VERIFICATION AND ANALYSIS OF AVIATION JET STREAM FORECAST FROM ARTIFICIAL INTELLIGENCE MODELS

Christy Y. Y. Leung[1], C. Y. Leung[2]

1. Introduction

Jet streams are strong, narrow bands of fast-moving upper air flow in the upper troposphere. According to the World Meteorological Organization, the jet stream is a strong narrow current concentrated along a quasi-horizontal axis in the upper troposphere or in the stratosphere characterized by strong vertical and lateral wind shears and featuring one or more velocity maxima. With reference to the Annex 3 of International Civil Aviation Organization, in the flight documentation, forecast areas with wind speeds greater than 40 m/s or 80 knots are highlighted as the jet using wind arrows and jet axis on the significant weather charts (SIGWX charts). On the Earth, the jet most commonly known is (i) the subtropical jet located near the edge of Hadley cell, located around the mid-latitudes (30°N) and (ii) the subpolar jet stream located near the polar cell, located around middle to high latitudes (60°N) [1]. These jets are formed due to differential heating between the tropics and the polar regions. Throughout the year, the location of jets will vary depending on solar insolation and land-sea differential heating, the two jets may even merge and combine into one strong wind band. Due to the rotation of the Earth, the winds in the jets usually travel from west to east and commonly known as strong westerly jet streams.

Jet streams play a crucial role in air navigation. As the cruising altitude of commercial flights is around the tropopause, flights can take advantage of high wind speeds within jet streams to significantly reduce flight time, thus reducing fuel consumption, especially for transcontinental flights. However, it should be noted that the strong wind shear near the jet stream may cause clear air turbulence (CAT), posing safety risks to aircraft and passengers. Climate model projections even suggest an increase in the vertical wind shear in the upper troposphere, possibly indicating an increase in CAT near jet streams [1]. Therefore, accurate forecasting of jet streams is essential for safe and efficient air travel.

In recent years, artificial intelligence (AI) data-driven deep learning models have become increasingly popular in various forecasting tasks and some proven to have higher accuracy or longer predictability ability than traditional Numerical Weather Prediction (NWP) model. The Hong Kong Observatory (HKO) has been implementing real time production of various AI data-driven deep learning models for operational weather forecast applications, including aviation weather forecasts. This study aims to verify, compare and analysis on the performance of three different AI data-driven deep learning models and a traditional NWP model [European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS)] on jet stream forecasting. The strengths and shortcomings of these AI models would be illustrated in this study which hopes to contribute to a better understanding of the usefulness of AI models as guidance in operational aviation weather forecasting.

2. Methodology

The verification analysis is performed by comparing the model forecast outputs to the corresponding analysis from the ECMWF Reanalysis v5 dataset (ERA5) [2], which is treated as the ground truth. The following models are used in the comparison study:

- ECMWF IFS (ECMWF) [3]: Physical NWP model with resolution of 0.125°
- Pangu-Weather (Pangu) [4][5]: AI data-driven model with resolution of 0.25°
- Fengwu [6]: AI data-driven model with resolution of 0.25°
- Fuxi [7]: AI data-driven model with resolution of 0.25°

Multiple metrics are used for determining the model performance on aviation jet streams. Further details will be discussed in the corresponding subsections below.

[1]Hong Kong Observatory, corresponding email: yyleung@hko.gov.hk
[2]The Chinese University of Hong Kong

2.1. Data Collection

The verification period in the study spans across around a month in the summer of 2024, from 23 May to 13 June 2024. The verification was performed against the 00 UTC and 12 UTC of ERA5 within the study period. The verification covers 6 different forecast hours (T+N), where N = 12, 24, 36, 48, 60, 72, and 7 different vertical levels (hPa), including 100, 150, 200, 250, 300, 400, 500.

For example, the ground truth of 23-05-2024 00:00 UTC was compared against the 12-hour forecast initialised by 22-05-2024 12:00 UTC run, the 24-hour forecast initialised by 22-05-2024 00:00 UTC, the 36-hour forecast initialised by 21-05-2024 12:00 UTC, the 48-hour forecast initialised by 21-05-2024 00:00 UTC, the 60-hour forecast initialised by 20-05-2024 12:00 UTC, and the 72-hour forecast initialised by 20-05-2024 00:00 UTC.

2.2. Verification Metrics

To provide a comprehensive analysis, different types of verification metrics are defined, including continuous variables, metrics for the jet stream position, metrics for jet stream coverage as an object, etc.

2.2.1. Metrics for Continuous Variables

To evaluate the performance of different models in predicting wind speed and wind direction within jet streams, various continuous variables are obtained from the output, including the U-component and V-component of wind, wind speed magnitude and wind direction. It should be noted that as the focus is on jet stream forecast performance, the verification was only calculated when the data points of either the ground truth or the forecast model showed a wind speed of over 80 knots, which are defined as part of the jet streams. The Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) are calculated for the four variables stated above, denoted as:

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

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

where N represents the total number of points taken into calculation, y_i represents the true values, and $\hat{y}_i$ represents the predicted values. Verification would only be performed when the magnitude for either y_i or $\hat{y}_i$ is above 80 knots. Lower RMSE or MAE indicates that the model perform better in predicting wind speed and wind direction.

2.2.2. Metric for Jet Stream Position

To evaluate whether the model can accurately forecast the position and area of the jet stream, the metric dice loss is used to provide an assessment of the overlapped area of jet stream. Dice loss is denoted as:

$$Dice\,loss = 1 - \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} = 1 - \frac{2\sum_{i=1}^{N}(y_i \cdot \hat{y}_i)}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon}$$

where Y and $\hat{Y}$ are the set of points outputted as part of a jet stream (magnitude >80 knots) by the forecast model or the ground truth. N represents the total number of points, while the epsilon (ϵ) is a small constant added to the denominator to avoid division by zero. y_i and $\hat{y}_i$ are denoted as:

$$y_i = \begin{cases} 1, & \text{if predicted magnitude} > 80 \text{ knots} \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \begin{cases} 1, & \text{if true magnitude} > 80 \text{ knots} \\ 0, & \text{otherwise} \end{cases}$$

with each point on the grid taken into the calculation. Lower dice loss indicates the model can better predict the area and location of the jet stream.

2.2.3. Object-based Verification

To evaluate different characteristics of jet streams (e.g. centroid, orientation, etc), object-based verification was employed to analyze the model performance on predicting the jets.

Resolving Objects Objects are resolved following two steps: thresholding and splitting. Only data points with wind speed >80 knots are kept to create a mask of jets. Then each connected piece in the mask is counted as a simple object. To study the performance of major jet stream, we also defined the ground truth objects with area >200 data points (about $3.5^o \times 3.5^o$ or 154000 km^2) as major objects, representing major jet streams that is more essential in flight planning.

Matching Objects To perform evaluation based on individual objects, it is necessary to match objects between model output and ground truth. An interest function $T(g, f)$ is defined between any ground truth object g and forecast object f:

$$T(g,f) = \begin{cases} \frac{(a_{g,f}+1)(r_{g,f})}{d_{g,f}} & \text{if } a_{g,f} \neq 0 \text{ or } d_{g,f} \leq 12.5^\circ \\ 0 & \text{otherwise} \end{cases}$$

where $a_{g,f}, d_{g,f}$ are the intersection area and centroid distance between g and f respectively. $r_{g,f}$ is the ratio of the two objects' areas. Each forecast object is matched to the ground truth object with the highest interest. If all the interest values are zero, then the forecast object will not be matched with any ground truth objects, and will be counted as "unmatched". For the case of multiple forecast objects being matched to one ground truth object, the involved forecast object will be merged into one object. On the other hand, if a ground truth object is not matched to any forecast objects, it will also be counted as "unmatched".

This method matches and merges objects without modifying the ground truth at all, ensuring fairness when comparing different models.

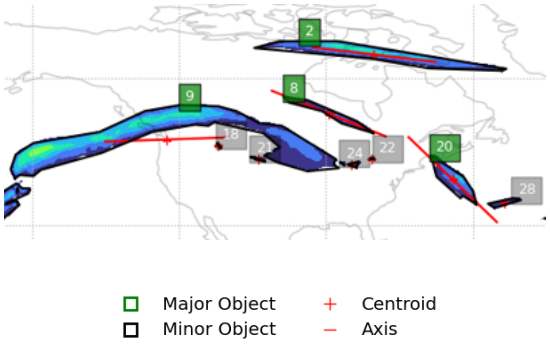


Figure 1: Example of resolved objects.

Metrics for Object-based Verification Various metrics are used in object-based verification to ensure multiple dimensions in the verification process:

1. **Centroid distance**. The mean Euclidean distance of centroids between matched objects for each model are calculated to examine its performance on predicting jet stream object location accurately. To provide additional insight, the average poleward and equatorward shifts of jets are also calculated. The centroid distance are computed in two ways, one with all the objects, and another with major objects only. This is to evaluate the model's capability in predicting major jets. Lower centroid distance shows that the model can better predict the jet object location.

2. **Orientation error**. The axis of each object are calculated by maximizing its second moment. The mean angular error between the axis of matched objects are then determined. As the axis of the smaller objects are rather fluctuating and may not lead to conclusive result, only major objects are taken in the calculation. Lower orientation error indicates the model can predict the jet orientation more accurately.

3. **Overlapping area ratio**. To compute whether two objects are similar, the overlapping area ratio are calculated. The ratio is given by
$$\frac{\sum a_{g,f}}{\sum a_f}$$
where $\sum a_{g,f}$ is the sum of intersection areas of matched objects, while $\sum a_f$ is the total area of forecast objects. All matched objects are taken into the calculation. The range of the ratio is from 0 to 1. Ratio that are closer to 1 means that the predicted jet area better resembles the actual jet area.

4. **Maximum magnitude error**. To evaluate the intensity of jets, the mean difference of maximum magnitude in each object are calculated. The calculation are computed in two ways, one with all the objects and one with major objects only. Lower error indicates higher performance for the model in predicting the intense and extreme values of the jets.

5. **Unmatched objects**. The number of unmatched ground truth objects and unmatched forecast objects are collected respectively. All objects, regardless of major or not, are taken into the calculation. The total area of these unmatched objects is also computed. A lower number or lower area of unmatched objects indicates that the model can make a more precise prediction.

3. Results

This following section presents analysis for selected cases of the verification results across the forecast models. The comparison is made across (1) forecast hours and (2) vertical levels.

3.1. Results by Forecast Hour

3.1.1. Continuous Variables

From Figure 2, it is observed that the RMSE increase with increasing forecast hours, indicating that the forecast becomes less predictable as time increases. Comparing the results for the four models, it is observed that ECMWF have the highest RMSE as compared with all the other deep learning based models. Even if increasing forecast hours, the performance of the ranking for the four models remains unchanged. The ranking are as follows: Fengwu >Fuxi >Pangu >ECMWF, with ECMWF performing the worst and Fengwu performing the best. Another observation is that the RMSE difference among the models are increasingly large as forecast hour increase. The RMSE of ECMWF increases more rapidly through forecast hours than FengWu, as well as the other deep learning models. Similar trend is observed for mean absolute error (MAE) for wind speed and wind direction (Figure 3).

Figure 4 shows the comparison of Fengwu and ECMWF forecast as compared with the ground truth for a selected case. Fengwu outperformed ECMWF in terms of RMSE or MAE. We can observe that Fengwu can predict the shape for the jet near Japan while ECMWF covers a much smaller area. For the jet south of South Africa, ECMWF overestimated the magnitude of jet when compared with ground truth, while Fengwu is similar to ground truth. This case study shows the subtle difference in the forecasts from ECMWF and Fengwu.

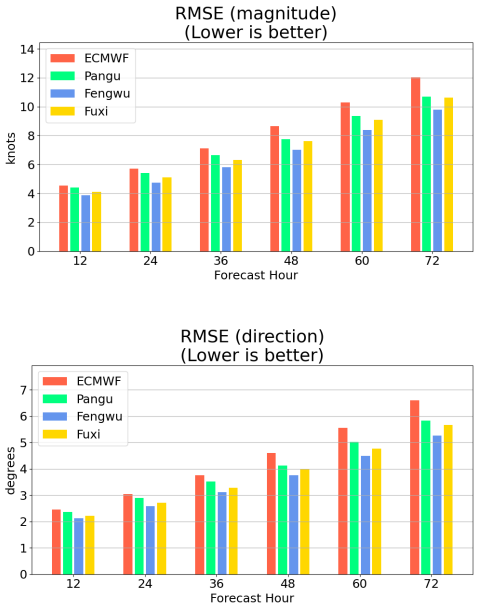


Figure 2: Root mean square error (RMSE) of wind speed (top) and wind direction (bottom) across forecast hour.

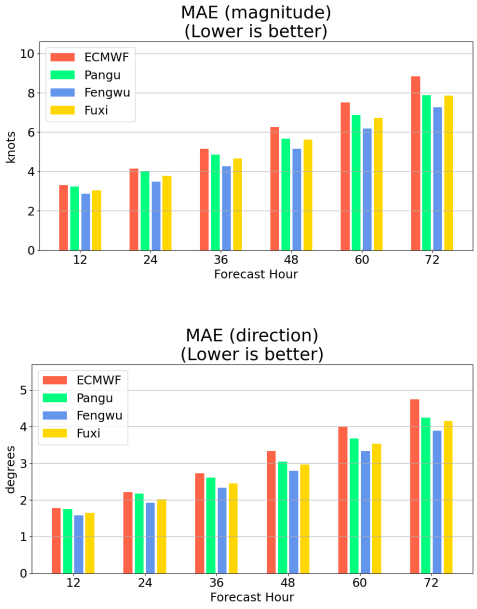


Figure 3: Mean absolute error (MAE) of wind speed (top) and wind direction (bottom) across forecast hour.

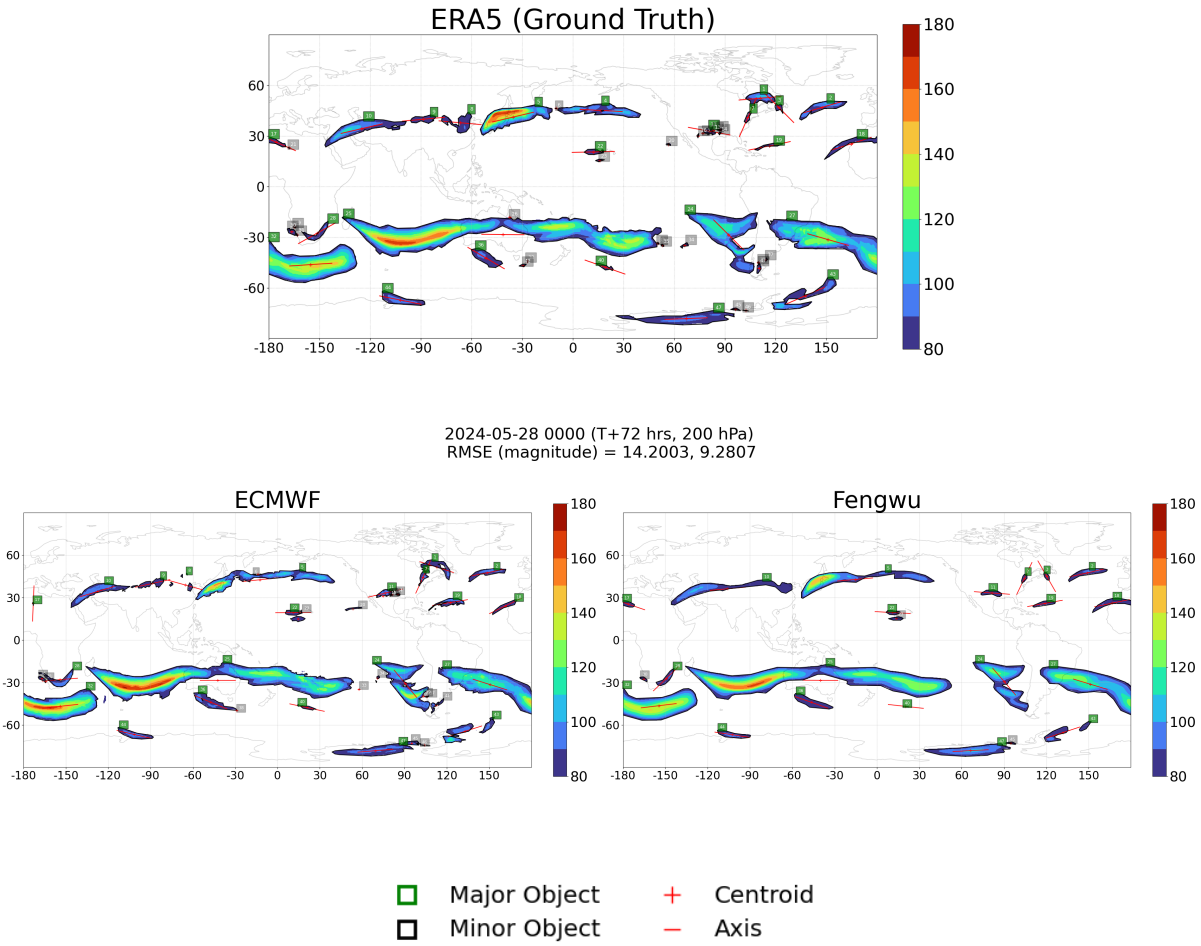


Figure 4: ERA5 ground truth (top), ECMWF 72 hour forecast (bottom left) and Fengwu 72 hour forecast (bottom right) for 200hPa level for valid time at 00Z 28/05/2024. The contours show the magnitude of wind speed in knots.

3.1.2. Centroid Distance Error and Orientation Error

Figures 5 and 6 show the performance comparison of two attributes of jet stream objects: centroid and orientation. The model performances for centroid distance and orientation error are similar to RMSE/MAE metric, Fengwu >Fuxi >Pangu >ECMWF, Fengwu still performs the best with minimum error. For the centroid distance for major objects, it is observed that Fuxi perform slightly better than Fengwu for T + 12, but overall speaking, Fengwu is still the best performing model for centroid distance error.

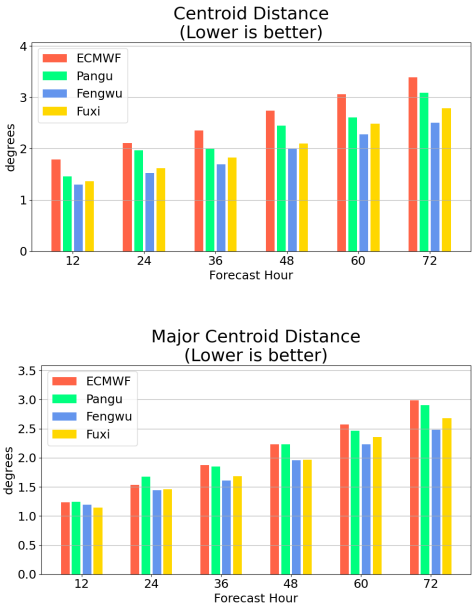


Figure 5: The Centroid Distance Error for all objects (top) and only major objects (bottom) across forecast hour.

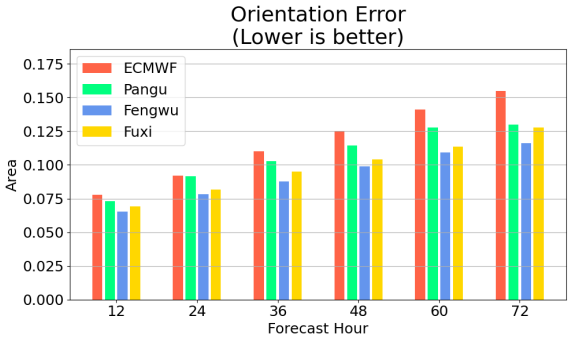


Figure 6: The Orientation Error across forecast hour.

Figure 7 shows the jet stream object from all the 4 models. All the models predict the position of the jets fairly accurately. However, more discrepancy with the ground truth was observed from ECMWF model, in particular to the jets south of New Zealand and near the south-east Pacific Ocean. The predicted shape of jets from the three deep learning models resembles the ground truth more as compared with ECMWF.

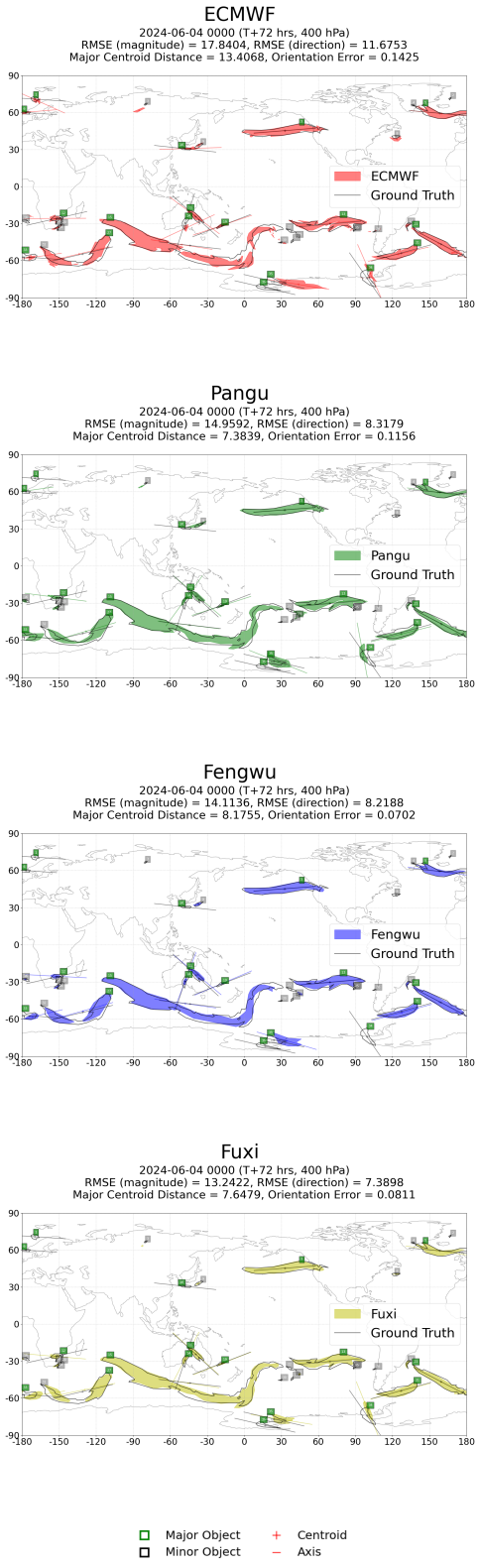


Figure 7: Jet stream object forecasts from all the 4 models overlaid with the ground truth object (in black line), showing the 72 hour forecast for 400 hPa for valid time at 00Z 04/06/2024.

3.1.3. Overlapping Area

Figure 8 shows the overlapping area ratio of the 4 models against the ground truth and a decrease in the ratio is observed across forecast hours. The degrade of predictability across forecast hours is evident in Figure 9. As forecast hours increase, ECMWF fails to predict the shapes and positions of the jets, with a lot of non-overlapped area while Fengwu can keep the shape of the jets even up to 72 hour forecasts.

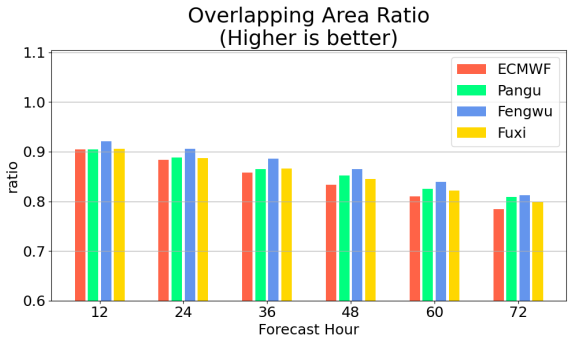


Figure 8: Overlapping area ratio across forecast hour.

3.1.4. Maximum Magnitude Error

Although the previously discussed metrics and examples all show that deep learning models (in particular Fengwu) outperform ECMWF on average, it is inappropriate to claim that Fengwu is the best performing model. During the verification, we have found two metrics for which ECMWF performed significantly better than the other models: maximum magnitude error, and unmatched objects. Figure 10 show the absolute maximum magnitude errors for all the 4 models across forecast hour. It is observed that the performance of ECMWF differs significant between all objects and only major objects. ECMWF obtained a similar level of maximum magnitude error for major objects as compared with Fengwu in the T+12 and T+24 and even outperformed the deep learning models from T+36 onwards.

From the bias chart (Figure 11), it is observed that all deep learning based models tend to underestimate the maximum magnitude of the jets with a negative bias, while ECMWF having a smallest bias both for all objects or major objects. This may possibly be related to the fact the deep learning models are built on transformers trained with past reanalysis via minimizing a loss function, this would tend to produce smooth results with lower average error but the extremes cannot be well captured or being smoothed out.

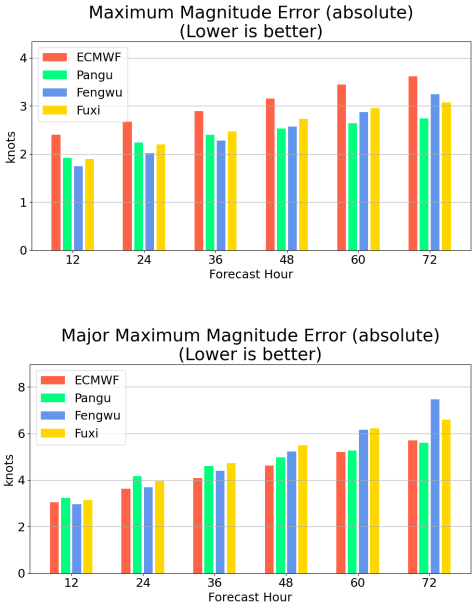


Figure 10: Absolute maximum magnitude error for all object (top) and only major object (bottom) across forecast hour.

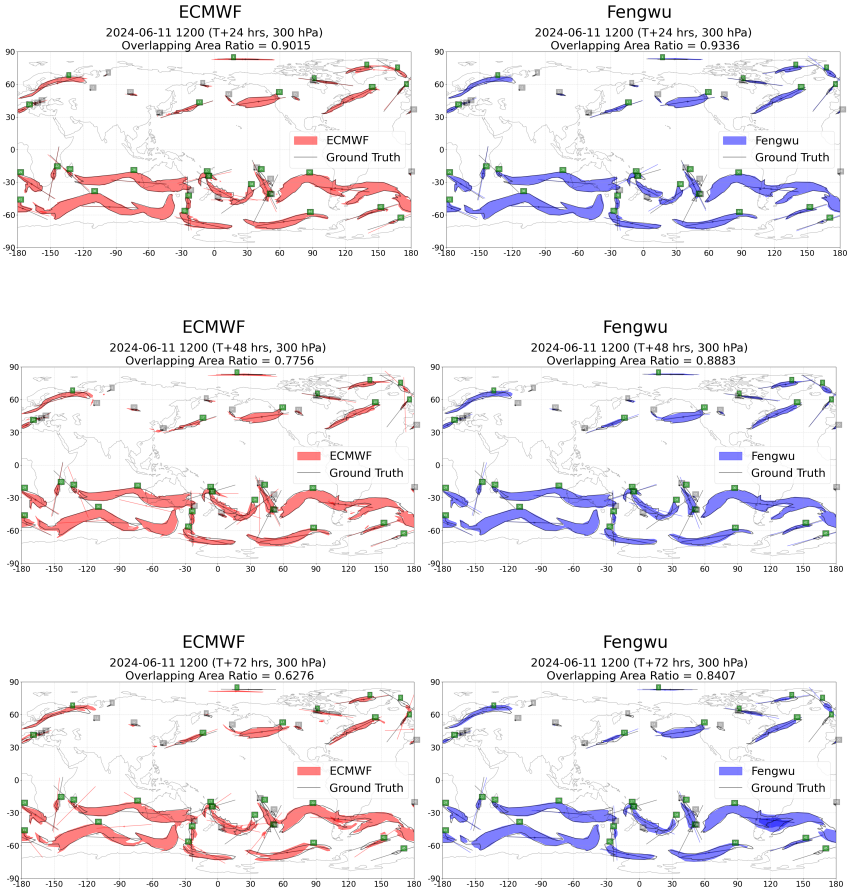


Figure 9: Jet stream forecasts from ECMWF (left) and Fengwu (right) overlaid on ground truth valid on 12Z 11/06/2024 at 300 hPa. The top, middle and bottom row show the 24, 48 and 72 hours forecast respectively.

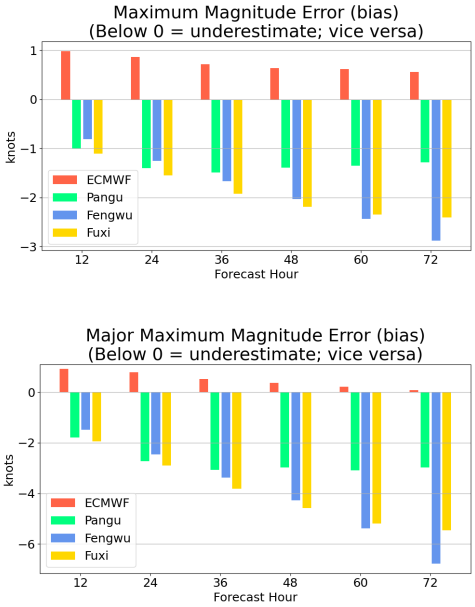


Figure 11: Maximum magnitude error (bias) for all object (top) and only major object (bottom) across forecast hour.

3.1.5. Unmatched Objects

The issue of overly smooth outputs from deep learning models also leads to its relatively inferior performance on unmatched ground truth objects. Smaller jet objects may be ”smoothed out” in deep learning models, thus are unable to capture the weak and small jets available in ERA5. Figure 12 show that ECMWF have a lower unmatched ground truth objects and area as compared with other deep learning models.

However, when investigating unmatched model objects, ECMWF has the highest number of unmatched model objects as area (Figure 13). This

implies that ECMWF might be forecasting jet objects that do not appear in ground truth. The case study in Figure 14 showed that the small jet objects over the Middle East area is only captured in ECMWF but not in Pangu, while ECMWF have additional jet areas highlighted near Hudson Bay over Northeast Canada and the south Indian Ocean near Antartica.

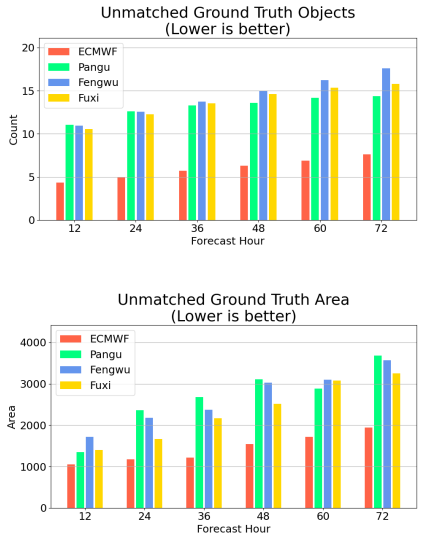


Figure 12: Unmatched ground truth objects (top) and area (bottom) across forecast hour.

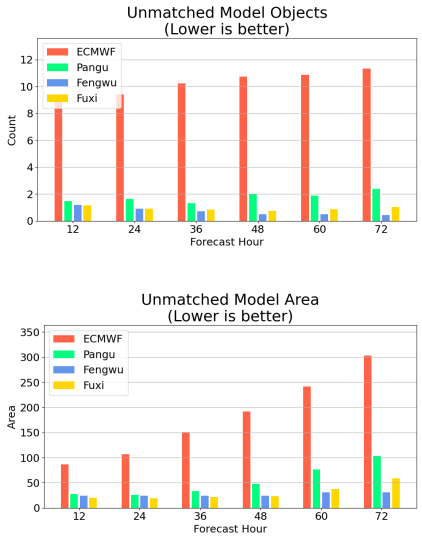


Figure 13: Unmatched model objects (top) and area (bottom) (based on pixel area where each pixel width is 0.25º) across forecast hour.

3.2. Results by Vertical Level

Verification was also performed for different vertical levels and found that the models perform differently across vertical levels. Figure 15 shows some selected metrics not involving magnitude of wind speed, including orientation error and overlapping ratio, by vertical level. It is observed that for these metrics, all models tend to perform better on levels where with more intense jets , such as 200-300 hPa. On the other hand, when the selected metric involves the magnitude, the errors are higher for those levels, as seen in Figure 16. This can possibly attribute to the model capability in forecasting the area or coverage of stronger jets, but not necessarily the actual magnitude or maximum value. The shortcoming of all the models mainly lies on the coverage of weaker jets and the maximum magnitude of intense jets.

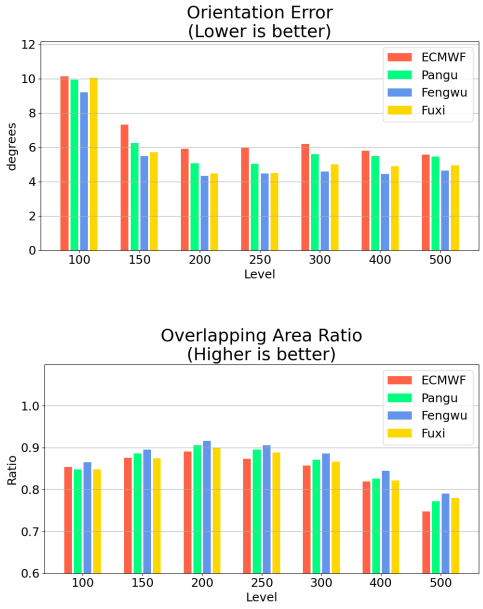


Figure 15: The orientation error (top) and overlapping area ratio (bottom) across vertical level in hPa.

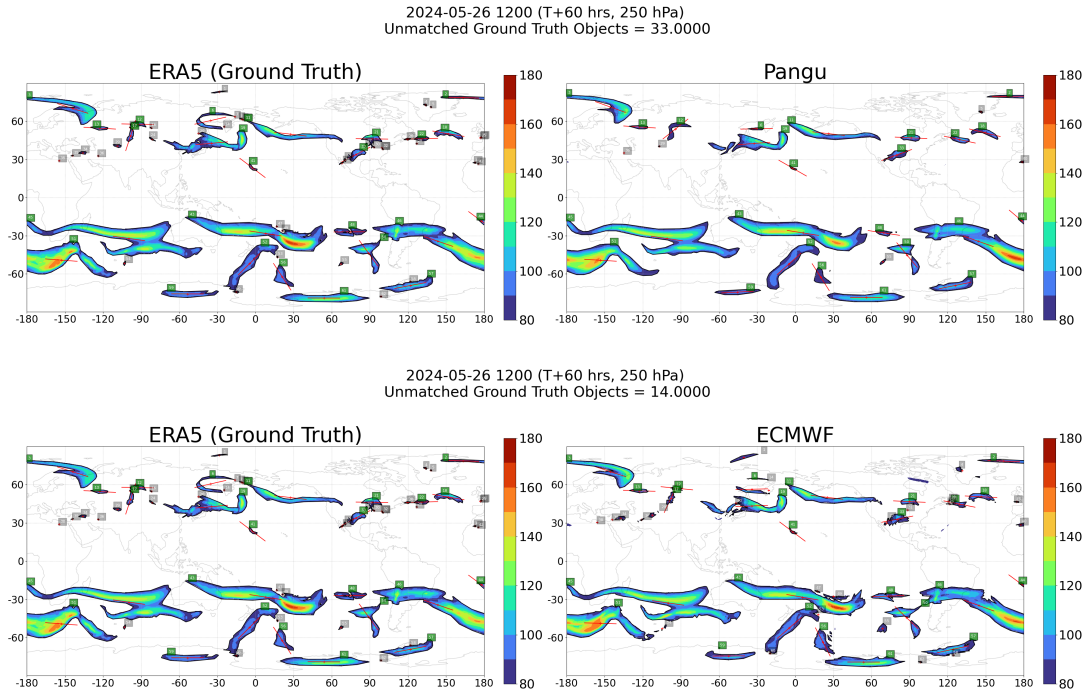


Figure 14: Ground Truth (left), Pangu 60-hour forecast (top right) and ECMWF 60-hour forecast (bottom right) for valid time 12Z 26/05/2024 at 250 hPa. The contours show the magnitude of wind speed in knots.

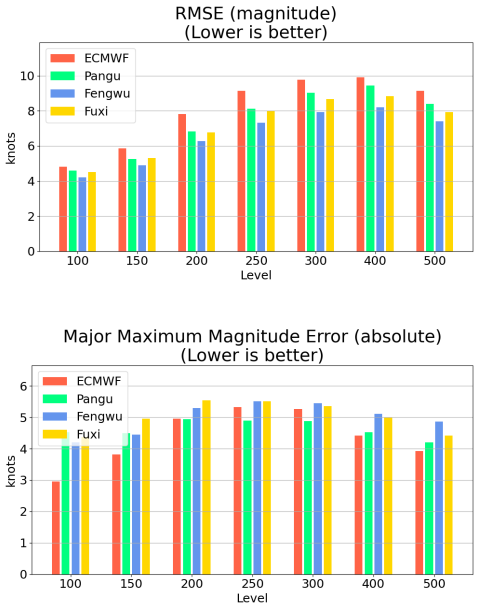


Figure 16: The RMSE of wind speed (top) and the absolute maximum magnitude error for major objects (bottom) across vertical level in hPa.

4. Conclusion

This short study attempts to evaluate the performance of aviation jet stream forecasts for three different deep learning medium range global models and a traditional physical Numerical Weather Prediction (NWP) model. The study reveals that the AI data-driven models outperformed the traditional NWP models in most of the metrics, including the continuous variables or object-based verification metrics. Among the AI models, Fengwu is more superior that the other two deep learning models (Pangu and Fuxi).

However, due to the nature of the transformers utilised in training the AI models, these models tend to produce overly smooth results and cannot capture the extreme or finer details of jets, for examples, the maximum intensity of jet or small and weak jet objects. Since both the coverage of jets and the absolute magnitude of jets are crucial in the flight planning process, each type of model still have their own advantage over the other type at this stage. In operational aviation weather forecasting, a combination of forecasts using both model types may lead to the best re-

sult. However, further study or post-processing are required to maximize the applications of AI models on aviation meteorology.

References

[1] M. Stendel, J. Francis, R. White, P. D. Williams, and T. Woollings, "The jet stream and climate change," p. 327–357, 2021.

[2] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, and J.-N. Thépaut, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, pp. 1999–2049, 2020.

[3] "Medium-range forecasts," https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts, [Accessed 31-12-2024].

[4] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast," 2022. [Online]. Available: https://arxiv.org/abs/2211.02556

[5] ——, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[6] K. Chen, T. Han, J. Gong, L. Bai, F. Ling, J.-J. Luo, X. Chen, L. Ma, T. Zhang, R. Su, Y. Ci, B. Li, X. Yang, and W. Ouyang, "Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead," 2023. [Online]. Available: https://arxiv.org/abs/2304.02948

[7] L. Chen, X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, "Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast," *npj Climate and Atmospheric Science*, vol. 6, 2023.