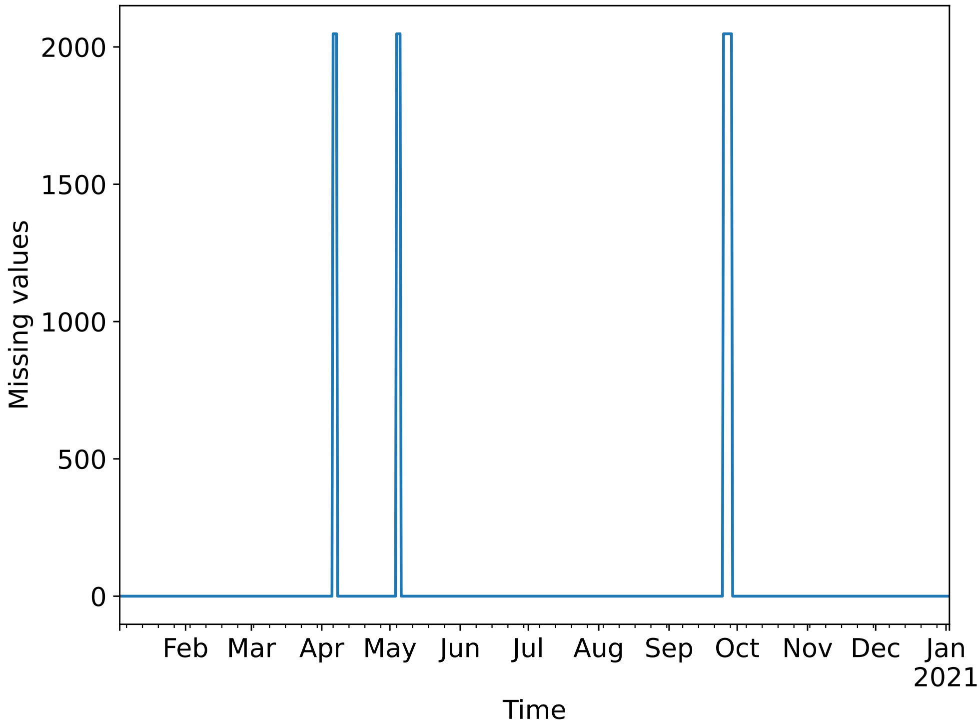

Figure B.1: Missing values over time of Keisler 2020 forecasts from WB2 for 48-hour lead time.

B.2 FuXi

FuXi is a cascade of three sequential artificial intelligence weather prediction (AIWP) models with a U-Transformer backbone for 0-5 days (FuXi-Short), 5-10 days (FuXi-Medium), and 10-15 days (FuXi-Long) forecasts developed by a research team from the Artificial Intelligence Innovation and Incubation Institute at Fudan University, Shanghai Qi Zhi Institute and contributions from Huawei Cloud (Chen et al., 2023).

The cascade architecture was developed to address the challenge of the accumulation of forecast errors as a result of using a single model over shorter and longer lead times. The output from one model is used as input for the next model; for example, the output from the 20th step (5th day forecast) of FuXi-Short is used as input for FuXi-Medium and similarly, the output from the 40th step (10th day forecast) of FuXi-Medium is used as input for FuXi-Long which goes ahead to the 15th day forecast (Chen et al., 2023).

The architecture of the pre-training FuXi base model consists of three main components: cube embedding, a U-Transformer and a fully connected layer. The model takes in two preceding time steps (X_{t-1} and X_t, where $t-1$ and t are the previous and current time steps, respectively) of the state of the weather, which is processed by the cube embedding layer to reduce temporal and spatial dimensions. The embedded data is then passed through the U-Transformer, which is based on the Swin Transformer

V2 architecture (Liu et al., 2022). Ultimately, the fully connected layer is used for prediction, and the output is scaled back to the original spatial and temporal dimensions (Chen et al., 2023). After pretraining, the base model is then fine-tuned for the specific forecast windows of FuXi-Short, FuXi-Medium, and FuXi-Long using an autoregressive training regime (Shih et al., 2022) and a curriculum training schedule (Soviany et al., 2022) similar to the fine-tuned GraphCast model (Chen et al., 2023).

At $0.25°$ longitude/latitude (1440×721) spatial grid resolution and 13 (50 hPa, 100 hPa, 150 hPa, 200 hPa, 250 hPa, 300 hPa, 400 hPa, 500 hPa, 600 hPa, 700 hPa, 850 hPa, 925 hPa, and 1000 hPa) pressure levels, FuXi model predicts and evaluates 5 atmospheric variables: geopotential height (Z), temperature (T), eastward component of wind (U), northward component of wind (V) and relative humidity (R) and 5 surface variables: 2-meter temperature (2T), 10-meter eastward wind component (10U), 10-meter northward wind component (10V), mean sea-level pressure (MSL) and 6-hourly total precipitation (TP) (Chen et al., 2023).

Fuxi ensemble forecast system (50-member ensemble generated through the introduction of Perlin noise for initial conditions and model parameters perturbation) revealed a state-of-the-art deterministic performance, which outperformed IFS HRES and had superior performance compared to GraphCast for longer lead times (Chen et al., 2023). However, the FuXi model's deterministic performance becomes slightly poorer than IFS HRES for forecasts beyond 9 days, while the ensemble performance is inferior to the ECMWF ensemble beyond 9 days.

In this study, we considered the FuXi model forecasts for 2020 at $5.625°$ longitude/latitude (64×32) spatial grid resolutions with equiangular conservative remapping, 12-hourly temporal resolution and upper-air atmospheric geopotential variable at 500 hPa. The Table (B.2) summarises the details of the FuXi forecasts dataset used in this study.

Table B.2: Details of the WeatherBench 2 FuXi forecasts dataset used.

Attribute	**Value**
Dataset Name	2020-64x32_equiangular_conservative.zarr
Dataset Location	gs://weatherbench2/datasets/fuxi/
Data Source	FuXi WeatherBench 2
Period	2020
Temporal Resolution	12-hourly
Spatial Resolution	64×32 (equiangular, conservative remapping)
Format	Zarr
Variables	Geopotential height
Vertical Levels	500 hpa pressure level
Use Case	WeatherBench 2 model evaluation (forecast)

Similarly, the Fuxi model forecast dataset was dropped because, after exploratory data analysis, we found out that the dataset was truncated at day '2020-12-16T00:00:00.000000000', hence insufficient for the forecast period in the study.

AppendixC. Source code and Additional Plots

C.1 Source Code

The source code of this research can be accessed on the Github repository AIMS_Masters_Thesis.

C.2 Additional Figures and Tables

Table C.1: Model RMSEs at Different Lead Times (Red values indicate the least RMSE in that lead time)

Lead Time (h)	PiggyCast	GraphCast	IFSHRes	NeuralGCM	Pangu
48	64.18	69.64	74.19	60.76	75.76
72	101.99	115.49	124.99	105.11	125.84
96	159.18	179.66	197.23	169.55	196.22
120	235.87	263.02	288.77	252.53	285.80
144	324.30	359.03	392.50	347.89	387.42
168	418.30	459.76	502.30	451.50	493.50
192	505.52	557.40	607.57	555.16	595.73
216	585.95	650.14	700.40	649.16	687.49
240	649.38	730.32	778.44	729.08	763.50

Table C.2: Percentage Improvement Over IFSHRes RMSE (Red values indicate the highest improvement in that lead time)

Lead Time (h)	PiggyCast	GraphCast	NeuralGCM	Pangu
48	13.49	6.13	18.10	-2.12
72	18.40	7.60	15.91	-0.67
96	19.30	8.91	14.04	0.52
120	18.32	8.92	12.55	1.03
144	17.37	8.53	11.37	1.29
168	16.72	8.47	10.11	1.75
192	16.80	8.26	8.63	1.95
216	16.34	7.18	7.32	1.84
240	16.58	6.18	6.34	1.92

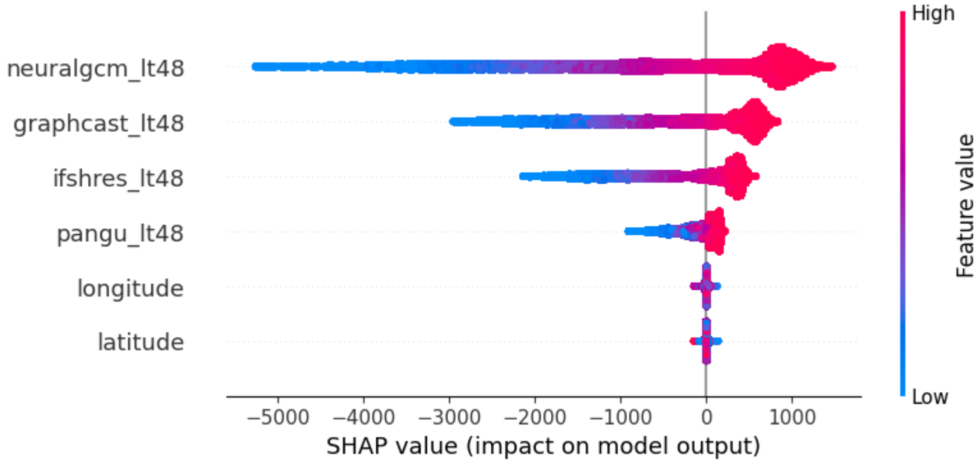

Figure C.1: A beeswarm plot of SHAPley values of PiggyCast's features at 48-hour lead time.

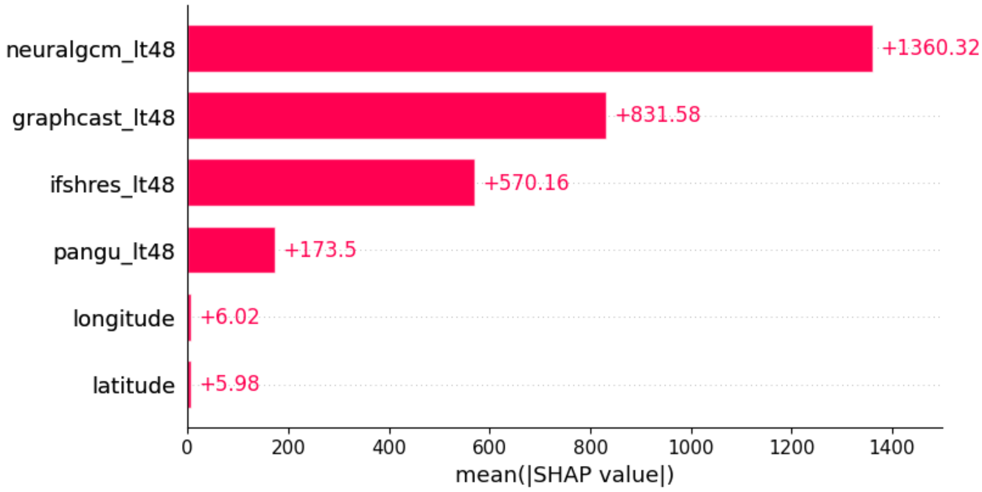

Figure C.2: A bar plot of mean absolute SHAPley values of PiggyCast's features at 48-hour lead time.

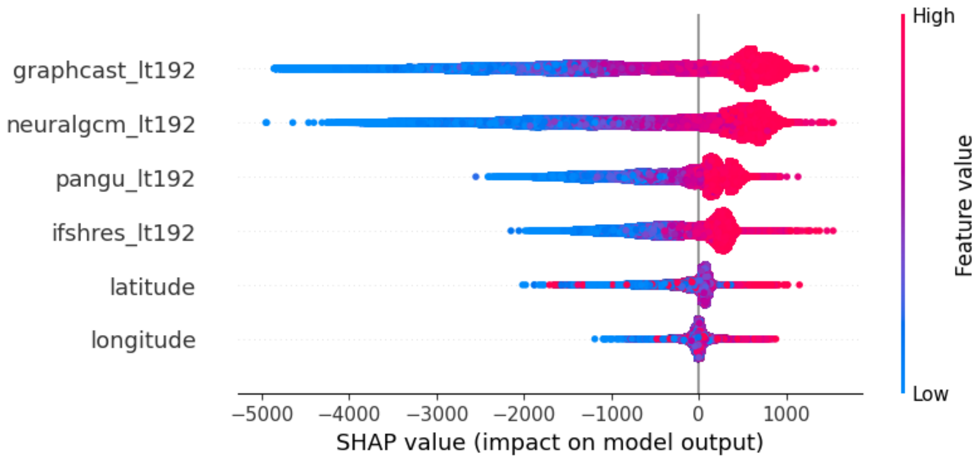

Figure C.3: A beeswarm plot of SHAPley values of PiggyCast's features at 192-hour lead time.

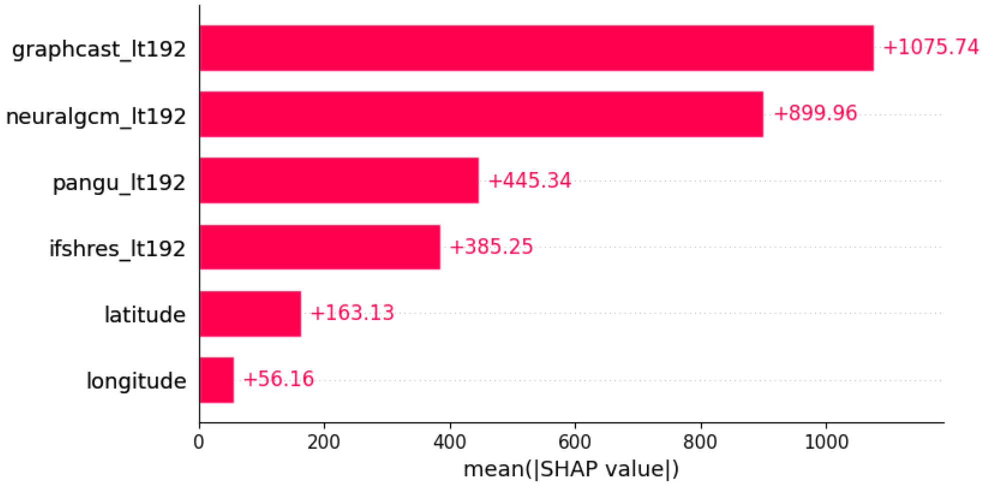

Figure C.4: A bar plot of mean absolute SHAPley values of PiggyCast's features at 192-hour lead time.

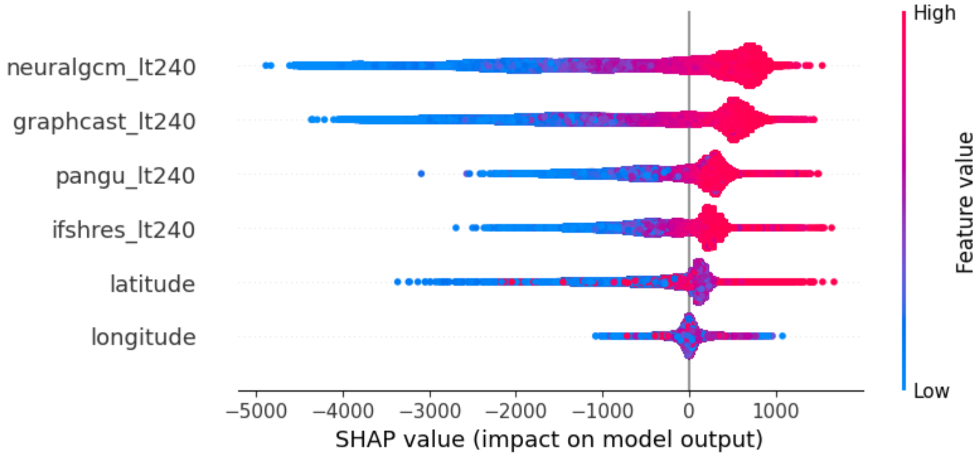

Figure C.5: A beeswarm plot of SHAPley values of PiggyCast's features at 240-hour lead time.

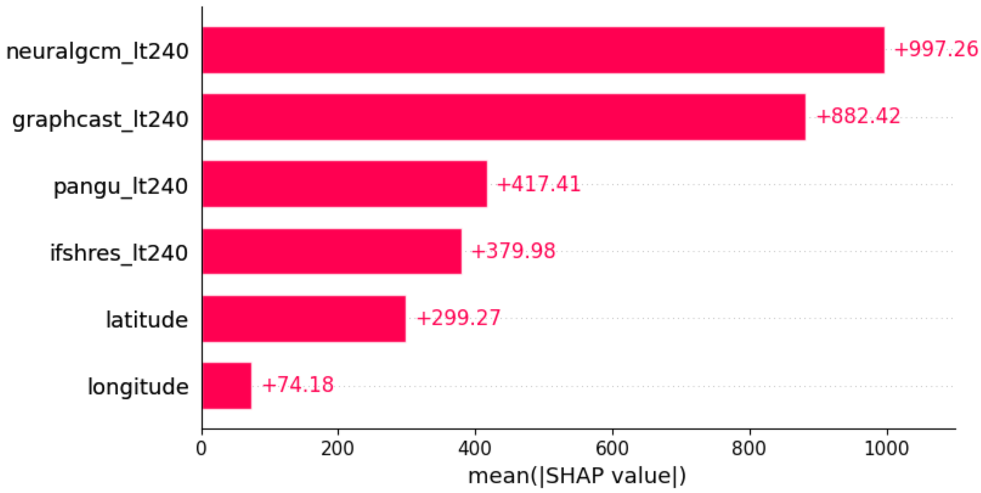

Figure C.6: A bar plot of mean absolute SHAPley values of PiggyCast's features at 240-hour lead time.