Australian Government
Bureau of Meteorology

Object-oriented verification of TC-Jasper rainfall forecasts: Machine learning model versus physical models

Hector Morisseau, Hongyan Zhu, Debra Hudson, Catherine de Burgh-Day

February 2025

Bureau Research Report - 106

Object-oriented verification of TC-Jasper rainfall forecasts: Machine learning model versus physical models

Hector Morisseau[2], Hongyan Zhu[1], Debra Hudson[1] and Catherine de Burgh-Day[1]

[1]Bureau of Meteorology, Australia

[2]Meteo France, France

Bureau Research Report No. 106

February 2025

National Library of Australia Cataloguing-in-Publication entry

Authors: Hector Morisseau, Hongyan Zhu, Debra Hudson, Catherine de Burgh-Day

Title: Object-oriented verification of TC-Jasper rainfall forecasts: Machine learning model versus physical models

ISBN: 978-1-925738-92-6

ISSN: 2206-3366

Series: Bureau Research Report – BRR106

Enquiries should be addressed to:

Corresponding Author: Hongyan Zhu

Bureau of Meteorology

GPO Box 1289, Melbourne

Victoria 3001, Australia

hongyan.zhu@bom.gov.au

Copyright and Disclaimer

Contents

List of Figures

Executive Summary

Forecasts of tropical cyclone Jasper, produced by a machine learning model (GraphCast) and global NWP physics-based models from the UK Met Office (UM GA8 and GA9) are compared. Evaluation of the rainfall forecasts are the focus of this work, given its significant impacts and the challenges associated with rainfall prediction for machine learning models. Forecasts are evaluated against multiple sources: the ERA5 and BARRA2 reanalyses, and GPM satellite rainfall data. The Contiguous Rain Area (CRA) method, which aims to objectively compare the properties of matched forecast and observed rain systems, was employed to evaluate model performance. The paper uses this case study to highlight some of the advantages and limitations of machine learning models relative to traditional physical models. For TC Jasper, GraphCast demonstrates comparable forecast skill for moderate rain when compared to GA8/GA9. However, it tends to underestimate intense rainfall, partly due to its coarser grid resolution. The study also shows the importance of careful consideration of observational/reanalysis data, as different datasets can lead to substantially different conclusions.

1. Introduction

Tropical cyclones (TCs) pose significant risks to lives and property, bringing heavy rainfall, strong winds, and storm surges that collectively cause extensive damage. Accurate forecasting of TC intensity, track, and rainfall is crucial for emergency management to mitigate impacts.

Physical models, grounded in the fundamental laws of physics, simulate the atmosphere's behaviour by solving complex mathematical equations. These models, such as the Unified Model (UM), have been the cornerstone of weather forecasting for decades. They provide detailed insights into the dynamics of tropical cyclones, including wind patterns, pressure systems, and thermodynamic processes. However, they often require substantial computational resources and can struggle with the rapid, non-linear changes characteristic of TCs.

On the other hand, machine learning (ML) models like GraphCast (Scarselli et al., 2009; Remi et al., 2023), PanguWeather (Bi et al. 2023), and Forecastle v2 (Bonev et al 2023), leverage vast datasets and advanced algorithms to identify patterns and relationships within historical and real-time data. These models excel in processing large volumes of information quickly and can adapt as new data becomes available. The current generation of ML models have shown promise in predicting the track of TCs, but tend to perform less well in predicting their intensity (e.g., Bonavita, 2023; Liu et al., 2024, Xie et al., 2024; DeMaria at al., 2024).

The comparison between ML models and physical (dynamical) models in TC forecasting represents a significant step forward in meteorological science. Both approaches offer unique strengths and face distinct challenges, making their comparison essential for advancing our understanding of prediction capabilities.

This study aims to compare the performance of an ML model, GraphCast, and physical models in forecasting TC Jasper, which impacted Australia in December 2023. In particular, we focus on systematic errors in location and intensity of the rainfall forecasts, identified using an object-oriented verification approach called the Contiguous Rain Area method (Ebert and McBride 2000). Although this is only a single case study, we can start getting insights into the relative strengths and weaknesses of each modelling approach. Unlike previous work that evaluated tropical cyclone track and intensity in ML models (e.g. DeMaria et. al 2024; Xie et. al 2024), this study focuses on the performance of rainfall forecasts.

(a) (b)

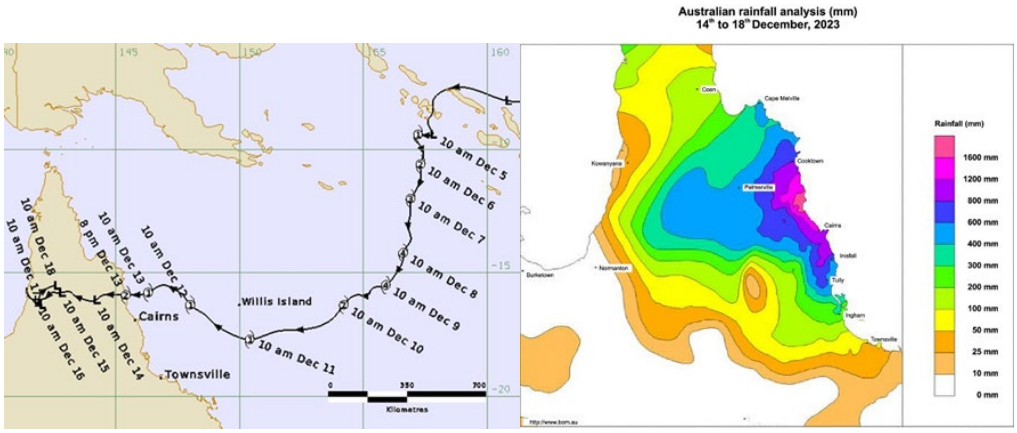

Figure 1: (a) Best track of Severe Tropical Cyclone Jasper, 2-17 December 2023 and (b) total rainfall (mm) recorded over the 5-d period (9am aligned) of December 14th to 18th, 2023, in Queensland, Australia (http://www.bom.gov.au/cyclone/history/jasper23.shtml).

TC Jasper developed over the northern Coral Sea on 5 December 2023. It moved southward and then southwest, intensifying to a Category 4 cyclone by 8 December. It weakened upon approaching the Queensland coast, making landfall near Bloomfield as a Category 2 storm on 13 December (Fig 1a). TC Jasper delivered very high rainfall along the far-north coast of Queensland after its landfall. It was one of the wettest tropical cyclones in Australian history with 5-day rainfall totals (14-18 December) exceeding 1 m over parts of Queensland (Fig. 1b) and more than 2 m at some stations (Pradad 2024). One of the reasons for the very high rainfall totals were that Jasper stalled over Cape York Peninsula before dissipating on 17 December. TC Jasper was responsible for unprecedented flooding and damage (https://www.abc.net.au/news/2023-12-18/ex-tropical-cyclone-jasper-floods-explained/103241528; https://www.abc.net.au/news/2023-12-17/qld-flooding-fnq-cyclone-jasper-sunday/103238806). Given the extreme nature of this TC-related rainfall event, it provides a good case study for examining the performance of an ML model in comparison to traditional physical models.

2. Models, data and analysis methods

2.1. Models

Three models were studied in this research: the GraphCast ML model and the Unified Model (UM) GA8 and GA9 global atmosphere (GA) physics-based models.

2.1.1. The GraphCast ML model

GraphCast (Remi et al., 2023) is a global atmosphere-only ML model developed by Google DeepMind, based on a Graph Neural Network (GNN, Scarselli et al., 2009) architecture. GraphCast demonstrates competitive skill compared to other leading ML models (Rasp et al., 2024) and has been the subject of numerous follow-on applications and studies (e.g., Olivetti & Messori, 2024; Subich, 2024; Yan et al., 2024; Colony & Andigani, 2024). For this work the operational version of GraphCast is used, which has 0.25º resolution and 13 pressure levels in the atmosphere. This version of the model is pre-trained on ERA5 from 1979 to 2017 and fine-tuned on European Centre for Medium-range Weather Forecasting (ECMWF) HRES data from 2016 to 2021. GraphCast was a natural choice for this study due to its competitive skill compared to other ML models, and the fact that it is one of the few ML models currently available to the wider research community which features precipitation as an output variable (in contrast to PanguWeather or FourCastNetV2 for example, which do not).

2.1.2. The UM physics-based models: GA8 and GA9

In this study, we use the latest scientific configurations of the atmospheric models from the Met Office: global atmosphere model version 8 (GA8) (Willett et al., 2020, Zhu et al. 2024), and version 9 (GA9). GA9 has improvements over GA8 in many areas of global model science compared to GA8, including several enhancements to the 6A convection scheme, latent heating in the gravity wave drag scheme, the "fountain buster" scheme, bimodal cloud initiation, and dust-dependent ice-nucleation temperature (Xavier et al., 2024). GA8 and GA9 are both tightly coupled to the JULES global land surface model (configuration GL9.0). The horizontal grid spacing of both models is 12km (N1024), and the vertical resolution consists of 70 levels. GA8 is the model configuration in the Bureau's APS4 global forecast system, ACCESS-G4/GE4, which will be implemented in Bureau operations towards the end of 2024 (Zidikheri et al., 2024 describe the global ensemble version of APS4).

2.2. Reanalysis and observations

To evaluate the model forecasts, we use ERA5 (Hersbach et al. 2020), the Bureau of Meteorology’s Atmospheric high-resolution Regional Reanalysis for Australia (BARRA, Su et al. 2019) and Global Precipitation Measurement data (GPM, Hou et. al 2014).

2.2.1. ERA5 reanalysis

The ERA5 (Hersbach et al. 2020) reanalysis is the fifth generation of atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and is available through the Copernicus Climate Change Service (C3S). It provides detailed hourly estimates of a wide range of atmosphere and land variables from January 1950 to the present.

ERA5 data is available on a ~31 km grid (~0.25º) and includes 137 vertical levels from the surface up to 80 km in the atmosphere. This reanalysis combines vast amounts of historical observations with advanced modelling and data assimilation systems to create a comprehensive and consistent record of the Earth's climate.

2.2.2. BARRA reanalysis

BARRA is the Bureau of Meteorology's regional reanalysis, covering Australia, New Zealand and Southeast Asia (Su et al. 2019; Su et al. 2022). Here we use the latest version, BARRA2, which provides a reanalysis at 12km horizontal resolution spanning 1979 to present (Su et al. 2022). BARRA2 is nested in the ERA5 global reanalysis and uses the UM atmospheric model. The model science configuration is based on model physics from global UM version GA7.2/GL8.1, along with additional packages from GA8 that enhance precipitation modelling (Su et al. 2022). The resolution of BARRA is 12 km.

2.2.3. GPM rainfall satellite data

The GPM (Hou et. al 2014) mission provides detailed data on global rainfall and snowfall. The resolution of the dataset is 0.1º. Using advanced satellite technology, GPM captures precipitation measurements across the globe, offering insights into weather patterns, storm intensities, and climate change. This data is crucial for improving weather forecasts, managing water resources, and understanding the Earth's water cycle.

2.3. Contiguous Rain Area (CRA) verification

In our study, we produced forecasts of TC-Jasper using the Unified Model (UM) and machine learning models, initialised on December 12th. By this time, TC-Jasper had weakened to a Category 1 storm and was moving westward. It made landfall on December 13th, bringing heavy rainfall to the coast and inland areas of northern Queensland. We used the Contiguous Rain Area (CRA) method to evaluate the rainfall forecasts on the Queensland coast produced by the models.

For the evaluation, all models and the BARRA2 and GPM data are first regridded (bilinear interpolation) to the ERA5 grid (0.25º) for a fair comparison.

The CRA verification method was developed by Ebert and McBride (2000) to evaluate systematic errors in rainfall forecasts. It was one of the earliest object-based methods to objectively compare the properties of matched forecast and observed rainfall systems.

The CRA method evaluates spatial forecasts by comparing forecasted and observed entities, which are defined by closed contours of a field (like rain areas) and uses pattern

matching techniques to determine errors in the forecast. Traditional verification methods will heavily penalise the forecast if the feature is mislocated. However, the object-oriented approach of CRA can take location errors into account and determine the accuracy if there were no position error. The CRA method facilitates a decomposition of the total error in the forecast into components due to location (displacement and rotation), rainfall amount (volume), and pattern.

CRA evaluation has been applied to various weather systems to identify systematic errors in rainfall predictions (e.g., Dube et al., 2014; Ashrit et al., 2015; Sharma et al., 2015, 2017, Chen et al. 2018). We follow the approach of Chen et al. (2018), who specifically applied the method to evaluating tropical cyclone rainfall. The first step in the method is to identify the "object" in the forecasts and observations. In this study, we define the object as heavy rainfall within a rectangular area (16°lon x 14°lat) associated with the tropical cyclone (TC). The continuous rainfall area is then defined by choosing a region enclosed by a specified rainfall amount in the forecasts and observations respectively. We use different rainfall thresholds to define the CRA for evaluating different intensities of rainfall. The next step is to find the best match between the object in the forecast and observation. To do this, the forecast is horizontally moved over the observations and then rotated until a best-fit criterion is met. In this study we use the maximum correlation coefficient best fit criterion. The spatial correlation coefficient measures how well the estimated values correspond to the observed values. It is a good measure of linear association or phase error. The correlation (r) is given by:

$$r = \frac{n(\sum XF) - (\sum X)(\sum F)}{\sqrt{[n\sum X^2 - (\sum F)^2]\ [n\sum F^2 - (\sum X)^2]}}$$

in which $\boldsymbol{F}$ is forecasted value and $\boldsymbol{X}$ is observed value across $\boldsymbol{n}$ grid boxes.

The total mean squared error (MSE) between the forecast and observed can be represented as:

$$\mathbf{MSE}_{\mathbf{total}} = \mathbf{MSE}_{\mathbf{displacement}} + \mathbf{MSE}_{\mathbf{rotation}} + \mathbf{MSE}_{\mathbf{volume}} + \mathbf{MSE}_{\mathbf{pattern}}$$

For the maximum correlation coefficient best-fit criterion, each MSE error component can be calculated as follows:

$$\mathbf{MSE}_{\mathbf{displacement}} = 2S_F S_X(r_s - r)$$

$$\mathbf{MSE}_{\mathbf{rotation}} = 2S_F S_X(r_{\mathbf{opt}} - r_s)$$

$$\mathbf{MSE}_{\mathbf{volume}} = (F'_m - X'_m)^2$$

$$\mathbf{MSE}_{\mathbf{pattern}} = 2S_F S_X(1 - r_{\mathbf{opt}}) + (S_F - S_X)^2$$

where $\boldsymbol{r}$ represents the original spatial correlation between forecast and observed objects, $\boldsymbol{r_{\mathbf{opt}}}$ denotes the maximum spatial correlation between forecast and observed objects after moving and rotating the forecast, and $\boldsymbol{r_s}$ indicates the maximum spatial correlation between forecast and observed objects after moving the forecast only (Chen et. al. 2018). $\boldsymbol{F'_m}$ and $\boldsymbol{X'_m}$ are the mean value of the forecast and observation respectively,

after moving and rotating. S_F and S_X are the standard deviations of the forecast and observed values respectively (Chen et. al. 2018).

Further details of the CRA method for evaluating tropical cyclone rainfall forecasts and systematic errors can be found in Chen et al. (2018).

3. Results

3.1. CRA evaluation of a GraphCast forecast

Before comparing models, and to fully demonstrate the CRA methodology, a verification of a GraphCast forecast is investigated here. Figure 2 demonstrates the CRA methodology by comparing GPM observations and GraphCast estimates of 24-hour rainfall distributions during TC Jasper, from 00:00 UTC on December 12, 2024, to 00:00 UTC on December 13, 2024. The forecast is verified using a 0.25° grid resolution and a 10mm CRA threshold. Below is a detailed description of how to interpret the CRA verification results for this case.

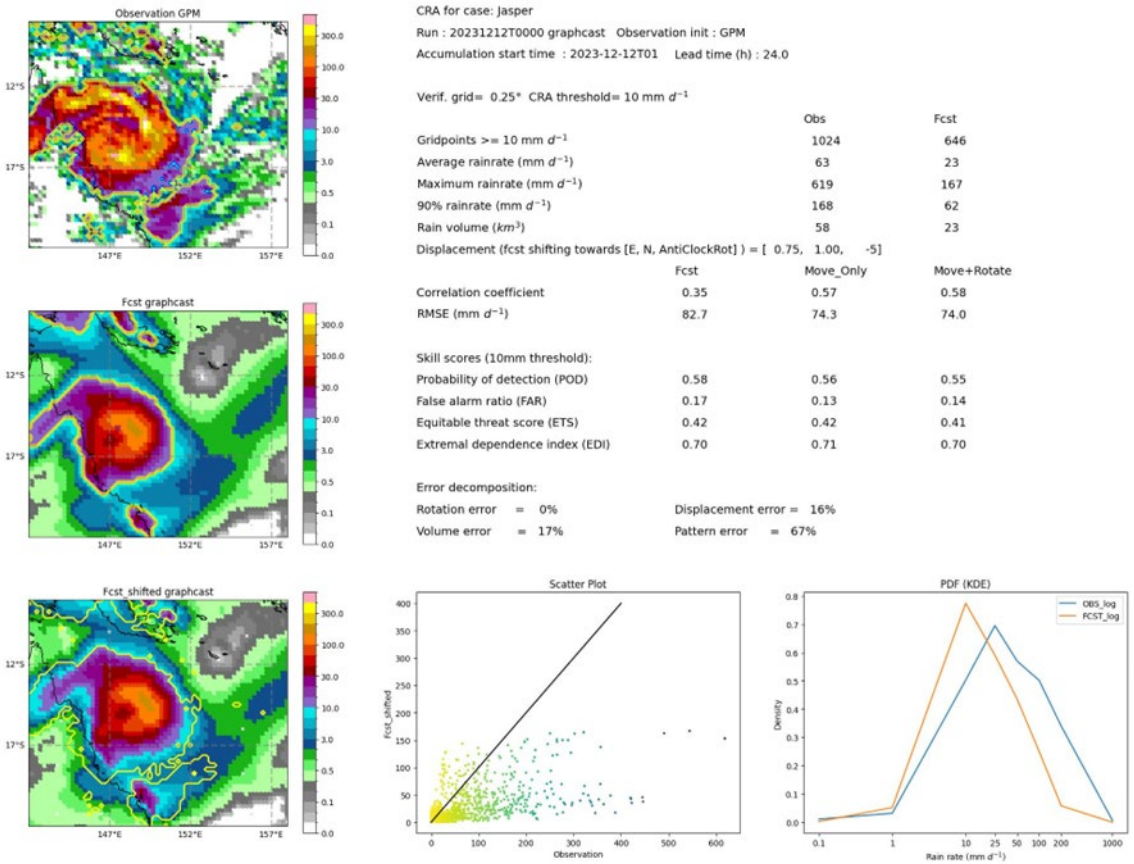

Figure 2: CRA verification of 24h forecast rain in TC-Jasper, accumulated from 2023:12:12:00 UTC to 2023:12:13:00 UTC relative to the GPM observation. Top-left panel: Reference data (GPM estimates in this study). Middle-left panel: Original forecast data. Bottom-left panel: Shifted forecast after moving and rotating. Bottom-middle panel: Rain rate scatter plot of shifted forecast vs. reference data. Bottom-right panel: Rain rate distribution (Kernel Density Estimation (KDE) PDFs). The yellow contour in the top-left and middle-left

panels encloses the rain exceeding the CRA threshold in each dataset. The union of these areas forms the CRA area, shown by the yellow contour in the bottom-left panel. Various statistics are displayed in the right panel.

The CRA verification package outputs several diagnostics and statistics, as shown in Figure 2. The yellow contour in the top-left and middle-left panels encloses the rain exceeding the CRA threshold in the GPM and GraphCast data respectively. The union of these areas forms the CRA area, shown by the yellow contour in the bottom-left panel. Various statistics are displayed in the right panel.

In this case, the GraphCast forecast shows a smaller area of rainfall exceeding the 10mm CRA threshold compared to GPM. According to three rain rate distribution metrics (average, maximum, and 90th percentile), GraphCast has a much lower rain rate than GPM. The underestimation of extreme rainfall is more pronounced, with the 90th percentile rain rate at 62mm in GraphCast compared to 168mm in GPM, and the maximum rain rate at 167mm in GraphCast compared to 619mm in GPM. The underestimation of heavy rain by GraphCast is also evident in the scatter and PDF plots. This underestimation in the GraphCast model is partly due to the coarser resolution of the training data (ERA5 ,0.25°) which is lower than that of GPM (0.1°), even though the GPM data were regridded to the ERA5 grid prior to this analysis.

For displacement error, the GraphCast rainfall forecast needs to be shifted 0.75° eastward, 1° northward, and rotated 5° anticlockwise to best match the GPM data (Fig.2). The correlation coefficient and root-mean-squared error (RMSE) between GraphCast and GPM are 0.35 and 82.7 mm/day respectively, before any shifting and rotation. These scores improve to 0.58 and 74 mm/day if the forecast rain area is both moved and rotated. The error decomposition attributes most of the error to pattern error (67%), followed by volume error (17%) and displacement error (16%), with rotation error being negligible.

3.2. Scatter plot comparisons of the models

In this study we verify the model forecasts against the ERA5 and BARRA2 reanalyses, as well GPM observations. Apart from the considerable observational uncertainty that exists for rainfall, different products have different advantages and disadvantages. For example, BARRA2 is produced at a higher resolution (12 km) than ERA5 (31 km). In addition, GraphCast is trained on ERA5 data, meaning that it will most closely reproduce the characteristics of ERA5, even if those characteristics are not accurate compared to observations.

Figure 3 presents scatter plots of the 24h rainfall estimates of the model forecasts with respect to the different reanalyses/observations for the CRA area. The GA8 and GA9 UM forecasts show a good linear relationship with BARRA2 rainfall estimates but tend to overestimate rainfall amounts compared to ERA5, especially for daily rainfall amounts exceeding 100 mm/day. The GraphCast forecast has the best linear relationship with ERA5 but underestimates rain (particularly for amounts greater than 100 mm/day)

relative to BARRA2. All models tend to underestimate rainfall amounts compared to GPM data, but this is more noticeable for GraphCast for totals exceeding 100 mm/day.

GraphCast is trained using ERA5 data and shares the same model resolution of ~31 km. Consequently, we might expect GraphCast to perform better when compared to ERA5 than when compared to BARRA2 or GPM – which indeed it does. BARRA2 (12 km native resolution) and GPM (0.1º resolution) include smaller-scale and higher amplitude weather/rainfall features than ERA5 due to their higher resolutions, despite being regridded to the coarser ERA5 grid prior to evaluation. However, it is instructive to see that GraphCast performs less well in capturing the intense rainfall when compared to BARRA2 and GPM than when compared to ERA5.

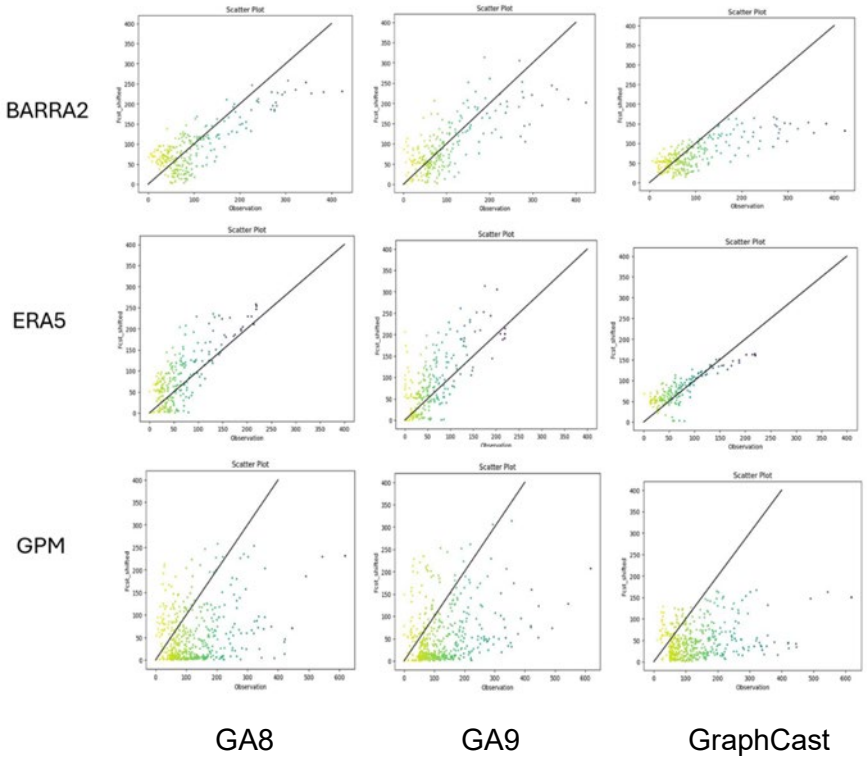

Figure 3: Scatter plots of the 24h rainfall forecast accumulations for GA8 (left column), GA9 (middle column) and GraphCast (right column) with respect to the BARRA2 (top row) and ERA5 (middle row) reanalyses and GPM observations (bottom row). Each point represents the rainfall in a grid cell that falls within the CRA area.

In contrast, the GA8 and GA9 forecasts have a better correlation with the BARRA2 reanalysis but overestimate rainfall compared to ERA5. The model used in BARRA2 has commonalities with these forecasts – it is based on the UM GA7 model configuration, with some improved model physics from the GA8 package (Su et. al 2022) and it operates on the same (12 km) resolution.

Perhaps the most independent reference data is GPM. Although all three models underestimate high rainfall amounts with respect to GPM, the underestimation is much more systematic with GraphCast. GraphCast also produces much lower rainfall totals in

general compared to the GA8 and GA9 models. The 24h maximum rain is 167 mm in GraphCast versus 258mm and 313 mm in GA8 and GA9 respectively (noting that all are evaluated on the 0.25º ERA5 grid).

3.3. Skill score comparisons

To summarise the respective model forecast skill relative to the different reference datasets, Figure 4 presents two widely used verification metrics: the pattern correlation coefficient and mean square error (MSE), plotted as functions of the rainfall threshold. The results pertain to the 72-hour accumulated rainfall amount for the forecast initialised at 00:00 UTC on December 12, 2023, and are done on the shifted objects (forecasts are moved and rotated as described in Section 2). Figure 4 highlights a common trend: as the rainfall threshold increases, the accuracy of the forecast decreases, underscoring the difficulties in predicting high-impact weather events.

As shown in the section above, the forecast skill is dependent on the reference data. When compared to ERA5, GraphCast demonstrates superior forecast skill, exhibiting the highest correlation coefficient and lowest MSE across all rain thresholds, outperforming both GA8 and GA9.

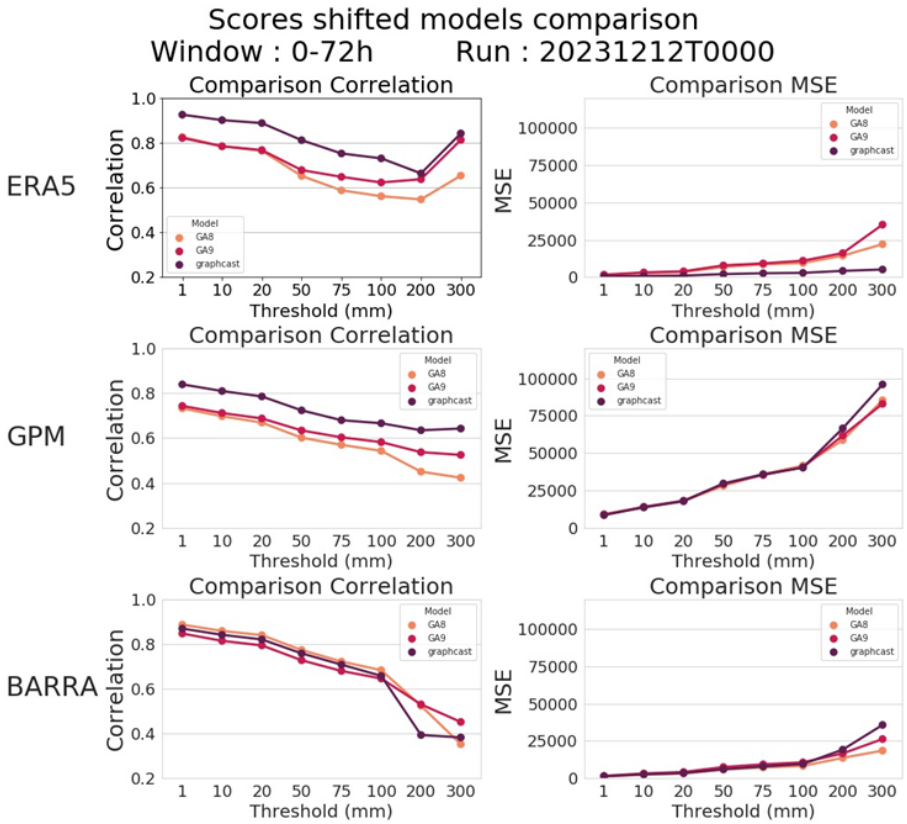

Figure 4: Comparisons of pattern correlation coefficients (left) and mean square errors (MSE) (right) for with different rainfall thresholds (x-axis) from GA8 (orange), GA9 (red) and GraphCast (purple) rainfall forecasts relative to the three reference datasets: ERA5, GPM, and BARRA2. The evaluation is for 72h accumulated rainfall since the initialisation of the forecast (20231212T00).

Compared to GPM, GraphCast achieves the highest correlation coefficients among the three model forecasts but has the highest MSE for rainfall amounts exceeding 100mm. GA9 shows slightly better skill than GA8 in terms of correlation.

When compared to BARRA2, GA8 and GA9 exhibit better skill for both correlation and MSE, while GraphCast shows the lowest skill for rainfall amounts greater than 100mm. This is likely related to the coarser native resolution of GraphCast and that GA8/GA9 have similar model configurations and resolutions to BARRA2. In this study, GA8 shows slightly better skill than GA9 when compared to BARRA2, probably because BARRA2 is based on a model physics package more similar to GA8 than GA9. The forecast skills are comparable for rainfall amounts less than 100mm across the three models when compared to BARRA2.

Interestingly, the MSE values from all of the models are much larger when compared to the more independent GPM data than when compared to ERA5 and BARRA.

3.4. Decomposition of error

As mentioned in Section 2, the CRA method allows decomposition of the total error into four components: rotation, displacement, volume, and pattern. Figure 5 shows the decomposition for GraphCast, GA8 and GA9 against the different reference datasets and using a CRA threshold of 50mm. In all three models, the pattern error dominates, especially for comparisons against ERA5 and BARRA (exceeding 90% for GraphCast and GA8 when compared to BARRA2). Rotation errors are negligible, and displacement errors are small, except for the ERA5 comparisons which show error contributions of ~20-30% across the models. The contribution of the volume error is only when forecasts are compared to GPM, with a value of 40% for GraphCast and slightly lower for GA8 and GA9.

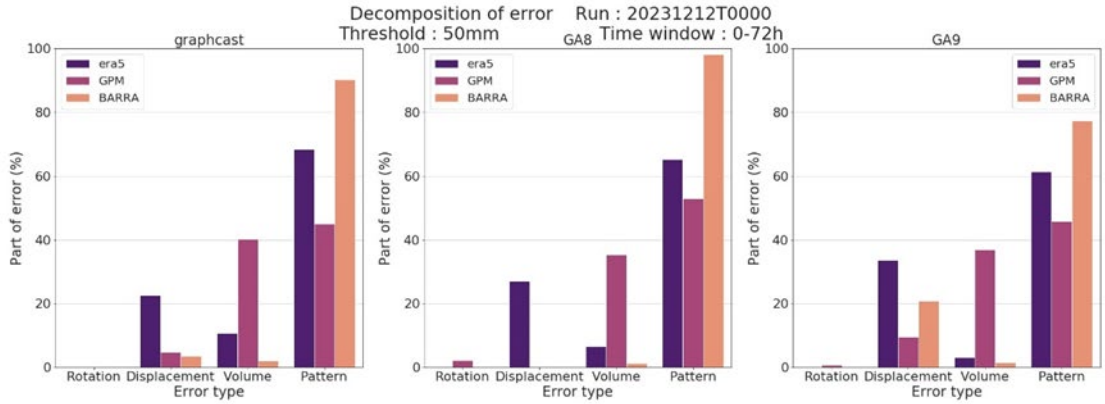

Figure 5: Error decomposition for 72-hour accumulated rainfall with a CRA threshold of 50mm. The error components are rotation, displacement, volume, and pattern errors. The three references used are ERA5, BARRA2 and GPM.

4. Summary

We have evaluated rainfall forecasts of TC-Jasper from an ML model (GraphCast), and dynamical models (UM GA8 and GA9). The Contiguous Rain Area (CRA) method was used to evaluate the performance of the model forecasts against three reference datasets: ERA5, BARRA2 and GPM. Key metrics for validation included the correlation coefficient and mean square error (MSE) as functions of the rainfall threshold.

Forecast performance showed significant sensitivity to the different reference datasets. GraphCast, trained using ERA5 data with a resolution of 0.25º, demonstrated the best performance when compared to ERA5, exhibiting the highest correlation coefficient and lowest MSE across all rain thresholds. Partly due to its coarse resolution, GraphCast underestimated the most intense rainfall, particularly when compared to BARRA2 and GPM. However, underestimation of an intense weather system by an ML model compared to physical NWP models is not merely a result of being trained on a "coarser" resolution dataset. Liu et al. (2024) states that ML models often produce smoother results due to several factors: regularization techniques (Goodfellow et al. 2016), data averaging (Breiman 1996), optimization objectives (Bishop 2006), and noise reduction (Ng 2004). Researchers are making progress in improving machine learning (ML) models to better predict intense weather systems. One approach is to use a diffusion model architecture (Sohl-Dickstein et al., 2015; Ho et al, 2020), as is done by Zhong et al. (2024) and Price et al. (2024) for example. Another method involves integrating ML models with traditional physics-based models (e.g., Niu et al. 2024), or using a hybrid ML and physics-based modelling system (e.g., Kochkov et al, 2024).

UM GA8 and GA9 forecasts outperformed GraphCast for intense rainfall and demonstrated better forecast skill when compared to BARRA2. A likely contributor to this outcome is the fact that BARRA2 is a reanalysis based on the UM and has a resolution of 12km, the same as GA8 and GA9 forecasts. For weak and moderate rain, GraphCast's forecast skill is comparable to that of GA8 and GA9.

When compared to GPM, GraphCast showed the best skill in terms of the correlation coefficient among the three models but had the highest MSE for intense rain.

In terms of error decomposition, pattern errors dominated across all models, while rotation errors were negligible. Volume errors were significant when compared to GPM, and displacement errors were significant relative to ERA5 for all three models.

Case studies exploring the performance of ML models in greater detail, and against a variety of benchmarks and other models, are still relatively scarce in the literature. More such work is desirable to better inform the community of the capabilities of these models.

It is worth noting that most ML models are trained on ERA5 and are also often evaluated against ERA5 (e.g. Liu et al., 2024). Our case study highlights that comparing (only) against ERA5 may lead to a misrepresentation of the performance of these models for phenomena that may not be well captured in ERA5 and/or for variables like rainfall for which there is considerable observational uncertainty (Cheung et al. 2024).

Overall, the study highlights the strengths and weaknesses of each model forecast in relation to different reanalyses/observations, providing insights into their performance and areas for improvement. Although this study is highlighting errors, it is notable that the rainfall forecast for TC Jasper from GraphCast is indeed comparable to the two dynamical models. This demonstrates the considerable promise for ML models into the future.

Acknowledgements

This study was undertaken while the lead author (Hector Morisseau) was on an internship at the Bureau from Météo-France. The Bureau and Météo-France have a longstanding partnership, marked by the regular integration of mid-career students from Météo-France's graduate school into the Bureau's scientific research environment for internships. We extend our gratitude to Jun Smith and Beth Ebert for their assistance with the CRA rainfall evaluation and the technical support from Griff Young, Harrison Cook and Anderson Murray.

References

Ashrit, R., E. Ebert, A. Mitra, K. Sharma, G. Iyengar, and E. Rajagopal (2015), Verification of Met Office Unified Model (UM) quantitative precipitation forecasts during the Indian monsoon using the Contiguous Rain Areas (CRA) method, Tech. rep., National Centre for Medium Range Weather Forecasting Ministry of Earth Sciences, Government of India.

Bi, K., Xie, L., Zhang, H. et al. (2023), Accurate medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bishop, C. M. (2006), Pattern Recognition and Machine Learning. Springer New York

Bonavita, M. (2023), On some limitations of data-driven weather forecasting models. https://arxiv.org/abs/2309.08473

Bonev, B., T. Kurth, C. Hundt, J.Pathak, M. Baust, K. Kashinath, A. Anandkumar. (2023), Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere, Proceedings of the 40 th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, https://arxiv.org/abs/2306.03838

Breiman, L, (1996), Bagging predictors. Mach. Learn. 24, 123–140

Chen, Y., Ebert, E. E., Davidson, N. E., & Walsh, K. J. E. (2018), Application of contiguous rain area (CRA) methods to tropical cyclone rainfall forecast verification. *Earth and Space Science*, *5*, 736–752. https://doi.org/10.1029/2018EA000412

Charlton-Perez, A. J. and co-authors (2024): Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. npj Clim. Atmos. Sci. 7, 93

Cheung, K. K. W., Ji, F., Nishant, N., Teng, J., Bennett, J., and Liu, D. L.: Comparison of BARRA and ERA5 in Replicating Mean and Extreme Precipitation over Australia, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2024-286, in review, 2024

Colony, C., & Andigani, R. (2024), Solarcast-ML: Per Node GraphCast Extension for Solar Energy Production. https://arxiv.org/abs/2406.13559

DeMaria, M., and co-authors (2024): Evaluation of Tropical Cyclone Track and Intensity Forecasts from Artificial Intelligence Weather Prediction (AIWP) Models, https://arxiv.org/abs/2409.06735

Dube, A., R. Ashrit, A. Ashish, K. Sharma, G. Iyengar, E. Rajagopal, and S. Basu (2014), Forecasting the heavy rainfall during Himalayan flooding—June 2013, Weather and Climate Extremes, 4, 22 – 34, https://doi.org/10.1016/j.wace.2014.03.004

Ebert, E., & McBride, J. (2000), Verification of precipitation in weather systems: Determination of systematic errors. Journal of Hydrology, 239(1–4), 179–202. https://doi.org/10.1016/S0022-1694(00)00343-7

Goodfellow, I., Bengio, Y. & Courville, A. (2016): Deep Learning. MIT Press, Chapter 7

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

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., & Iguchi, T. (2014), The Global Precipitation Measurement Mission. Bulletin of the American Meteorological Society, 95(5), 701-722. https://doi.org/10.1175/BAMS-D-13-00164.1

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., & Hoyer, S. (2024). Neural general circulation models for weather and climate. Nature, 632(8027), 1060-1066.

Liu, C. C., Hsu, K., Peng, M. S., Chen, D. S., Chang, P. L., Hsiao, L. F., & Kuo, H. C. (2024), Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific. *Climate and Atmospheric Science*, *7*(1), 221 https://doi.org/10.1038/s41612-024-00769-0

Ng, A. Y. (2004), Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML)

Niu, Z., and co-authors, (2024): Improving Typhoon Predictions by Integrating Data-Driven Machine Learning Models with Physics Models Based on the Spectral Nudging and Data Assimilation, https://arxiv.org/abs/2408.12630v1

Olivetti, L., & Messori, G. (2024), Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast. *EGUsphere*, *2024*, 1-35

Prasad V (2024): Severe Tropical Cyclone Jasper (02U), http://www.bom.gov.au/cyclone/history/jasper23.shtml

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., & Willson, M. Gencast: Diffusion-based ensemble forecasting for medium-range weather (2024). URL: https://arxiv.org/abs/2312.15796, 2312

Remi Lam et al. (2023), Learning skilful medium-range global weather forecasting. Science, 382, 1416-1421. https://doi.org/10.1126/science.adi2336

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., & Sha, F. (2024), Weather Bench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modelling Earth Systems*, *16*(6), e2023MS004019.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

Su, C.-H., Eizenberg, N., Steinle, P., Jakob, D., Fox-Hughes, P., White, C. J., Rennie, S., Franklin, C., Dharssi, I., and Zhu, H. (2019): BARRA v1.0: the Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia, Geosci. Model Dev., 12, 2049–2068, https://doi.org/10.5194/gmd-12-2049-2019

Su, C.-H. and coauthors (2022): BARRA2: Development of the next-generation Australian regional atmospheric reanalysis. Bureau Research Report No. 067

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., and Monfardini, G. (2009): The graph neural network model, https://ro.uow.edu.au/infopapers/3165

Sharma, K., R. Ashrit, E. Ebert, G. Iyengar, and A. Mitra (2015), NGFS rainfall forecast verification over India using the Contiguous Rain Area (CRA) method, Mausam, 66, 415–422

Sharma, K., R. Ashrit, G. R. Iyengar, A. Mitra, B. Ebert, and E. N. Rajagopal (2017), Spatial verification of rainfall forecasts during Tropical Cyclone 'Phailin', Springer, Cham. Short, C

Subich, C. (2024). Efficient fine-tuning of 37-level GraphCast with the Canadian global deterministic analysis. https://www.arxiv.org/abs/2408.14587

Willett M. R., T. Graham, M. Brooks and D. Copsey, 2020: GC4 and GA8GL9 Acceptance Report, Technical report, UK Met Office

Xavier, P., and co-authors (2024): Assessment of the Met Office Global Coupled model version 5 (GC5) configurations. Met Office report

Xie, Y.N., and co-authors, 2024: Evaluation of typhoon forecasts using the Pangu-Weather Neural Network Model: A comparative analysis with ECMWF and other numerical prediction models. 23rd Conference on Artificial Intelligence for Environmental Science, Baltimore, MD, Jan-Feb, 2024, Baltimore, MD, https://ams.confex.com/ams/104ANNUAL/meetingapp.cgi/Paper/430780

Yan, Z., Lu, X., Wu, L., Liu, F., Qiu, R., Cui, Y., & Ma, X. (2024). Evaluation of precipitation forecasting base on GraphCast over mainland China. https://doi.org/10.21203/rs.3.rs-4645037/v1

Zhong, X., Chen, L., Liu, J., Lin, C., Qi, Y., & Li, H. (2024). FuXi-Extreme: Improving extreme rainfall and wind forecasts with diffusion model. Science China Earth Sciences, 1-13

Zhu H et al. (2024): Impacts of the new UM convection scheme, CoMorph-A, over the Indo-Pacific and Australian regions. Journal of Southern Hemisphere Earth Systems Science 74, ES23011. https://doi.org/10.1071/ES23011

Zidikheri, M. J., Steinle, P.J., Xiao, Y. Villardon, E.A., 2024: An objective evaluation of the Bureau's ACCESS GE4 global ensemble model. Bureau Research Report Number 091