Numerical vs AI Models in Global Hurricane Forecasting

ALAN NESSIPBAYEV, University of Twente, The Netherlands

Natural disasters, such as hurricanes and floods are often recognized as catastrophic events that have a major impact on socio-economic and environmental sectors. They are considered to be difficult to control with the current level of technology that humans have. Thus, it is necessary to have a good prediction mechanism that can provide an early alert to save as many lives as possible and reduce damage. Artificial intelligence (AI) is considered to be one of the most promising solutions to potential problems that humans may face as a civilization. It is not uncommon for agencies such as the National Centers for Environmental Prediction (U.S.) or the European Centre for Medium-Range Weather Forecasts (ECMWF) to develop and deploy different types of AI technologies that include machine learning and neural networking to analyze a large amount of data received via satellites or on-ground sensors. With the help of AI, the agencies are looking for potential improvements in the accuracy of prediction compared to times when most things were calculated by complex machines that required significant time and resources.

This research specifically focuses on performance comparison of the modern numerical models that are used for accurate weather prediction and their AI-enhanced counterparts. The findings highlight the potential of AI-enhanced models to improve the prediction of hurricanes, which ultimately leads to better preparation and more efficient efforts of local communities. Moreover, the work provides insights into model performances and demonstrates that ,even though the AI-system provide acceptable and positive results, they do not show crucial difference in prediction accuracy to state that they are significantly better than the deployed numerical systems. Lastly, the summary of the findings gives a following result: The IFS took the 1st place, with the GFS Graphcast in 2nd, closely followed by AIFS in 3rd, and the GFS as the last one on the list.

Additional Key Words and Phrases: Hurricane, Artificial Intelligence, forecasting, analysis, prediction, comparison, Numerical Weather Prediction.

1 INTRODUCTION

The first successful implementation of computational techniques in weather forecasting was carried out by a team of scientists led by Jule Charney in the 1950s, with the help of the ENIAC computer. It symbolized the beginning of a new era of forecasting systems that are still used to this day. With the help of computers, people's time to prepare for natural threats has significantly increased due to early warning systems that can track and predict natural disasters. Even with this, several studies suggest that the old models used for weather forecasting are too time-consuming and costly [7]. Thus, a new generation of AI-assisted models was presented as a solution for accurate predictions based on historical data. However, historical data becomes less relative due to the significant climate changes that the earth is currently facing, and it is something that needs to be addressed. Several modern weather prediction models are deployed to issue early warnings if necessary. However, it still needs to be determined what methods provide the most accurate results to ensure preparedness for potential threats. Modern approaches include Numerical Weather Prediction (NWP), deep learning, ensemble forecasting, statistical methods, and data assimilation. Almost all of these methods can potentially involve work with Artificial Intelligence, whether to make a whole prediction or to exclude biases from the output of a mathematical model. Nonetheless, it is still uncertain how the location of a forecast can affect the accuracy of the models. Thus, more research must be done. In this research, current mathematical approaches that are used to identify and predict hurricanes will be analyzed to understand their accuracy. It will be followed by an analysis of the new promising AI-supported systems that have the potential to replace current methods. Such analyses will help to further enhance existing models and deal with the current issues that these systems face. The research will focus on comparing mathematical and hybrid (AI-assisted) models. To the best of the author's knowledge, as of March 2024, there has been no research done that covers a comparison of AI-enhanced and numerical models in forecasting hurricanes in different parts of the world.

2 PROBLEM STATEMENT

Despite significant improvements in weather prediction models, accurate hurricane prediction remains a challenging task. Numerical models that are based on mathematical and physical formulas require significant computational powers and still show certain limitations when it comes to predicting complex weather phenomena such as hurricanes. Thus, recently developed AI-enhanced models that utilize machine learning and neural networking techniques are presented as promising solutions that will increase accuracy and decrease the costs of the forecasting processes.

The goal is to determine whether AI-enhanced models provide a significant advantage over the classical models to replace them in hurricane forecasting. It involves assessing the accuracy and reliability of such models in predicting data variables that contribute to hurricane detection in various hurricane-prone regions of the world. Furthermore, an assessment of the performance of the models in recent real-life cases of hurricanes is crucial for the evaluation of the capabilities of the models.

3 RESEARCH QUESTION

The problem statement gives rise to the following research question:

"How do AI-enhanced weather prediction models perform in comparison to numerical models in terms of accuracy and reliability for hurricane forecasting in various hurricane-prone parts of the planet?"

To answer this, the following sub-questions are stated:

TScIT 41, July 5, 2024, Enschede, The Netherlands

(1) What are the most effective and promising models that are deployed to predict specific natural disasters such as hurricanes?
(2) What data is being used during the processes of prediction and how is it related to hurricane activities?
(3) What are the most hurricane-prone regions of the world?
(4) How do the models perform across different regions and in real-life scenarios when compared to each other?

4 RELATED WORK

To conduct the research it was decided to use several popular research domains such as IEEE, Google Scholar, Scopus, and the University of Twente (UT) online library. Moreover, for specific technical data, the ECMWF website, which includes recent publications, and the Copernicus Climate Data Store were used for documentation access.

In the field of weather prediction based on mathematical and/or AI-enhanced models, there have been a significant number of studies that cover both approaches. The studies can be split up into two groups where the first group focuses on upgrading existing methods [21, 27] that involve mathematical computations, and the other discusses a more promising topic that involves AI-based solutions [5, 10, 24, 26]. However, some sources discuss attempts to improve existing models by applying deep-learning processes for data analysis [5]. The studies confirm that there is an existing problem where the current solutions that involve mathematical models either have a relative inaccuracy in the results or are not cost/time efficient. Moreover, some challenges are present for AI-based solutions as well. Such challenges involve data inaccuracy, biases, and irrelevant historical data on which the modules are trained. All of these problems lead to studies [5, 26] that try to find a middle ground for both approaches and apply new technologies to already existing methods to investigate the outcomes. Nonetheless, it is important to mention that it is still unclear what methods make the most promising combinations to predict hurricanes and if there is a reason to make such combinations of systems in the first place. Therefore, more research has to be done to evaluate the potential of AI systems in weather forecasting.

5 METHODS OF RESEARCH

The research consists of three main parts that answer the main and sub-research questions.

5.1 Literature review

The first part of the research involves understanding how the weather-predicting models work, the difference between them, what models are considered the best, what regions are considered the most affected by hurricane activities, and what data is used to predict hurricanes when using such models. It also involves analyzing the literature to see the current views and ideas on AI-based or AI-assisted models as a replacement for NWP models.

5.2 Comparative Experiment

The second part involves collecting weather data from the most hurricane-prone areas of the world, acquiring weather forecasts made by the models that were selected for the experiment, comparing the forecasts, evaluating accuracy, and analyzing the results. Furthermore, it involves the application and comparison of the models on two real-life cases of the most recent hurricanes that occurred near Australia (Cyclone Megan) and Bangladesh (Cyclone Remal). The experiment is based on a comparison of the ERA5 dataset which is a reliable reanalysis dataset provided by the ECMWF and the outputs of the weather-predicting models.

The data variables that are used for the comparison include 2-meter temperature, surface pressure, sea-level pressure, wind speed, and cloud coverage. The experiment provides insights into the behavior of each model in every region and evaluates their accuracy and reliability if used around the globe or in real-life case scenarios.

5.3 Results Analysis

The third part includes the evaluation of the results and implementation of error and accuracy metrics such as Mean Absolute Error (MAE) and Normalized Mean Absolute Error (NMAE). The MAE metric gives an overview of how accurate the model is in terms of performance, and it allows a fair comparison of the models. The NMAE metric is a derivative of MAE and serves for model comparison across all data variables.

The analysis of the results in various parts of the world gave trustworthy insights into how accurate and reliable the chosen weather-predicting models are. The results and findings were turned into graphs and tables for better representation.

6 LITERATURE REVIEW

The purpose of this section is to provide valuable insights into previous studies that create a good basis for the experiment. This section covers experiment-related topics such as hurricane-prone regions of the world, previous NWP models' tests, AI-enhanced models' tests, and data variables for hurricane prediction.

6.1 Regions

Several trustworthy articles [11, 22, 25] supported by the National Oceanic and Atmospheric Administration (NOAA) and based on historical data provide a well-supported reason to suggest that the Gulf of Mexico, the Caribbean Sea, the East Coast of the United States, the Western Pacific, and the Eastern Pacific are some of the most hurricane-prone regions of the world. The studies further suggest that the disasters in this region cause severe damage and loss of life, especially in areas with poor infrastructure or low levels of emergency response sources. Thus, it is acceptable to use these regions as a fundamental key of the experiment.

6.2 Models

This subsection defines the two main parties of the experiment and discusses their current state. Those parties are NWP and AI-enhanced models.

6.2.1 NWP models. Numerous studies [19, 26] describe and test the efficiency of the NWP models such as ECMWF IFS and GFS. The results confirm the expectations and give a reason to assume that these numerical models are some of the best-performing models in the world. It is important to note that the local forecasting models

usually provide higher forecast resolutions, which causes the models to produce more accurate results. Nonetheless, the IFS and the GFS models are global systems and can be deployed anywhere. The studies verify the reliability of the models in predicting several data variables at different times of the day, at distinct places, and under numerous circumstances. Therefore, there is a strong basis to believe that the models chosen for this experiment have a well-supported background.

6.2.2 AI-enhanced models. AI-enhanced and AI-based weather predicting models are a recent phenomenon that attracts a lot of attention from weather agencies around the world. Several studies [6, 7, 9, 18] were conducted to evaluate the accuracy of these models and compare them to each other to demonstrate the potential that AI has in the future of weather forecasting. Studies [1, 18] on models such as AIFS, developed by ECMWF, already suggest that the model outperforms the traditional IFS model in several metrics and is expected to significantly increase its accuracy in the future. The GFS Graphcast consists of two parts that can potentially work separately but are united in this hybrid model. The first part is the standard GFS model and the second part is the AI-based Graphcast model developed by Google Deepmind. The GFS Graphcast takes the best of two worlds and is expected to increase its accuracy with future updates.

It is important to understand that these AI-enhanced models were developed quite recently, with the GFS Graphcast model being released by Google DeepMind in November 2023 and the AIFS model being released in early 2024. Therefore, not many studies provide a good overview or comparison of these models.

6.3 Data Variables

Hurricane prediction is a complex phenomenon and the forecasts for such natural disasters are based on a significant analysis of contributing data variables. For the experiment, it was decided to test five data variables that directly impact hurricane formation and development. These data variables are described and explained in detail in Section 7.2. The studies [4, 12, 14, 15, 23] confirm the importance of these data variables in model testing. Therefore, based on the recent studies, documentation, and other model tests, it is reasonable to conclude that the chosen data variables for this experiment are suitable for hurricane prediction model testing.

7 EXPERIMENT

This chapter goes into detail about the experiment which describes why certain locations were used as a primary example of hurricane-prone areas, what weather variables contribute to hurricane formation, what models are being deployed, and what the process of the experiment itself is.

7.1 Locations

The locations that were used in the experiment are united by several factors such as warm water temperature, which contributes to hurricane formation as it provides the energy, and humidity, which supplies hurricanes with moisture that further intensifies them. The list of locations in this experiment includes the USA, Japan, Honduras, Australia, and Puerto Rico.

7.2 Data

The choice of data is a crucial step in model performance analysis and comparison. It is necessary to understand what variables contribute to hurricane formation, development, and intensity the most. This chapter provides insights into the data variables that were used to compare the forecasting models as well as the reason why such variables were included.

Most of the forecast data was obtained from open-source APIs such as Copernicus Climate Data Store (CDS) and Open-Meteo. The ERA5 is a reliable reanalysis dataset provided by ECMWF in CDS. It was used as an observation dataset that contains the actual observed weather data for comparison with forecast datasets.

7.2.1 Two-Meter Temperature. Two-meter temperature is one of the key aspects of hurricane predictions that drive convection processes in the atmosphere. Higher surface temperatures indicate an increased rate of evaporation, which adds moisture to the atmosphere. Moisture is necessary for cloud formation. Moreover, the rise of warm air creates low pressure at the surface level, which contributes to the development of cyclonic systems.

7.2.2 Sea-Level Pressure. One of the main characteristics of a cyclone is low sea level pressure at their centers. Sea-level pressure data helps identify, predict, and follow the hurricanes. The lower the sea pressure is - the more intense a hurricane becomes. Thus, this variable is a significant comparative component between weather-predicting models.

7.2.3 Surface Pressure. Similar to sea-level pressure, surface pressure helps with the identification of hurricanes, their direction, and their intensity. However, the surface pressure provides more localized data for the overall hurricane overview.

7.2.4 Cloud Coverage. Cloud coverage is a direct representation of the moisture content in the atmosphere. Cloud coverage has a direct impact on the amount of precipitation and gives an overview of the intensity of the cyclones.

7.2.5 Wind Speed . Wind speed is a primary criterion for hurricane classification on the Saffir-Simpson scale (Category 1 to Category 5). High wind speeds serve as indicators of hurricane intensity and destructiveness. The analysis of wind speeds helps with assessing potential damage and preparation for any future natural disasters.

7.3 Models

This section describes what models were used in the experiment. It explains how the models work and their differences

7.3.1 ECMWF IFS 0.25. ECMWF's Integrated Forecasting System (IFS) is one of the most frequently used global NWP models and is known for its accuracy and reliability. The model includes two options of spatial resolution: ECMWF IFS 0.4° and ECMWF IFS 0.25°. This experiment focuses on the use of ECMWF IFS 0.25°, which provides an area of a single grid point of approximately 28 km by 28 km, which implies a higher resolution and accuracy compared to ECMWF IFS 0.4°, which has a grid area of approximately 44 km by 44 km at the equator. The model deploys advanced data

assimilation techniques to simulate and work with processes within the atmosphere.

7.3.2 GFS. The Global Forecast System (GFS) is an American global NWP model that was deployed by the National Centers for Environmental Prediction (NCEP). The model specializes in the simulation and prediction of atmospheric processes around the globe with a spatial resolution of approximately 0.25°. The GFS model utilizes sophisticated assimilation methods to integrate weather data and simulate atmospheric behavior. The model is widely accessible and is frequently used as a valuable tool for meteorologists worldwide.

7.3.3 ECMWF AIFS. It is one of the most recently developed AI-enhanced models which was introduced by ECMWF as a potential upgrade for IFS. The model operates at the same spatial resolution of 0.25° as IFS. It is capable of producing forecasts for up to 10 days and is believed to be one of the most reliable and precise models. The key difference between AIFS and IFS is the integration of artificial intelligence in data assimilation and the deployment of machine learning to improve the results of the forecasts. The use of artificial intelligence significantly reduces the costs and time that numerical models such as IFS must deal with to generate forecasts.

7.3.4 GFS Graphcast. GFS Graphcast is a state-of-the-art model developed to replace existing numerical models as it is believed to offer lower costs and increased accuracy. Graphcast deploys traditional NWP approaches of the GFS model combined with machine learning techniques from Google DeepMind's Graphcast AI-based model to generate reliable and highly accurate weather forecasts. The model is capable of operating worldwide and maintains a spatial resolution similar to GFS of 0.25°. The GFS Graphcast is believed to decrease costs and time per forecast, which is highly beneficial for meteorologists around the world.

7.4 Models Testing

The testing involved gathering predicted data from the given models and comparing them to the ERA5 dataset that contains reanalysis data. Each data variable received its own accuracy and error estimates. It was then used to compare the average accuracy of the models to determine reliability in different regions of the world.

7.5 Case Studies

To finalize the comparison between the models it was decided to test them on two real-life hurricanes that occurred in 2024: Cyclone Megan and Cyclone Remal.

Cyclone Megan made landfall in northern territories of Australia in March 2024 where it brought destructive winds and heavy rainfalls.

Hurricane Remal, which made landfall in Bangladesh and India, has caused 38 deaths, widespread destruction, and floods. The cyclone was hammering the country for more than 36 hours and caused considerable damage.

The case studies analysis followed a similar approach to model testing as described before. The predicted weather data was gathered from different models and compared to the ERA5 dataset that uses data reanalysis provided by ECMWF. The results of the case studies were compared and outlined in section 8.2.

8 RESULTS

The main focus point of the model comparison in this experiment is Mean Absolute Error (MAE) metrics. MAE is a reliable way to assess the models' performance since it represents the average of the absolute error between predicted and actual values. The MAE values are given in the same measurement units as the data variables. Thus, it is possible to assess the performance across the models in the prediction of the same data variable. Lower values indicate better performance of the model. Nonetheless, to compare the overall performance of a model across several data variables in different regions it was necessary to implement Normalized Mean Absolute Error (NMAE), which removes the influence of the units of measurement within the data.

The formulas for MAE and NMAE are provided below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|a_i - b_i| \tag{1}$$

where:

n is the number of observations

a_i is the actual value

b_i is the predicted value

$$\text{NMAE} = \frac{\text{MAE}}{\text{range of actual values}} \tag{2}$$

8.1 Model Comparison Results

Most of the data in this section is provided in table format where each table summarizes findings of a tested data variable predicted by different models in various regions. The tables are accompanied by short explanations of what each of them demonstrates. Some tables contain acronyms that represent the following countries/regions: the United States of America, Japan, Honduras, Australia, and Puerto Rico.

The acronyms of the data variables:

- 2-meter Temperature (2m Temp) in °C
- Surface Pressure (SP) in hPa
- Sea-level Pressure (SLP) in hPa
- 10-meter Wind Speed (10m WS) in km/h
- Cloud Cover (CC) in %

8.1.1 2-meter Temperature. The results of the error analysis for the "2-meter Temperature" variable are presented in Table 2. It is accompanied by Table 1, column 2, which provides data on the average MAE of the models.

The results show superior performance of the IFS model in temperature prediction with an average MAE of 0.49 °C. It is followed by two AI-enhanced models GFS Graphcast and AIFS. The GFS came out as the least accurate among the models. Thus, the results partially confirm the findings of the previous studies that implied that the AI-enhanced models begin to show better predictions when compared to some NWP models. Nonetheless, the performance of all models leaves a good impression on their accuracy and reliability in temperature prediction across the regions. Additionally, it is

important to notice the increase of MAE in the Honduras region testing for AIFS, GFS, and GFS Graphcast models.

Table 1. Average MAE of models in all regions

Model	2m Temp	SP	SLP	10m WS	CC
IFS	0.49	0.68	0.33	1.87	19.26
AIFS	1.21	1.40	0.73	3.49	23.19
GFS	1.29	1.266	1.28	3.36	39.68
GFS GRPH	1.11	0.67	0.72	3.13	23.37

Table 2. MAE 2-meter Temperature in °C

Model	USA	JPN	HS	AU	PR
IFS	0.56	0.43	0.50	0.48	0.50
AIFS	1.17	1.06	2.05	0.96	0.83
GFS	1.75	1.06	1.63	1.13	0.90
GFS GRPH	1.50	0.70	1.54	0.95	0.87

8.1.2 Surface Pressure. The results of the error analysis for the "Surface Pressure" variable are provided in Table 3 for MAE values and Table 1, column 3, for the average MAE of the models.

The results demonstrate a good performance of the IFS and GFS Graphcast models with their MAE not exceeding 1 hPa. Interestingly, Honduras demonstrates a higher average mean error than other regions. Overall, the average performance of the models across all regions in surface pressure prediction leaves satisfying results.

Table 3. MAE Surface Pressure in hPa

Model	USA	JPN	HS	AU	PR
IFS	0.28	0.33	1.55	0.27	0.99
AIFS	0.66	3.12	1.39	0.41	1.41
GFS	0.46	0.39	3.32	0.35	1.81
GFS GRPH	0.73	0.57	0.78	0.45	0.84

8.1.3 Sea-Level Pressure. The results of the error analysis for the "Sea Level Pressure" variable are outlined in Table 4, which summarizes the MAE of each model in the chosen regions, and Table 1, column 4, which provides the average MAE of each model.

The results point out that the IFS model on average demonstrates a better performance compared to other models and stays consistent across the chosen regions. It is followed by the GFS Graphcast and AIFS models that demonstrate similar average results for sea level pressure prediction. The GFS model takes the last place with the outcome of 1.28 hPa, which is still considered to be a good result. It is also important to mention the higher average MAE of sea level pressure in Honduras. More specifically, the tendency of the GFS model to give larger MAE in Honduras when predicting this and two previous variables.

Table 4. MAE Sea Level Pressure in hPa

Model	USA	JPN	HS	AU	PR
IFS	0.28	0.31	0.26	0.26	0.52
AIFS	0.66	0.59	0.99	0.39	1.02
GFS	0.46	0.39	3.34	0.33	1.87
GFS GRPH	0.73	0.57	1.01	0.44	0.89

8.1.4 10-meter Wind Speed. The results of the error analysis for the "10-meter Wind Speed" variable are showcased in Table 5 for the MAE of each model and the region and in Table 1, column 5, for the average MAE of each model across all regions.

The IFS model takes the leading position in terms of average accuracy of 10-meter wind speed prediction across the chosen regions. It has a significant gap from GFS Graphcast which is the second most accurate model in this case. The GFS and AIFS models take 3rd and 4th places respectively. Honduras comes out as the most problematic region for the GFS model with the MAE of 5.09 km/h which is a significant error.

Table 5. MAE 10-meter Wind Speed in km/h

Model	USA	JPN	HS	AU	PR
IFS	2.68	2.23	1.53	1.61	1.31
AIFS	4.05	5.07	3.64	2.17	2.54
GFS	4.80	2.53	5.09	2.29	2.09
GFS GRPH	4.81	2.25	3.77	2.48	2.32

8.1.5 Cloud Cover. The results of the error analysis for the "Cloud Coverage" variable are illustrated in Table 6 with the MAE of each model for every region and in Table 1, column 6, with the average MAE of the models.

In general, the IFS, AIFS, and GFS Graphcast show acceptable results if compared to the industry standards. The IFS model provided the most accurate results on average. Still, the results of the GFS Graphcast and AIFS are considered to be of good quality. Cloud coverage is an inherently difficult variable to predict due to the chaotic and complex nature of cloud formation and movement. Nonetheless, the average result of the GFS model of 39.68 indicates a poor performance within the experiment on this data variable. Interestingly, Honduras repetitively stands out as the most problematic for GFS to predict most of the data variables.

Table 6. MAE Cloud Coverage in %

Model	USA	JPN	HS	AU	PR
IFS	21.40	18.17	18.69	18.80	19.24
AIFS	21.23	29.86	20.51	22.14	22.22
GFS	27.18	35.91	54.22	53.92	27.19
GFS GRPH	23.95	31.71	20.27	19.82	22.01

8.1.6 Average NMAE. Table 7 provides the NMAE values that can be used for direct comparison of the models across the regions. It provides insights into the overall performance of the models in each region across all data variables. The summary of the findings is shown in Table 8, where the average NMAE of all models across the regions serves as a direct indicator of the performance within the experiment. The IFS model showed the best performance with the lowest NMAE and was placed as the best model on the list. It is followed by GFS Graphcast which provided good results in weather prediction across several variables and locations. The AIFS showed consistent results with most of the predictions. When compared in this part of the experiment, the GFS Graphcast model outperformed the AIFS model and emerged as a better AI-enhanced model on average. The GFS model demonstrated good results that often satisfy industry standards. Moreover, the experiment provided some interesting insights, where the performance of the GFS model tends to be less accurate in Honduras when compared to other parts of the world. Therefore, it appeared to be the least effective among the models.

Table 7. Average NMAE of models

Model	USA	JPN	HS	AU	PR
IFS	0.08	0.08	0.12	0.08	0.12
AIFS	0.12	0.20	0.17	0.12	0.19
GFS	0.137	0.14	0.36	0.18	0.23
GFS GRPH	0.142	0.12	0.15	0.11	0.16

Table 8. Average NMAE of models across all regions

Model	Avg NMAE	Rank
IFS	0.096	1
AIFS	0.16	3
GFS	0.21	4
GFS GRPH	0.14	2

8.2 Case Studies Results

This section provides an overview of the results of two case studies. The information of each cyclone and the models' results are described in respective subsections.

8.2.1 Cyclone Remal. The Cyclone with the given name Remal made its landfall on May 26, 2024. It caused severe damage to local communities with 38 people losing their lives and more than 150.000 houses destroyed or partially damaged.

The data shows that all the models in this study case demonstrated good performance, and all successfully predicted the cyclone. The surface pressure and sea-level pressure in this case are united into one data variable since observations were taken by the coast of Bangladesh. Thus, the collected data for surface pressure is the same for sea-level pressure. The graphs of data variables make it clear that the landfall was expected to occur approximately on the 26th of May. For instance, Figure 1 demonstrates pressure observations collected near Bangladesh when cyclone Remal occurred. A significant pressure drop can be noticed around the 26th of May which signifies a cyclone activity in the region. Similarly, Figure 2 shows the wind speed increase when the cyclone hit the region.

Despite larger MAE values in "Wind Speed" and "2m Temperature" predictions that can be seen in Table 9, the GFS's output still had a clear signal of the incoming cyclone. When compared, the IFS model performed the best in the prediction of all variables. The AI-enhanced models provided good results that satisfy the standards of the industry.

Table 9. Comparison of MAE of Data Variables across Models (Remal)

Data Variable	IFS	AIFS	GFS	GFS GRPH
2m Temp (°C)	0.65	1.07	1.52	1.23
SP (hPa)	0.63	0.98	0.68	1.09
SLP (hPa)	0.60	0.96	0.68	1.09
10m WS (km/h)	2.25	2.84	6.21	3.13
CC (%)	17.82	19.61	26.7	20.85

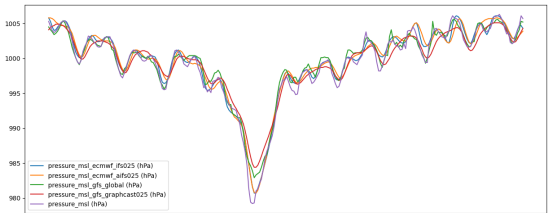

Fig. 1. Sea-level and Surface Pressures during Cyclone Remal

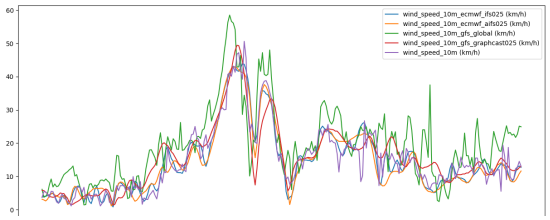

Fig. 2. 10-meter Wind Speed during Cyclone Remal

8.2.2 Cyclone Megan. The cyclone caused significant damage and floods in the northern part of Australia. The data taken to inspect the models' performance was taken from the 15th of March until the 23rd of March 2024. Similar to the first study case, the surface and sea-level pressures are united into one variable since the observations were collected on the coast of Australia. Therefore, the surface and the sea-level pressure have the same values.

The outcome of this case study provides some surprising results provided in Table 10. Initially, AI-enhanced models show superior results in 2-meter temperature forecasting. Moreover, similar performance occurs in predictions of wind speed and pressure variables. However, the cloud cover remains a problematic aspect of both numerical and AI-enhanced models. Nonetheless, GFS Graphcast still shows impressive results that are considered fairly accurate.

Figure 3 highlights an increased variance in pressure during the dates when the cyclone was hitting the region. Even though all the models make it clear that there was a hurricane threat, the precision of those predictions fluctuates. It is especially noticeable in the prediction of the GFS model, which is represented by the green line. Figure 4 demonstrates increased wind speed in the region at the same time as the pressure level dropped. Similarly to the previous observation, the GFS demonstrated high variance during the peak of the cyclone activity.

Table 10. Comparison of MAE of Data Variables across Models (Megan)

Data Variable	IFS	AIFS	GFS	GFS GRPH
2m Temp (°C)	1.06	0.36	0.70	0.64
SP & SLP (hPa)	2.36	1.06	2.56	1.14
10m WS (km/h)	10.58	5.58	9.80	7.38
CC (%)	17.98	28.10	20.20	17.07

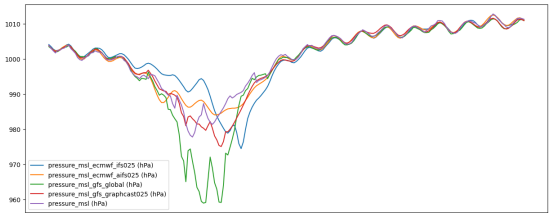

Fig. 3. Sea-level and Surface Pressures during Cyclone Megan

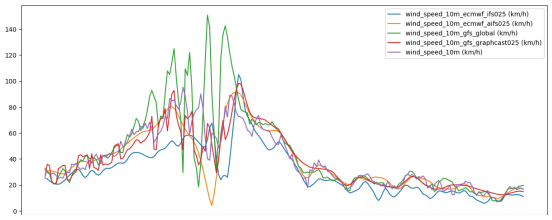

Fig. 4. 10-meter Wind Speed during Cyclone Megan

9 DISCUSSION

This section of the paper addresses each of the sub-questions individually and answers the main research question. Additionally, it presents the limitations of this work and explains why future research is needed.

9.1 Tested models (SQ1)

Due to the global nature of the experiment, using ECMWF IFS, GFS, and their AI-enhanced counterparts, ECMWF AIFS, and GFS Graphcast, was the right choice. Not only, according to the reviewed literature, the models are considered to be some of the best in the world when it comes to global weather forecasting, but they also demonstrated good accuracy across several data variables in different regions of the world. Even though in some cases the models' performance might have been poor, the overall outcome leaves a good impression of their efficiency. Overall, the tested models created a good basis for the experiment and provided meaningful insights.

9.2 Data Variables (SQ2)

The choice of data variables was mostly based on their contribution to hurricane formation and development. Based on the results of the experiment, the chosen data variables give a direct insight into hurricane prediction and development. Thus, the temperature, pressure, wind speed, and cloud coverage were rational choices for model assessments in hurricane prediction scenarios. Nonetheless, the implementation of other variables will have a positive impact on future experiments.

9.3 Regions (SQ3)

The regions chosen for this experiment have an extensive history of hurricane occurrences, making them suitable for the study. However, it is still recommended to conduct further research with the models in other parts of the Caribbean Sea and the Indian Ocean for the validation of this research.

9.4 Models' Performance (SQ4)

The models of the experiment showed a different and diverse range of results. Both NWP and AI-enhanced models showed good results and provided meaningful insights into the final models' performance evaluation. The details of the performances of each model are described in respective subsections.

9.4.1 ECMWF IFS 0.25. The IFS model proved its accuracy and reliability throughout the whole experiment. The model demonstrated high accuracy in all data variable predictions and frequently came out as the one with the lowest MAE. It showed good results in temperature prediction with an average MAE of 0.49. Moreover, the results of surface pressure, sea-level pressure, and wind speed, experiments show that the model does not lose its performance and works just as well as for temperature. Cloud cover is a difficult variable to predict due to its chaotic nature. Nonetheless, the IFS showed fairly good results with 19.26%, which is considered to be within the acceptable range.

However, the model has its flaws and times of uncertainty which were demonstrated in the second study case. The model showed poor performance in the prediction of several data variables, especially in the prediction of wind speed and pressure. Thus, in general, the model performs well but is not error-proof.

9.4.2 AIFS. The IFS model showed fairly accurate results with well-made predictions and relatively low MAE values. It was noticeable that the model lacks consistency and sometimes produces unexpected extreme values that affect the forecasts.

However, it is important to note that the model performed well during the study cases. The performance of the model in the first study case showed that the model has good potential and needs further development for accuracy improvements. The second study case had similar results with only deviation in cloud cover prediction, which is a difficult-to-predict variable. Overall, even though

the model showed good performance, it is clear that further improvements are needed.

9.4.3 GFS. The GFS model came out as the worst-performing model during the experiment. However, the model has both strong and weak points that need to be addressed. For instance, the model had a problem when predicting different data variables in Honduras. Abnormally high MAE values occurred in every data variable in Honduras. Thus, it may be suggested that the GFS finds it problematic to conduct weather predictions in Honduras. At the same time, the model showed good performance during the work on study cases. The model received low MAE values and high graph accuracy that showed its reliability in times of real-life events.

9.4.4 GFS Graphcast. The GFS Graphcast appeared to be a well-performing model with good accuracy and fair reliability. Consistently low MAE values for all data variables across several locations give a reason to rely on the model during daily weather predictions.

The study cases further support the reliable performance of the GFS Graphcast, but this time in real-life scenarios. In both cases, the model gave valuable and fairly accurate information that could be used to predict the cyclones. Thus, it is important to test the model on more cases to evaluate its consistency in weather forecasting and hurricane prediction practices.

9.5 Research Question

Based on the discussions in sections 9.1-9.4 it becomes clear that AI-enhanced models act as strong competitors to classic solutions. They provided several reasons to believe that further development of the technologies may lead to the eventual replacement of the numerical models in the future. Nonetheless, currently, the NWP model ECMWF IFS outperforms the AI-assisted models on almost every level. High accuracy and reliability prove the IFS to be one of the best in the world. Moreover, the classical models tend to output less bias due to their limited work with historical data, which is a potential problem for AI-enhanced solutions.

To conclude, the NWP models remain the most reliable choice when it comes to hurricane predictions around the world, with AI-assisted models holding an opportunity to replace these models in the future due to their increasing accuracy and rapid development.

9.6 Limitations

Some of the key limitations of the study were limited time, limited resources, and a limited number of data variables. The research involved work with several weather data sources for data comparison and analysis. However, this is not enough to draw sharp conclusions about the efficiency of each model. While it is true that the paper gives a good overview of the models' performances in different regions, the key is to go into detail and see how each variable, each location, and each model update influences the final result. Moreover, an external influence that causes massive events such as global warming can have a significant impact on the state of AI-enhanced models in the future. Thus, future research needs to consider that before concluding and providing any strong statements.

9.7 Further Research

This research acts as one of the first steps in the evaluation of AI's influence on hurricane forecasting capabilities. It provides insights into the current state of the models and their performance. Further research with deeper evaluation is crucial for understanding how effective the systems can be in times of need. Moreover, long-term research to validate reliability is important to ensure that the communities at risk are notified from a trustworthy source of data. Finally, an important problem that needs to be addressed in future research is global warming which can directly affect the effectiveness and accuracy of AI-enhanced models that consistently work with historical data that might be not as reliable as in the past.

ACKNOWLEDGMENTS

I would like to thank my supervisor Alessandro Chiumento for answering all of the questions and navigating me during the research process. It allowed me to work efficiently and meet all necessary deadlines.

REFERENCES

[1] AIFS: a new ECMWF forecasting system: https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system.
[2] Bouallègue, Z.B. et al. 2023. Statistical modeling of 2-m temperature and 10-m wind speed forecast errors. Monthly Weather Review. 151, 4 (Apr. 2023), 897–911. DOI: https://doi.org/10.1175/mwr-d-22-0107.1.
[3] Bouallègue, Z.B. et al. 2023. Statistical modeling of 2-m temperature and 10-m wind speed forecast errors. Monthly Weather Review. 151, 4 (Apr. 2023), 897–911. DOI: https://doi.org/10.1175/mwr-d-22-0107.1.
[4] DeMaria, M. et al. 2005. Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). Weather and Forecasting. 20, 4 (Aug. 2005), 531–543. DOI: https://doi.org/10.1175/waf862.1.
[5] Fu, Q. et al. 2019. Multi-Stations' Weather Prediction Based on Hybrid Model Using 1D CNN and Bi-LSTM.
[6] Gong, B. et al. 2022. Temperature forecasting by deep learning methods. Geoscientific Model Development. 15, 23 (Dec. 2022), 8931–8956. DOI: https://doi.org/10.5194/gmd-15-8931-2022.
[7] GraphCast: AI model for faster and more accurate global weather forecasting: 2023. https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/.
[8] HadISD: Sub-daily, non-interpolated weather station data | Climate Data Guide: 2023. https://climatedataguide.ucar.edu/climate-data/hadisd-sub-daily-non-interpolated-weather-station-data.
[9] Heikkilä, M. 2023. Google DeepMind's weather AI can forecast extreme weather faster and more accurately. MIT Technology Review.
[10] Hennayake, K.M.S.A. et al. 2021. Machine Learning Based Weather Prediction Model for Short Term Weather Prediction in Sri Lanka.
[11] Hurricane FAQ - NOAA/AOML: 2024. https://www.aoml.noaa.gov/hrd-faq/.
[12] Hurricane forecasting: NOAA: 2023. https://www.noaa.gov/explainers/hurricane-forecasting.
[13] Hurricane Modeling and Prediction Program - NOAA/AOML: 2023. https://www.aoml.noaa.gov/hurricane-modeling-prediction/.
[14] Hurricane tracking Technology: Advancements and opportunities - IEEE Public Safety Technology Initiative: https://publicsafety.ieee.org/topics/hurricane-tracking-technology-advancements-and-opportunities.
[15] Hurricanes: Science and Society: Dynamical models: https://hurricanescience.org/science/forecast/models/modeltypes/dynamicalmodels/.
[16] Improving Hurricane Forecasts with Near Real-Time Imagery and Data: 2024. https://www.earthdata.nasa.gov/learn/articles/improving-hurricane-forecasts-near-real-time-imagery-and-data.
[17] Lam, R. et al. 2023. Learning skillful medium-range global weather forecasting. Science. 382, 6677 (Dec. 2023), 1416–1421. DOI: https://doi.org/10.1126/science.adi2336.
[18] Lang, S. et al. 2024. AIFS - ECMWF's data-driven forecasting system. arXiv (Cornell University). (Jun. 2024). DOI:https://doi.org/10.48550/arxiv.2406.01465.
[19] Liu, Y. et al. 2021. Evaluation of forecast performance for four meteorological models in summer over northwestern China. Frontiers in Earth Science. 9, (Dec. 2021). DOI: https://doi.org/10.3389/feart.2021.771207.

[20] Nagaraj, P. et al. 2023. Weather Report Analysis Prediction using Machine Learning and Data Analytics Techniques.
[21] Omary, A. et al. 2012. An interactive predictive system for weather forecasting.
[22] Puerto Rico Natural Hazards: Hurricanes | Peligros naturales de Puerto Rico: Huracanes | U.S. Geological Survey: 2022. https://www.usgs.gov/mission-areas/natural-hazards/science/puerto-rico-natural-hazards-hurricanes-peligros-naturales-de.
[23] Saunders, M.A., Klotzbach, P.J. and Lea, A.S.R. 2017. Replicating annual North Atlantic hurricane activity 1878–2012 from environmental variables. Journal of Geophysical Research. Atmospheres. 122, 12 (Jun. 2017), 6284–6297. DOI: https://doi.org/10.1002/2017jd026492.
[24] Sharma, U. and Sharma, C. 2022. Deep Learning Based Prediction Of Weather Using Hybrid_stacked Bi-Long Short Term Memory. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence). (Jan. 2022). DOI: https://doi.org/10.1109/confluence52989.2022.9734133.
[25] The most Hurricane-Prone regions in the world: 2023. https://hurricaneinsider.org/basic-knowledge-about-hurricanes/the-most-hurricane-prone-regions-in-the-world/.
[26] V, V. et al. 2023. Ensemble Machine Learning based Weather Prediction System.
[27] Wardah, T. et al. 2011. Statistical verification of numerical weather prediction models for quantitative precipitation forecast.
[28] Wong, C. 2023. DeepMind AI accurately forecasts weather — on a desktop computer. Nature. (Nov. 2023). DOI: https://doi.org/10.1038/d41586-023-03552-y.

A USE OF AI

During the preparation of this work, the author used Grammarly in order to check and avoid any potential grammatical or structural mistakes within the text. Moreover, the author used ChatGPT to evaluate the readability of the paper. Nonetheless, all of the problematic parts of the paper were edited and re-written by the author.

After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.