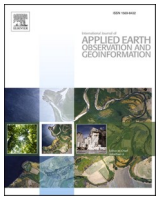
International Journal of Applied Earth Observation and Geoinformation 139 (2025) 104473

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Spatiotemporal inhomogeneity of accuracy degradation in AI weather forecast foundation models: A GNSS perspective☆

Junsheng Ding [a], Wu Chen [a], Junping Chen [b,c,*], Jungang Wang [d,e], Yize Zhang [b], Lei Bai [f], Yuyan Wang [a], Xiaolong Mi [a], Tong Liu [a], Duojie Weng [a]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong 999077, China
[b] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[c] School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China
[d] Technische Universität Berlin, Institut für Geodäsie und Geoinformationstechnik, Berlin 10623, Germany
[e] Department of Geodesy, GeoForschungsZentrum (GFZ), Potsdam 14473, Germany
[f] Shanghai AI Laboratory, Shanghai 200030, China

ARTICLE INFO

Keywords:
Foundation models
GNSS tropospheric delay
Spatiotemporal inhomogeneity
Accuracy degradation
Weather forecast

ABSTRACT

The artificial intelligence (AI) weather forecast foundation models can infer and generate precise global atmospheric state forecasts on the user's device and with speed over 10,000 times faster than the operational Integrated Forecasting System (IFS), and it is making increasingly significant contributions to geodetic applications represented by the Global Navigation Satellite System (GNSS). However, existing studies on the investigation of these AI models are typically carried out by concentrating on specific one or several meteorological events in certain regions or by comparison with physical models, and the evaluation results obtained in this manner are not comprehensive and universal. Additionally, we find that the results obtained by the foundation models through the "rollout" method for forecasting are not uniform in terms of time and space. This temporal and spatial inhomogeneity of accuracy and accuracy degradation are related to AI algorithms and attributes of training data, etc., but these characteristics have not been thoroughly explored and analyzed. In this study, we obtained the global forecast results of foundation models for 2022 and subsequently derived the GNSS tropospheric delay through numerical integration. We calculated the mean deviation, mean absolute error, and root mean square error of these data. Using these metrics, we analyzed the spatiotemporal inhomogeneity in the accuracy degradation of foundation models, represented by Huawei Cloud Pangu-Weather, Google DeepMind GraphCast, and Shanghai AI Lab FengWu. We evaluated how this inhomogeneity changes with forecast time and identified the best-performing models across different regions and forecast durations. From the results, we find that taking topography into account when training the model enhances its accuracy at high altitudes, and the facilitating influence between the high related atmospheric variables such as precipitation and water vapor. The contributions of this study are twofold: it serves as a valuable reference for geodetic and remote sensing users employing foundational models, and offers insights and case supports for AI practitioners aiming to develop more accurate models for weather forecasting.

1. Introduction

Over the past two years, the technological revolution driven by artificial intelligence (AI) has permeated various industries, exerting profound impacts across numerous sectors globally (Bommasani et al., 2021; Moor et al., 2023). Leading technology giants such as NVIDIA, Google, Microsoft and Huawei have actively entered the field, achieving significant breakthroughs in areas such as computer vision (CV), multimodal learning, and scientific computing (Puladi et al., 2023; Feng et al., 2024). Numerical weather prediction (NWP) is one area where advances in AI technology are especially notable (Bonavita 2024; Charlton-Perez et al., 2024). Weather forecast foundation models like FourCastNet, Pangu-Weather and GraphCast, have shown to be able to execute inference locally on user devices and outperform traditional

☆ This article is part of a special issue entitled: 'Foundation models EO' published in International Journal of Applied Earth Observation and Geoinformation.

* Corresponding author.
E-mail address: junping@shao.ac.cn (J. Chen).

https://doi.org/10.1016/j.jag.2025.104473
Received 12 January 2025; Received in revised form 15 February 2025; Accepted 7 March 2025
Available online 21 March 2025

methods over 10,000 times faster (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023). These models have achieved forecasting accuracy comparable to, or even better than, the Integrated Forecasting System (IFS) and have been integrated into the operational suite of the European Centre for Medium-Range Weather Forecasts (ECMWF), which now offers these services to the public (Bi et al., 2023; Chen et al., 2023a). Due to their superior performance, AI-based weather forecasting foundation models have garnered significant attention from scholars in the meteorological community, prompting extensive discussion, evaluation, and in-depth research (Hsu et al., 2024; Feldmann et al., 2024a; Yan et al., 2024).

Scholars have extensively evaluated AI foundation models from diverse perspectives. A study by Feldmann et al. (2024a) shows that the spatial structure of AI models is smoother compared to IFS and ERA5 (ECMWF Reanalysis v5). In the global seasonal analysis, GraphCast and Pangu-Weather have the highest performance, which is comparable to or even exceeds the IFS. Hsu et al (2024) evaluated five foundation models from synoptic-scale prediction, typhoon predictions as extreme weather cases, and local rainfalls induced by a typhoon. They concluded that FengWu is the best performing model, followed by FuXi and GraphCast, with FCN2 and Pangu-Weather at the bottom of the list. There are also studies that are pessimistic about AI models, such as Bonavita (2024), who argues that the three current leading AI models, Pangu-Weather, FourCastNet, and GraphCast, are not able to properly reproduce sub-synoptic and mesoscale weather phenomena and lack the fidelity and physical consistency of physics-based models. Feldmann et al. (2024b) used the example of tornado outbreaks in the United States and argued that the spatial structure of the AI model is smoother compared to the IFS and the reanalysis of the ERA5, which all exhibit different biases in their predictions. In addition, there are also some more neutral views, such as Shu et al. (2024) highlighted the integration of AI with traditional meteorological techniques as promising to improve the accuracy of weather forecasts, and they argued that synergistic effects have made large-scale models (foundation models) pivotal in the evolving field of weather forecasting. Olivetti and Messori (2024) showed that data-driven models may already now be a useful complement to physics-based forecasts in those regions where they display superior tail performance, but that some challenges still need to be overcome before widespread operational implementation can take place.

These studies mentioned above can already give us a relatively clear picture of AI models. However, most of the aforementioned studies evaluate specific weather or meteorological events in certain regions. Furthermore, these evaluations have three main limitations: 1). The results are often one-sided and lack comprehensive spatial and temporal coverage, making it difficult to provide a comprehensive and universal assessment of these models. This is one reason why the performance of these models varies across different studies. 2). There is a lack of analysis on how differences in algorithms, the number of pressure levels in datasets, and the types and quantities of atmospheric variables used by different models impact their performance. 3). These researches do not account for the accumulation of errors in multi-step forecasts, nor do they consider the spatiotemporal inhomogeneity of accuracy degradation. Given the vast amount of data generated by AI model outputs, overcoming the aforementioned limitations is not an easy task. Fortunately, tropospheric delays in the field of satellite navigation can show them in a reduced dimension, thus being able to reflect the overall performance of these AI models. To comprehensively capture the spatiotemporal inhomogeneity of AI model accuracy, we designed an experiment leveraging tropospheric delay data from the Global Navigation Satellite System (GNSS).

In this study, we selected the three most representative AI weather forecast foundation models and used the ERA5 data for the entire year of 2022 as the initial input. We performed a 60-step (15-day) inference and then transformed the results into tropospheric zenith delays through numerical integration. We computed the three metrics of the tropospheric delays and their inter-step differences. Finally, we analyzed the variation of these metrics with forecast time, longitude, latitude, and height, and discussed the factors leading to the experimental results. We present the data and methods in Section 2, as well as giving the variation of AI model accuracy with forecast time. The regions and times when the different models perform best are given in Section 3, and the spatial and temporal distribution characteristics are analyzed with respect to the degree of degradation of the accuracy, and finally conclusions and perspectives are given in Section 4. Due to space constraints, some detailed experimental procedures and results are provided in the supplementary information. We recommend reading this article in conjunction with the supplementary information.

2. Data and methods

In this section, we initially present the AI weather forecasting foundation models, concentrating on three representative foundation models employed in this research, namely Pangu-Weather, GraphCast, and FengWu. Subsequently, we introduce the operation modes of these foundation models and the methodology for obtaining the tropospheric delays in the GNSS view as delineated in this research, and a general flowchart is provided. Finally, we introduce three accuracy metrics and present the results of these three metrics the inter-step differences of these metrics along with the variation with forecast time, and analyze the results.

2.1. AI weather forecast foundation models

AI foundation models exhibit remarkable prowess in handling extensive volumes of meteorological data and identifying intricate patterns and trends, which has revolutionized the accuracy and timeliness of weather forecasting. AI models such as FourCastNet (Pathak et al., 2022), Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FengWu (Chen et al., 2023a), FuXi (*Chen et al., 2023b*), ClimaX (Nguyen et al., 2023), have emerged in the past two years, among which the first three are already accessible on the ECMWF official website as part of the ECMWF's operational suite. On 3 June 2024, the ECMWF also released preprints of their self-developed AI versions of IFS (AIFS) (Lang et al., 2024). AIFS products are now obtainable on ECMWF's OpenCharts (https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system). We have compiled the information of these models within Tables S1 to S3 and Fig. S3 in the supporting information, which encapsulate the models' profiles regarding their release dates, affiliations, resolutions, atmospheric variables supported, pressure levels covered, and so on. The following is a brief description of the three models utilized in this paper.

Pangu-Weather, developed by Huawei Cloud, is constructed on the basis of a 3D Earth-Specific Transformer (3DEST), which can effectively handle complex 3D meteorological data and minimize iteration error (Bi et al., 2023). GraphCast, put forward by Google DeepMind, is a state-of-the-art weather forecasting foundation model based on graph neural networks (GNN) technology. Its "encode-process-decode" modeling structure demonstrates outstanding performance, surpassing Pangu-Weather on 99.2 % of its targets (Lam et al., 2023). FengWu, proposed by the Shanghai Artificial Intelligence Laboratory, is yet another advanced weather forecast foundation model. FengWu is founded on multi-modal and multi-task deep learning methods and performs more favorably than GraphCast in predicting 80 % of the 880 reported predictands (Chen et al., 2023a). It is important to highlight that GraphCast incorporates an additional single-level weather variable, total precipitation (TP), distinguishing it from the other two models. Precipitation is characterized by its discontinuity in both temporal and spatial dimensions, rendering it one of the most challenging weather variables to predict. Moreover, due to known biases in ERA5 precipitation data (Lavers et al., 2022), GraphCast does not assess the predictive performance of total precipitation. However, this does not imply that total

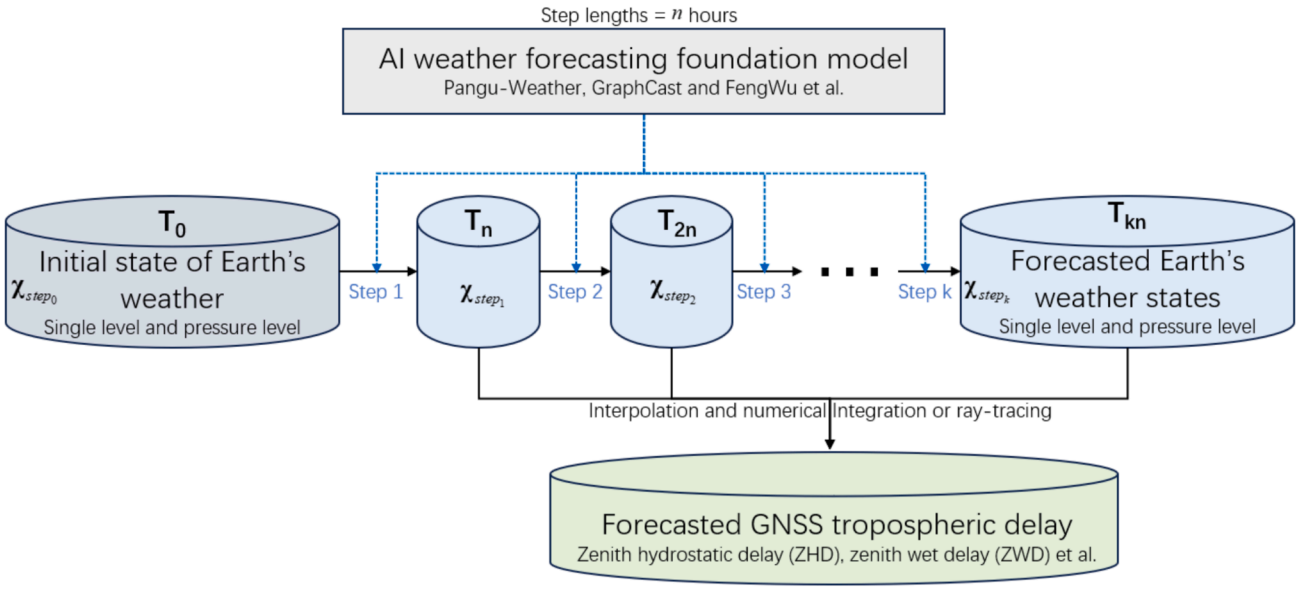


Fig. 1. A general flowchart for calculating forecasted GNSS tropospheric delay via AI weather forecast foundation model inference. The bold T represents the time, its subscript is the forecast time length relative to the initial state time, n is the step length, and k is the number of inference steps. In this research, n = 6 h and k = 60 steps.

precipitation is an ineffective variable. Instead, it can enhance the prediction accuracy of other related variables, and we will elaborate on this in subsequent sections.

We selected Pangu-Weather, GraphCast, and FengWu as representative models for this research for the following three reasons: (1) Pangu-Weather is the first artificial intelligence weather forecast foundation model to outperform the traditional numerical forecasting methods (operational IFS) in terms of medium- and medium-to long-term weather forecasting accuracy (Bi et al., 2023). The meteorological and artificial intelligence communities have been quite interested in it since it was published in the journal Nature. (2) GraphCast and FengWu are the only two models that currently support the full 37-layer pressure level NWM data input (see Fig. S3 in the supplementary information). Even the latest release of AIFS from ECMWF only supports 13 layers of pressure level, and the sparse pressure level layers will introduce systematic errors in the interpolation calculation of the GNSS tropospheric delay (Ding et al., 2024a; Ding et al., 2024b). (3) These three models have common data structures and commonly supported pressure levels and all support the computation of tropospheric delays, but use their unique AI algorithmic structure. Note that this research uses data from 2022 as the initial input for the AI models, which are not used in model training and hyperparameter tuning (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023a). Such an experimental setup ensures a fair and accurate assessment of the models' performance and avoids the risk of overfitting and data leakage.

2.2. The operation mode of foundation models and the data processing flow in this research

The AI weather forecast foundation models use "rollout" mode to perform multi-step inference. This means that, with the exception of the first step, which calls for an external initial state of Earth's weather, the input for each subsequent step of inference is the result of the step before it, as indicated by equation (1):

$$\begin{cases} \chi_{step\ i} = Model(\chi_{step\ i-1}) \\ \chi_{step\ i+1} = Model(\chi_{step\ i}) \end{cases}, i = 1, 2, 3... \quad (1)$$

Additionally, there are AI weather forecast foundation models like GraphCast and FengWu whose inputs require the Earth's weather state at two neighboring epochs. The inference mode of such models is shown in equation (2):

$$\begin{cases} \chi_{step\ i} = Model([\chi_{step\ i-1}, \chi_{step\ i-2}]) \\ \chi_{step\ i+1} = Model([\chi_{step\ i}, \chi_{step\ i-1}]) \end{cases}, i = 1, 2, 3... \quad (2)$$

where χ is the Earth's weather state and '*Model*' refers to the foundation model, in this paper it is Pangu-Weather, GraphCast or FengWu. Moreover, i is an integer between 1 and 60 in the experiments in this study. This means that a 15-day forecast, or a 60-step inference, is carried out with a step length of 6 h. We selected the 6 h step length for the experiment because, apart from Pangu-Weather, all other models only support 6 h interval (see Table S2).

Fig. 1 shows a general flowchart for calculating forecasted GNSS tropospheric delay via AI weather forecast foundation model inference, taking a single epoch input mode of Eq. (1) as an example. The experimental data presented in this work, specifically the predicted tropospheric delays, were derived from the computations illustrated in Fig. 1. The method of calculation of zenith hydrostatic delay (ZHD) and zenith wet delay (ZWD) are given in the supporting information (see Eq. S1–S3 in the supplementary information, note that an equation or figure number with the letter S means it is in the supporting information.), and a more detailed description is also available from Hofmeister, (2016). We employ ERA5 as the initial Earth's atmospheric state input in the AI model inference, and the assessment is likewise based on the ERA5 results (on the global grids). Additionally, we have chosen to use the Nevada Geodetic Laboratory's (NGL) GNSS tropospheric delay product as a reference (on the GNSS site). Due to their similar accuracy to that of the International GNSS Service (IGS), tropospheric products from the NGL provide a reliable reference for evaluating general tropospheric models (Ding and Chen, 2020; Ding et al., 2023).

2.3. Accuracy degradation in foundation models

To evaluate the accuracy of the tropospheric delays estimated using the above-described methods, we employ three metrics: mean bias (BIAS), mean absolute error (MAE), and root-mean-square error (RMSE) (see Eq. S4). RMSE is more sensitive to outliers than MAE. A much larger RMSE compared to MAE indicates the presence of more extreme values, whereas similar values for both metrics suggest a relatively smooth

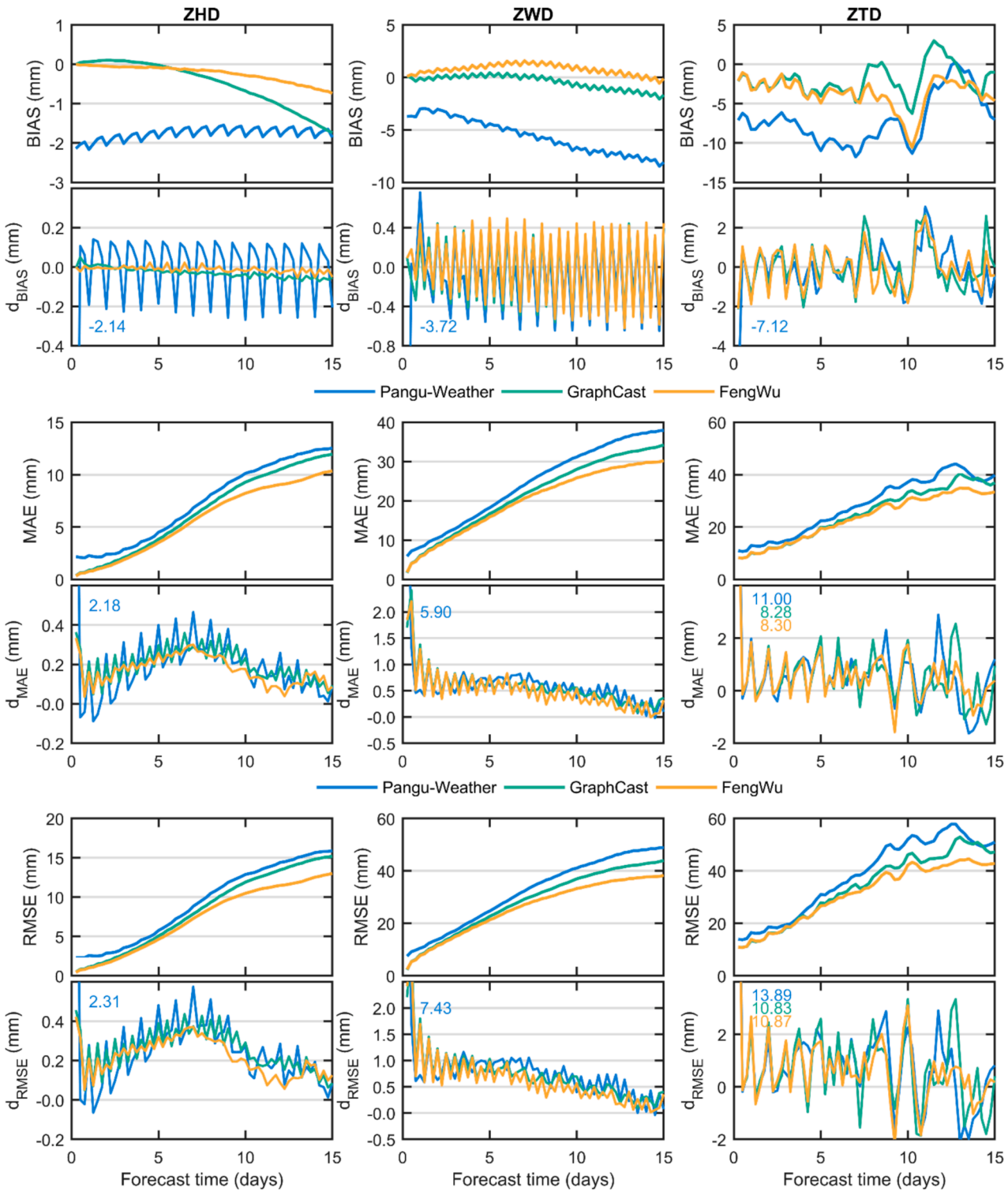


Fig. 2. Variation of BIAS, MAE and RMSE and their inter-step differences with increasing forecast time for ZHD (upper panels), ZWD (middle panels), and ZTD (lower panels). The values of texted in the figure are the results that exceed the y-axis display range.

result. Note that in the study of AI foundation models, more attention is paid to the performance of the algorithms used in the models themselves, whereas the measurement field is more concerned with how well the predictions themselves fit the reference values, and thus the evaluation metrics are calculated slightly differently, as detailed in section "Evaluation metrics" of supporting information. In addition, we calculated the inter-step differences of these three-evaluation metrics, i.e., the single differences of the metrics between neighboring forecast steps (For example $RMSE_{step\ n+1}$ minus $RMSE_{step\ n}$), which were used to assess the degradation of the accuracy of the model over multi-step inference. The three inter-step differences are represented by d_{BIAS}, d_{MAE} and d_{RMSE}, respectively. The reason we added these three metrics is that, in the "rollout" inference mode of foundation models, the input data does not need to include time labels, and the AI model itself is not aware of which step the current prediction is at during its operation. Therefore, the variation in step-to-step difference is crucial information. Furthermore,

Table 1
The step-by-step differences in the average bias (BIAS), mean absolute error (MAE), and root mean square error (RMSE) of ZHD, ZWD, and ZTD calculated using the foundation models Pangu-Weather, GraphCast, and FengWu, when predicting first step (χ_1), and averaged over multiple steps ($\chi_{2..n}$, step 2 to step n, n= 60).

Foundation models		d_{BIAS} (mm)		d_{MAE} (mm)		d_{RMSE} (mm)	
		χ_1	$\overline{\chi_{2..n}}$	χ_1	$\overline{\chi_{2..n}}$	χ_1	$\overline{\chi_{2..n}}$
ZHD	Pangu-Weather	-2.141	0.005	2.18	0.18	2.31	0.23
	GraphCast	-0.006	-0.029	0.36	0.20	0.46	0.25
	FengWu	-0.001	-0.013	0.33	0.17	0.42	0.21
ZWD	Pangu-Weather	-3.725	-0.074	5.90	0.54	7.43	0.70
	GraphCast	0.096	-0.030	1.70	0.55	2.21	0.71
	FengWu	0.089	-0.004	1.84	0.48	2.41	0.61
ZTD	Pangu-Weather	-7.120	0.001	11.00	0.489	13.89	0.63
	GraphCast	-2.155	0.019	8.28	0.486	10.83	0.62
	FengWu	-2.115	-0.043	8.30	0.426	10.87	0.54

*The worst accuracy is marked in red, and the highest accuracy is marked in blue.

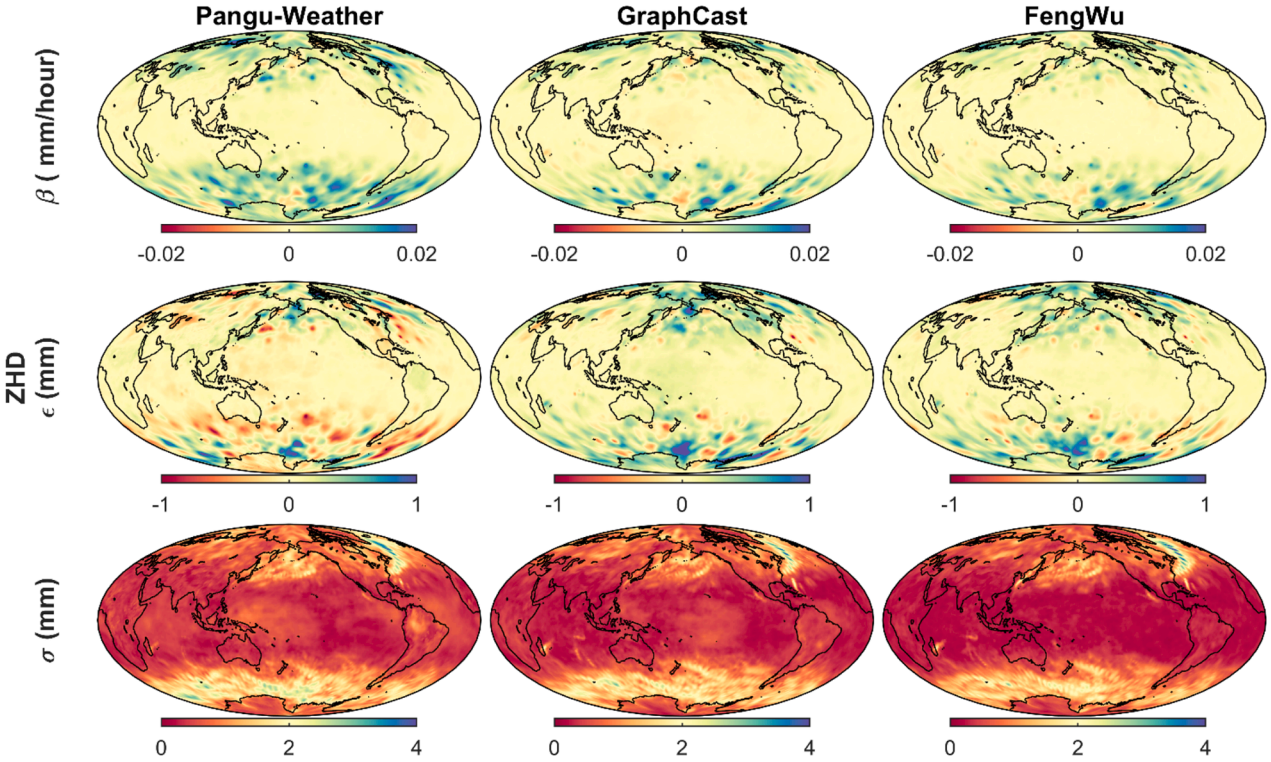


Fig. 3. Global distribution of linear fit coefficients (slope β and intercept ε) and root mean square error of fitted residuals σ for d_{RMSE} of ZHD phase 1, using ERA5 results as a reference.

this single difference method can somewhat mitigate the systematic error caused by Pangu-Weather's 13 layers of pressure level, making it more equitable in the follow-up comparisons with Pangu-Weather.

Fig. 2 shows the variation of BIAS, MAE, RMSE, d_{BIAS}, d_{MAE} and d_{RMSE} with increasing forecast time for ZHD, ZWD and ZTD. The ZHD and ZWD in the figure were carried out with the ERA5 as a reference on global 1°×1° grid points. Additionally, ZTD was carried out on 1142 GNSS stations worldwide, with the GNSS tropospheric delay product from NGL as a reference. The distribution of these GNSS stations is depicted in Fig. 9 and they were chosen from over 10,000 sites since only

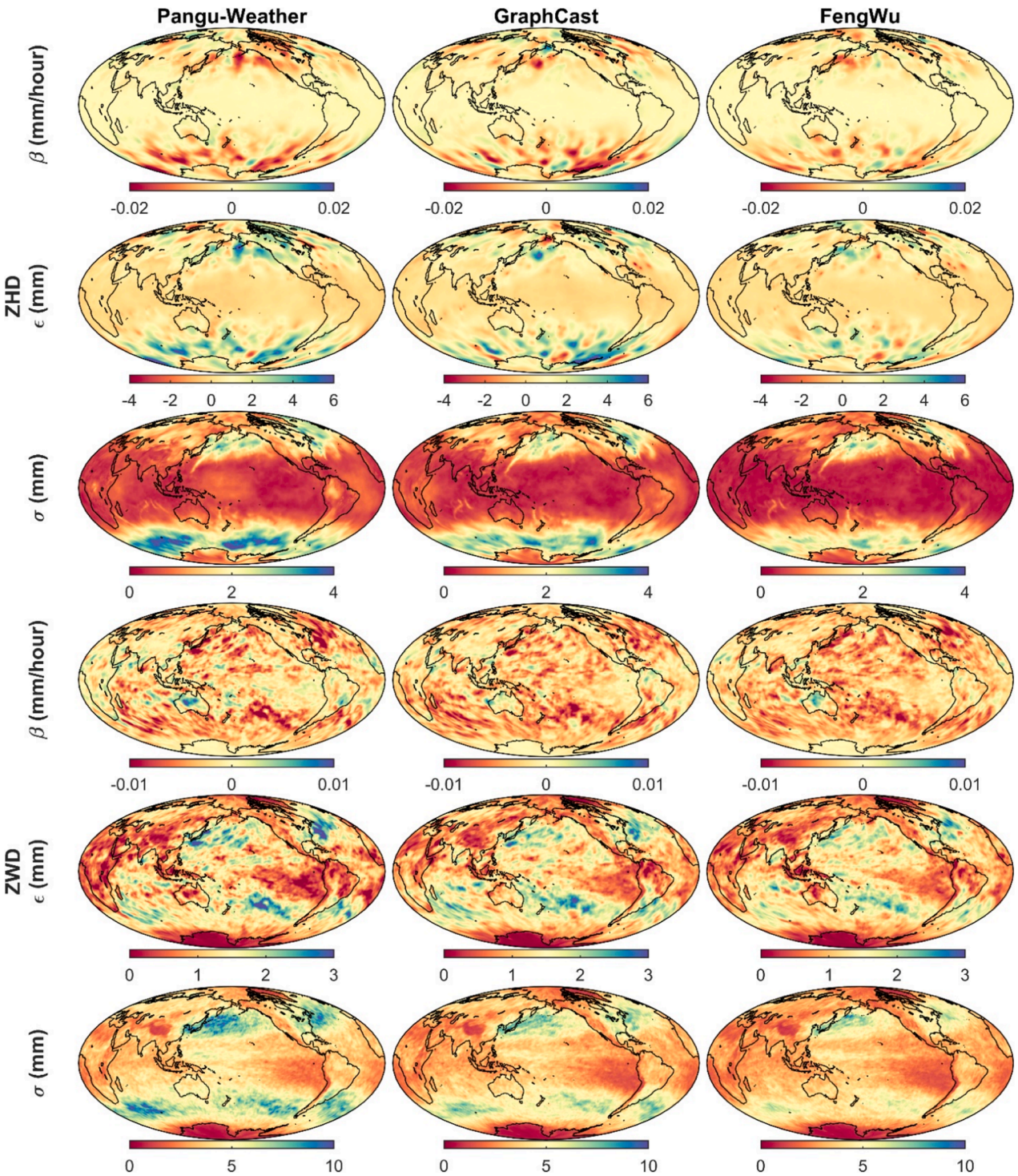


Fig. 4. Global distribution of linear fit coefficients (slope β and intercept ε) and root mean square error of fitted residuals σ for d_{RMSE} of ZHD phase 2 and ZWD, using ERA5 results as a reference.

their data completeness in 2022 satisfies the requirements of our experiments. Furthermore, the results mentioned above for ZHD and ZWD of the global distribution are shown in Figs. S7–S18 of the supporting information. Note that all global statistical results in Fig. 2 are latitude-weighted (Eq. S4).

From the BIAS in Fig. 2, the BIAS decreases as the forecast time increases, primarily because the BIAS becomes increasingly negative. This means that in multi-step forecasts, AI's results are underestimated, and this underestimation accelerates as the number of forecast steps increases. From the results of d_{BIAS}, the first notable finding is the significant daily cycle, especially Pangu-Weather. There are two reasons for this. Firstly, the accuracy of ERA5, which is used as a reference value, is different in different epochs. Secondly, Pangu-Weather utilizes 4 models with different step sizes and uses a greedy algorithm to minimize error accumulation. This research uses 6 h and 24 h. Obviously, the errors of these two step sizes are different, which leads to this periodic phenomenon. We counted the results according to the 00z, 06z, 12z and 18z epochs respectively (Figs. S4–S6 in the supporting information). It

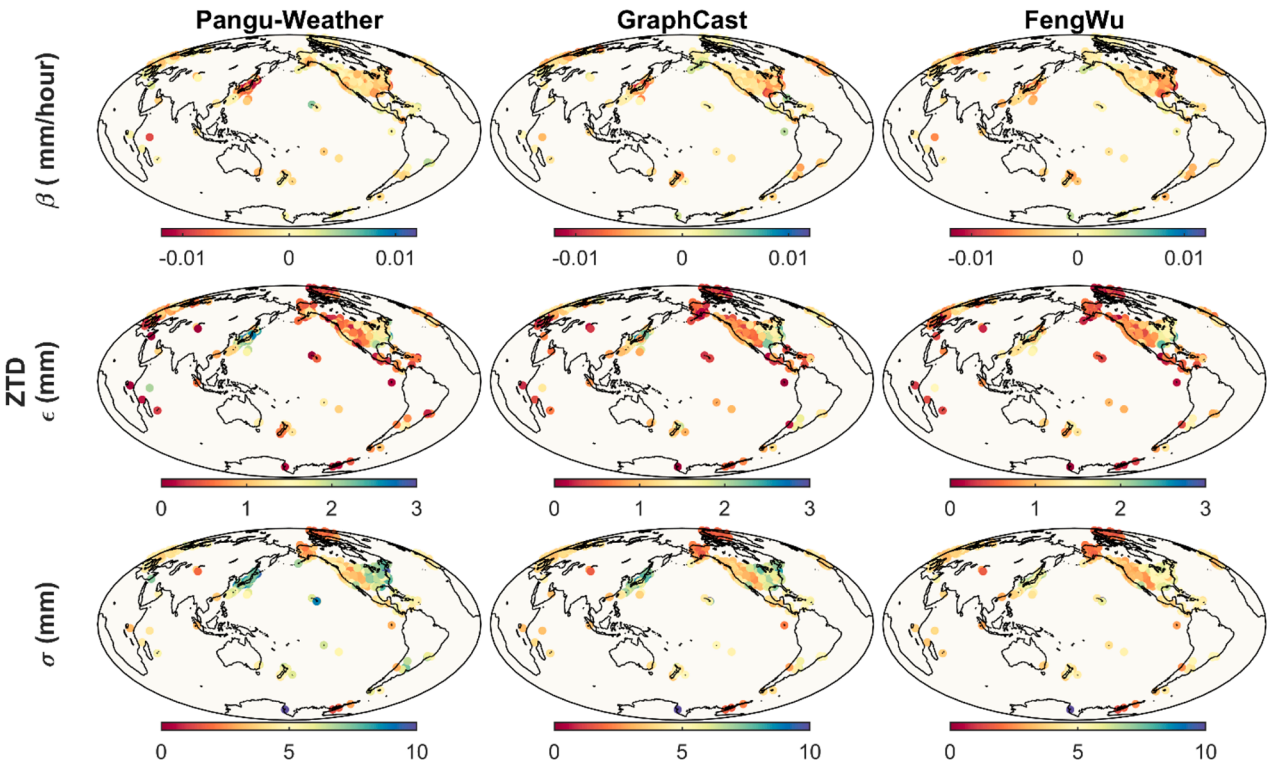


Fig. 5. Global distribution of linear fit coefficients (slope β and intercept ε) and standard deviation of fitted residuals σ for d_{RMSE} of ZTD, using GNSS ZTD results as a reference.

should be noted that the error accumulation mentioned in this research is the sum of the inter-step differences.

From the MAE and RMSE in Fig. 2, it can be observed that their variations are highly similar, and the MAE result is smaller than the RMSE result. This can explain the following two. First, the accuracy of the AI model results exhibits a small numerical range, that is, there are no or very few particularly large outliers (i.e. the data structure is relatively smooth). Second, the spatial heterogeneity of the accuracy of the AI model results does not increase significantly with the increase of the forecast length. Fig. S7 to S18 of supporting information give the global results at each forecast time, and it can be noticed that the spatial differences between the different steps are not significant after more than 1 day of forecast time, except for the value span. In addition, only Pangu-Weather showed significant terrain-related performance, with higher accuracy in high-altitude regions than in low-altitude regions. We believe this is mainly due to Pangu-Weather incorporating three constant masks, namely the topography mask, land-sea mask, and soil-type mask, during training. In particular, the terrain data led to the terrain-height-related accuracy of Pangu-Weather. However, it is also affected by the model architecture, as this terrain dependence becomes less significant as the forecast time increases. It is worth noting that this terrain-related performance is only prominently reflected in the short-term forecast accuracy of ZHD.

From the d_{MAE} and d_{RMSE} in Fig. 2, ZWD and ZTD exhibit a descending trend, while ZHD displays an ascending and then falling trend that peaks on day 7. Oddly enough, however, we inspecting the ZHD data at each single grid point and could not find the previously reported phenomenon. This suggests that the amplitude is very modest, hidden by epoch-to-epoch variations, and only visible in statistical results like Fig. 2. In summary, the phenomenon that the degree of deviation of BIAS from the reference value is increasing, while the increase in RMSE is slowing down, can explain the fact that as the forecast time increases, the overall results of the AI model deviate more from the actual results, but its spatial inhomogeneity is decreasing (i.e., the structure is getting smoother). We believe that the reason for the segmented degradation of this accuracy is that this type of AI model is typically trained to optimize a weighted mean squared/absolute error (L2/L1) norm of forecast errors. By generating forecasts closer to the average of the forecast, imposes constraints on the model inference process. This constraint was not significant within the first 7 days due to the relatively small forecast bias. However, as the forecast bias increases, this constraint begins to become prominent.

Another interesting phenomenon is that the step-by-step (inter-step) differences of the three metrics in the first step are considerably larger than those in other steps. This is due to that the result at step 0 is 0, and the step-by-step difference in the first step is corresponding metric itself (a value minus 0 equals the value itself). We computed this value and the average value of the remaining steps 2 to 60, and the results are shown in Table 1. From the results in the table, we can see that Pangu-Weather consistently exhibits the worst results among all metrics. This is because Pangu-Weather only supports 13 layers of pressure level at most, and the sparse data leads to an inherent negative bias (see Fig. 2). The model with the best forecast results in the first step is FengWu most of the time. But in the d_{MAE} and d_{RMSE} results of ZWD and ZTD, GraphCast outperforms FengWu. In the statistics of multi-step prediction results, FengWu has an absolute advantage. Except for the d_{BIAS} of ZHD and ZTD, the other metrics are the best among the three models. While in the multi-step forecast results, GraphCast lags behind Pangu-Weather (which contains only 13 layers) in ZHD and ZWD, while in ZTD, Pangu-Weather again becomes the worst performer. In summary, when predicting the first step, GraphCast is outperforms FengWu and outperforms Pangu-Weather on the variables containing water vapor (i.e.

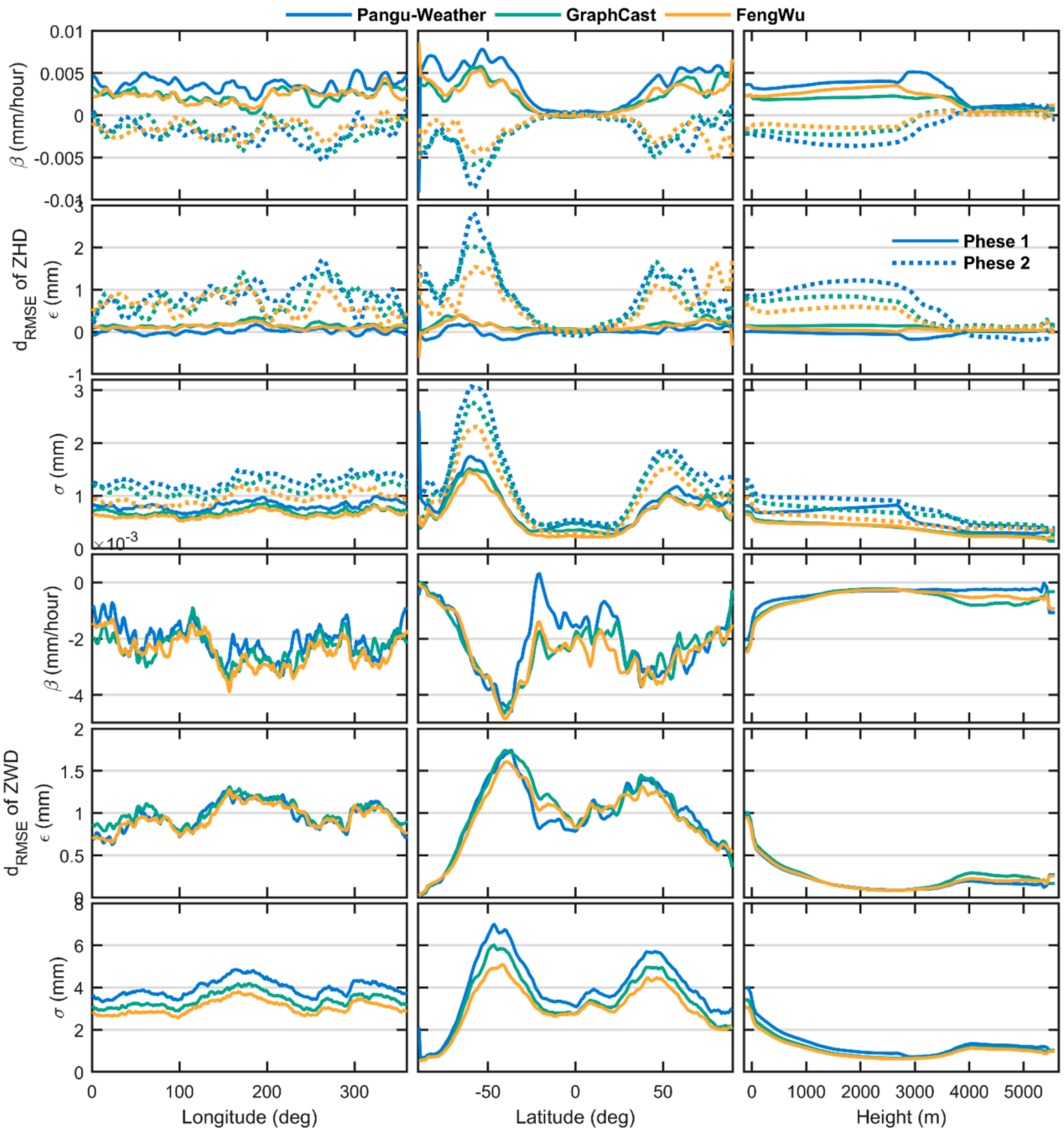


Fig. 6. Linear fitting parameters of the d_{RMSE} for ZHD (upper panels) and ZWD (lower panels), i.e. slope β, intercept ε and standard deviation (σ) of the fitted residuals, with longitude (left panels), latitude (middle panels) and height (right panels). Note that since the ZHD is fitted in segments, where days 0–7 are the first and 7–15 are the second.

ZWD and ZTD), while when predicting multiple steps, FengWu outperforms Pangu-Weather and outperforms GraphCast. We believe that there are two possible causes for this phenomenon: first, the total precipitation (TP), one of distinctive model training variables of GraphCast, is extremely discontinuous in time and space, particularly in time, it helps predict water vapor accurately in the short term but has little effect on medium- or long-term forecasting; second, FengWu's algorithm is significantly optimal in terms of accumulation of errors, and Pangu-Weather is also superior to GraphCast because it considers the effect of topography. This suggests to some extent that Transformer may surpass graph neural networks (GNN) in terms of weather forecasting algorithms.

We fit the d_{RMSE} linearly using the following equation (3) in light of the aforementioned phenomenon. The d_{RMSE} of ZHD rises and then falls, therefore we fit it in two parts: the first lasts for 0–7 days and the second, for 7–15 days. In the next section, we analyze and discuss the spatial distribution of the fitting results.

$$d_{\mathrm{RMSE}} = \beta \cdot (6 \times N_{steps}) + \varepsilon \tag{3}$$

where N_{steps} is the steps of inference, β and ε refer to the slope and intercept of the linear fit, respectively. In addition, we use σ to refer to the standard deviation of the fitting residuals.

3. Results and discussion

In this section, we conduct a linear fit to the step difference of the RMSE of the forecast results of the tropospheric delay and examine the temporal and spatial distribution characteristics of the linear fit parameters. In addition, the frequency of occurrence of the best performing model at different locations around the globe at 1–60 forecast steps are counted and the spatial distribution of this result is analyzed.

3.1. Spatiotemporal inhomogeneity of fitting parameters

Figs. 3 to 6 display the global distribution of the linear fitting parameters, namely slope β, intercept ε and standard deviation of the fitted residuals σ. Specifically, Fig. 3 presents the results for the first segment of the ZHD (increasing with forecast time), Fig. 4 shows the results for the second segment of the ZHD and the ZWD (decreasing with forecast time), and Fig. 5 shows the results for the ZTD.

The slopes in Fig. 3 show that all three models are positive in the vast

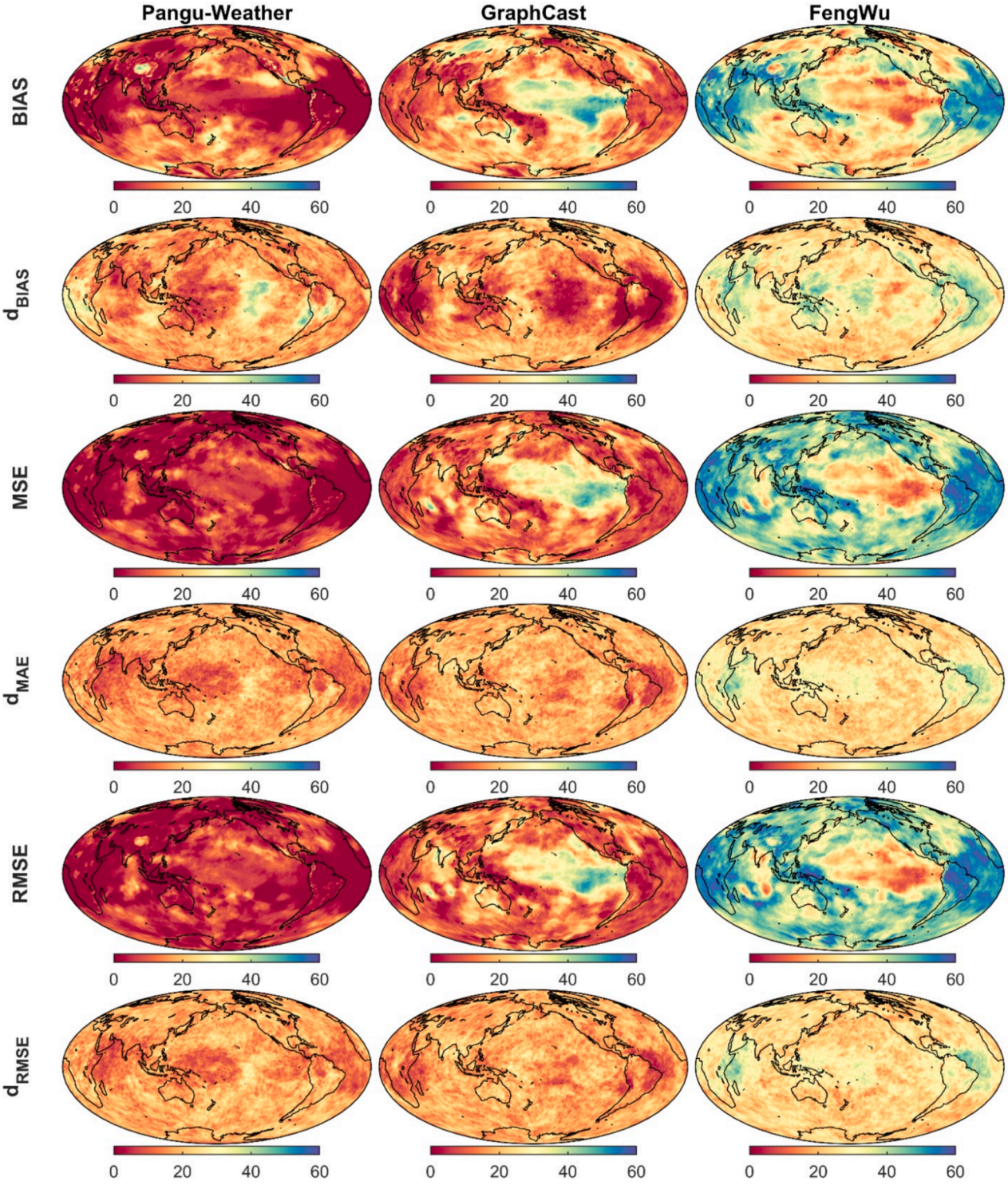


Fig. 7. The total number of times when the three foundation models, Pangu-Weather, GraphCast and FengWu performed best in 60 steps (times) of prediction using BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZHD as metrics, respectively (the bluer the better).

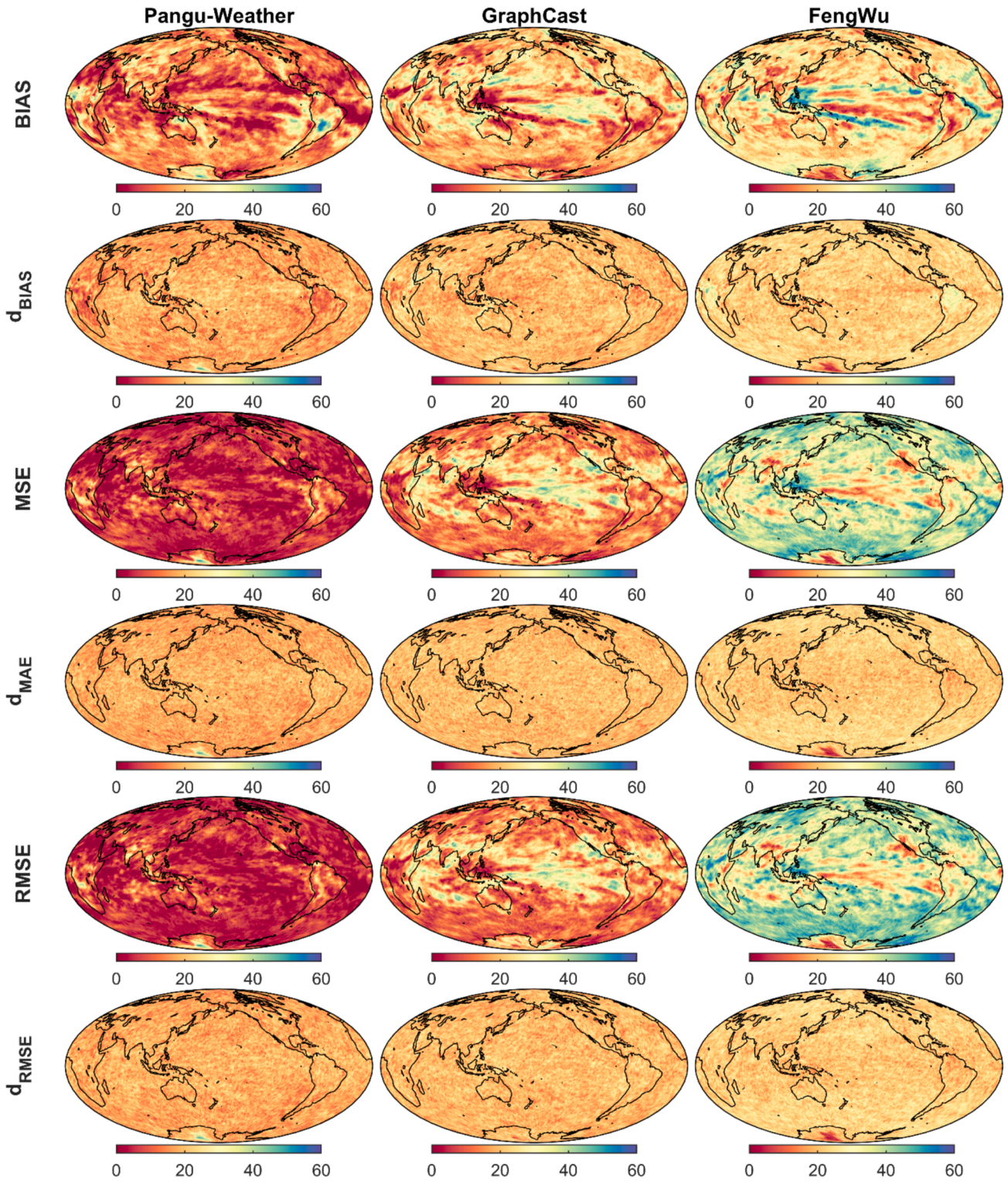


Fig. 8. The total number of times when the three foundation models, Pangu-Weather, GraphCast and FengWu performed best in 60 steps (times) of prediction using BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZWD as metrics, respectively (the buler the better).

majority of regions, and have small values at low latitudes and large values at mid- and high-latitudes. This feature is also shown in the intercepts; however, Pangu-Weather has significantly more regions with negative intercepts than the other two models. In terms of the standard deviation of the fitted residuals, there are regions with large values at high latitudes near Antarctica and in the northeastern part of North America. There are no significant differences in other regions of the globe. The above phenomenon indicates that ZHD forecasting is more difficult in mid- and high-latitude regions than in low-latitude regions.

Fig. 4 can be divided into two parts, ZHD and ZWD. From the results of ZHD, the regional distribution characteristics are similar to the differences of Fig. 3, but the positives and negatives of the values are opposite to those of Fig. 3. In addition, the magnitude of the values is slightly larger than the results of Fig. 3, which is due to the fact that the

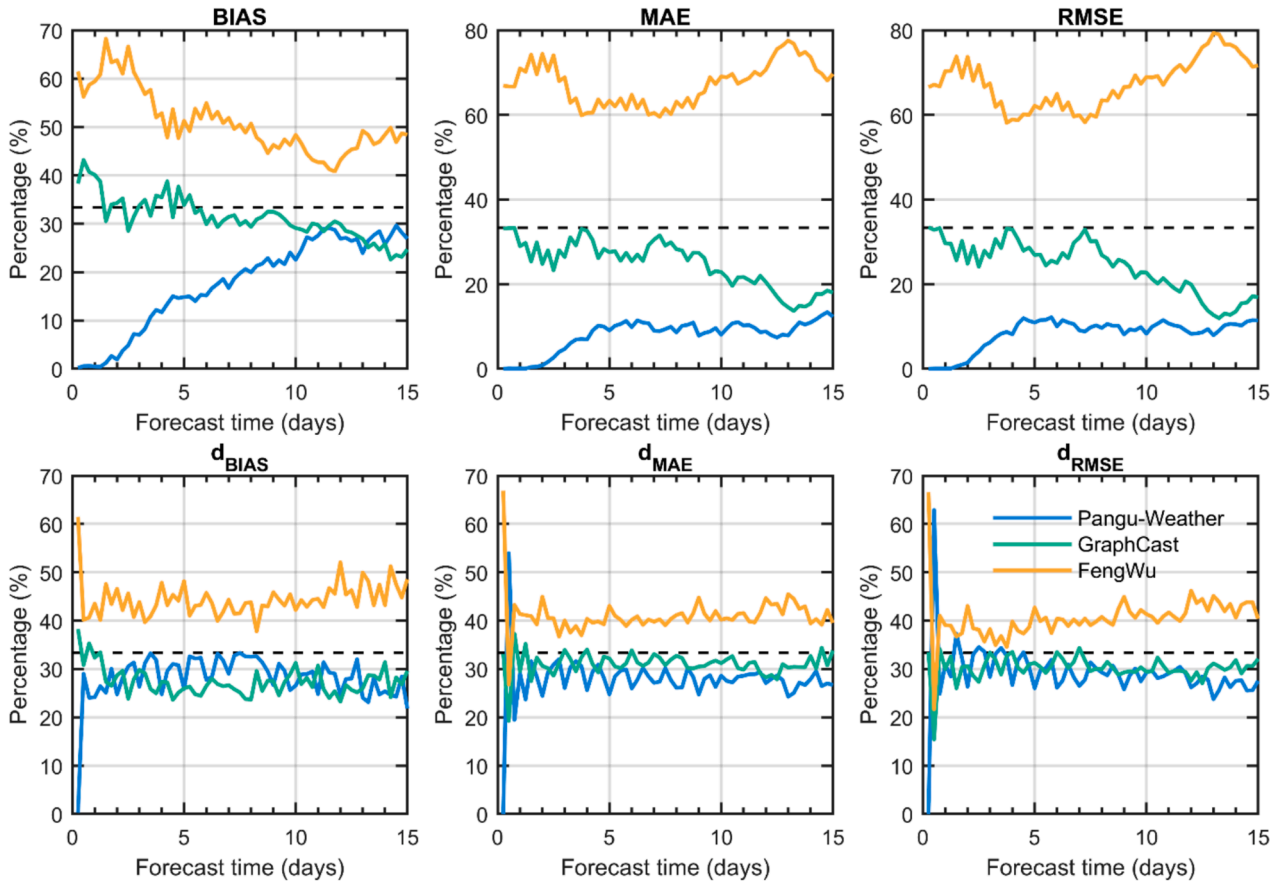


Fig. 9. Variation in the percentage of grid points with forecast time for the best performing model for BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZHD between Pangu-Weather, GraphCast and FengWu (the higher the better).

forecast time is further away from the initial epoch than the former. In addition, the case of σ starts to show larger values in the northern part of the Pacific Ocean, in addition to the northeastern waters of North America. The results for the slope of the ZWD do not show a significant latitude/longitude correlation, but it is clear that a large area in Antarctica is very close to a value of 0. The intercepts and standard deviations of the fitted residuals of the ZWD show significantly higher values in the mid-latitude ocean than in the other regions, while in Antarctica, Greenland and the Tibetan Plateau both are close to 0. Furthermore, the intercepts of the ZWD show positive values, while the results of the ZHD are both positive and negative.

Fig. 5 shows the results on 1142 GNSS sites around the world, with RMSEs calculated using the measured GNSS ZTD results as a reference. As can be seen from the figure, the slopes are mostly negative, but the GraphCast and FengWu results have small positive values in Alaska in the US and in the northern region of Canada. The intercept results show that the stations with large values for all three models are mainly located in Japan and the southeastern United States. This distribution is also reflected in the standard deviation of the fitted residuals. The results closely mirror the ZWD patterns observed in Fig. 4, demonstrating that GNSS ZTD forecast errors primarily stem from ZWD variations, which is dominated by the water vapor. The initial forecast errors in these regions are larger, and the forecast instability is also greater. This is because these regions are areas where cold and warm air masses frequently converge, with more precipitation weather and frequent water vapor activities. Pangu-Weather shows slightly worse performance than GraphCast and FengWu since it only uses data from 13 pressure levels, but the difference is not significant. This is because most of the missing pressure levels are above 50 hPa, where the water vapor is already extremely thin.

3.2. Spatiotemporal inhomogeneity at different latitudes and altitudes

Fig. 6 presents the linear fitting parameters of the d_{RMSE} for ZHD and ZWD, namely slope β, intercept ε and standard deviation of the fitted residuals σ, as functions of longitude, latitude and elevation. Since ZHD performs a segmented fit, in the figures, the first segment (0–7 days) is shown as a solid line and the second segment (7–15 days) is shown as a dashed line.

From the slopes of the ZHD results, it can be observed that the first and second segments are symmetrically distributed on the zero-value axis. The results of both segments are not significantly correlated with longitude, while they are stronger correlation with latitude, being almost zero in the low latitude region within ±20 degrees, and increasing significantly in the middle and high latitudes. In addition, the results begin to decrease rapidly in regions with heights greater than 3000 m, and are essentially above and below the value of 0 up to heights greater than 4000 m. The intercepts are similar to the slopes, but the slopes are significantly larger in the second segment than in the first because the fit starts on the seventh day and the slopes are less than zero. The standard deviation of the fitted residuals can reflect how good the fit is, and it is clear from the results that 0–7 days is better than 7–15 days. In addition, Pangu-Weather > GraphCast > FengWu, which indicates that FengWu has the smoothest error accumulation in multi-step prediction, while GraphCast is second and Pangu-Weather is the least smoothed.

From the ZWD results, it is clear that the three parameters are not significantly correlated with longitude. While slope is positively correlated with elevation, intercept and standard deviation are negatively correlated with elevation. This is because the higher the altitude, the less water vapor there is, and the thinner the water vapor is, the smaller the

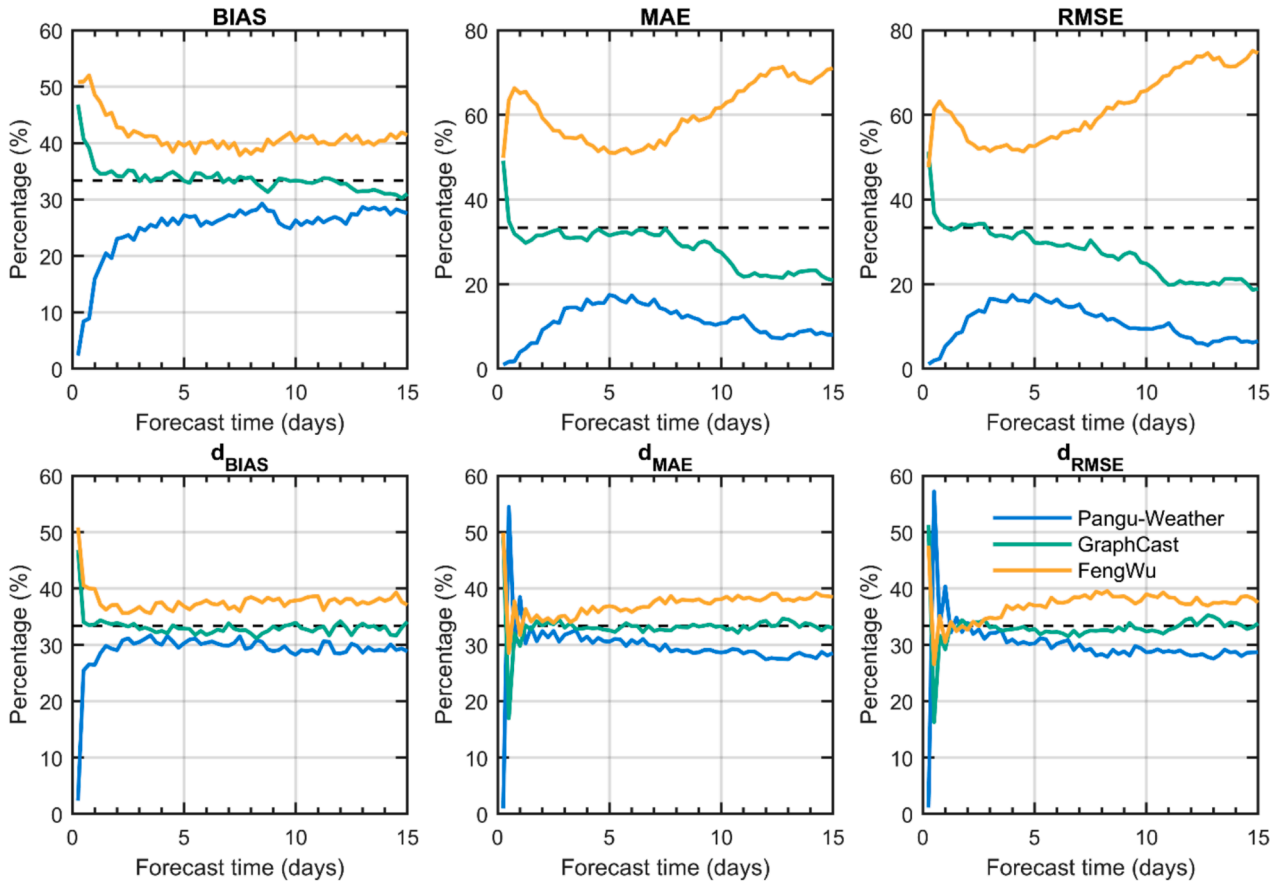


Fig. 10. Variation in the percentage of grid points with forecast time for the best performing model for BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZWD between Pangu-Weather, GraphCast and FengWu (the higher the better).

variation is and the easier it is to predict. More interestingly is the variation of the three parameters with latitude, the intercept and standard deviation have a peak at around 40 degrees north and south latitude. And in these two peak regions, it is also the time when the slope is the smallest. The above phenomenon can somewhat illustrate that the water vapor changes in the low altitude region of mid-latitude are the most difficult to predict.

3.3. Best performing models in different regions and different forecast times

Fig. 2 illustrates that, on average, FengWu outperforms GraphCast and Pangu-Weather. However, it is important to recognize that their performance varies across locations and forecast times. Consequently, we have identified the top-performing models for each specific grid point at various forecast times. Subsequently, we conducted a statistical analysis of the tagging data, and the outcomes are depicted in Figs. 7 and 8, which correspond to the ZHD and ZWD results, respectively. For a forecast horizon of 60 steps across all three models, the figures utilize color coding to represent the frequency (ranging from 0 to 60 times) with which the highest performance was achieved at each grid point.

From the ZHD results (Fig. 7), it can be seen that for BIAS, MAE, and RMSE, FengWu is significantly superior to the other two models in terms of the number of times it performs optimally in most regions. The regions where its performance is slightly inferior are the low latitude regions of the Pacific Ocean, India and the southern part of India Ocean, and the Tibetan Plateau region of China. The first two are dominated by GraphCast, while for the Tibetan Plateau region, Pangu-Weather is the best performer. In addition, it can be observed that starting from the 13th day of the forecast, Pangu-Weather starts to gain the upper hand in the two regions mentioned above except for the Tibetan Plateau (see global results of all steps in Figs. S19–S30). For d_{BIAS}, d_{MAE}, and d_{RMSE}, the best model among the three models is still FengWu, which outperforms GraphCast and Pangu-Weather, but the differences among regions are much smaller, and FengWu outperforms the other two models at low latitudes, especially in the regions of South America and Africa, and also in the eastern region of Indonesia.

Fig. 8 shows the results for ZWD in terms of BIAS, MAE, and RMSE. FengWu also outperforms the other two models by a large margin in most regions. However, at low latitudes, FengWu features a large number of east–west striped areas and performs less well than GraphCast, and in the center of Antarctica, Pangu-Weather, which has hitherto been the worst performer, performs much better than the other two models. GraphCast outperforms FengWu in the middle and low latitude regions at 1 step of prediction, while GraphCast rapidly loses its accuracy advantage in the middle and low latitudes as the number of prediction steps increases (Figs. S19–S30). We attribute this to the fact that the atmospheric variables trained by GraphCast contain TP, which contributes to water vapor prediction, but the contribution of this variable decreases with increasing prediction steps until it becomes negligible. For d_{BIAS}, d_{MAE} and d_{RMSE}, there are no significant regional differences between the three models. The only significant difference is that Pangu-Weather performs best in the Antarctica region (significantly outperforming the other regions), while FengWu performs worst

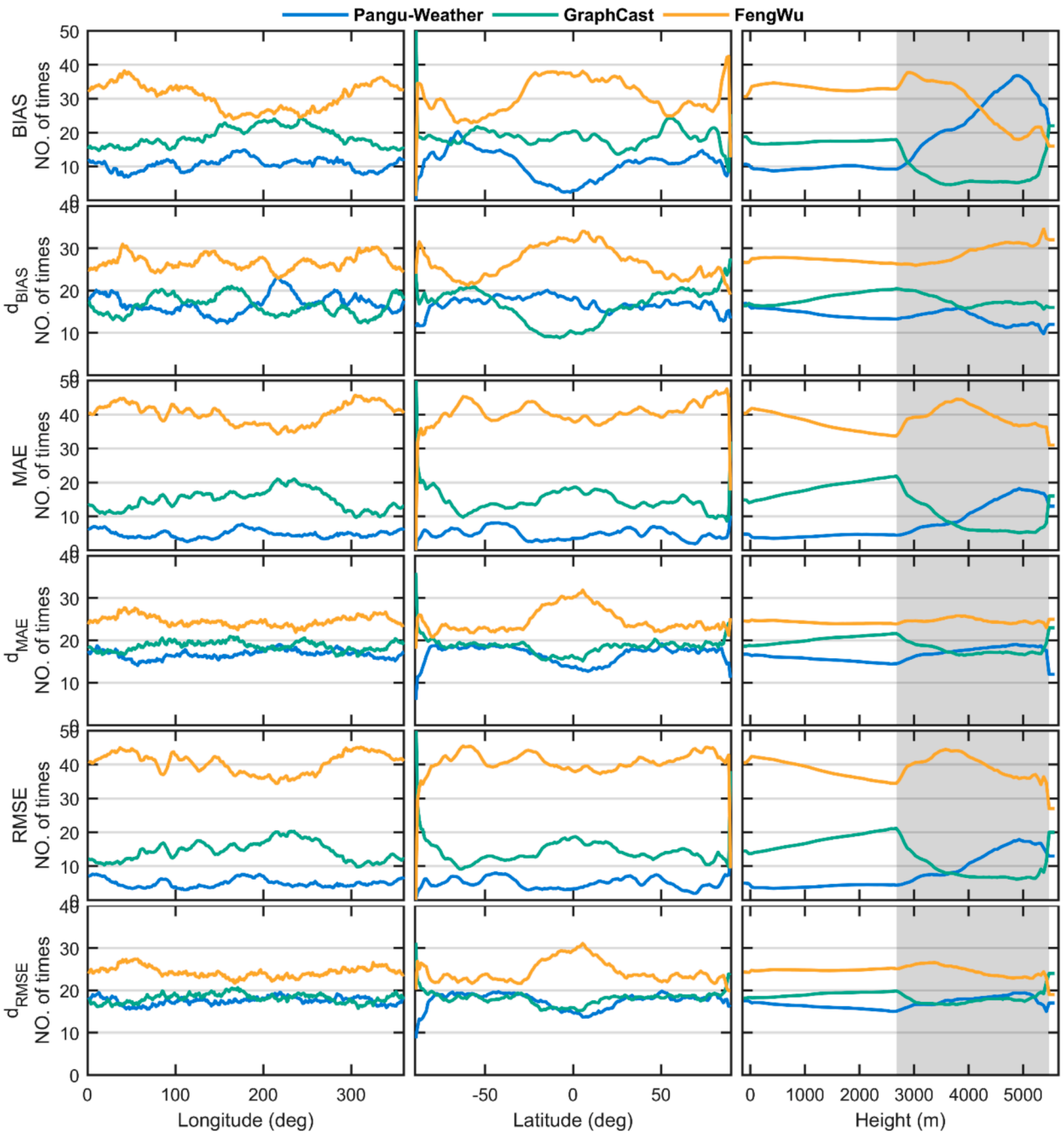


Fig. 11. The total number of times when the three foundation models, Pangu-Weather, GraphCast and FengWu performed best in 60 steps (times) of prediction using BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZHD as references, at different longitude, latitude and height (the higher the better).

(significantly underperforming the other regions).

Fig. 9 and Fig. 10 illustrate how the number of grid points occupied by the best-performing models as a percentage of all grid points varies with forecast time. From the BIAS, MAE, and RMSE of ZHD, Pangu-Weather performs best in almost no region in the 0–2 day forecast. As the forecast time increases, Pangu-Weather begins to gradually be able to prevail in about 10 % of the regions. In the other two models, FengWu consistently outperforms GraphCast, where the number of regions where GraphCast performs best is decreasing as the forecast length increases. In terms of d_{BIAS}, d_{MAE}, and d_{RMSE}, except for the first 2 steps, the proportion of the three models has been stable for the forecast time afterwards, with FengWu at ~ 40 % and GraphCast and Pangu-Weather at ~ 30 %.

From the ZWD results of Fig. 10, the BIAS results start to stabilize after 4 days of forecasting, while the results from d_{BIAS}, d_{MAE} and d_{RMSE} start to stabilize after 2 days of forecasting. After stabilization, FengWu is larger than GraphCast and larger than Pangu-Weather, whereas MAE and RMSE show that Pangu-Weather increases and then decreases, FengWu decreases and then continues to increase, and GraphCast keeps decreasing. In addition, we have statistically calculated the results of the best-performing models with longitude, latitude and height in Fig. 11 and Fig. 12. From the performance of ZHD (Fig. 11), all evaluation metrics exhibit no significant correlation with longitude. FengWu is significantly superior to the other two models in the low-latitude region, with the exception of MAE and RMSE. This implies that in multi-step inference, FengWu not only demonstrates overall superiority over the

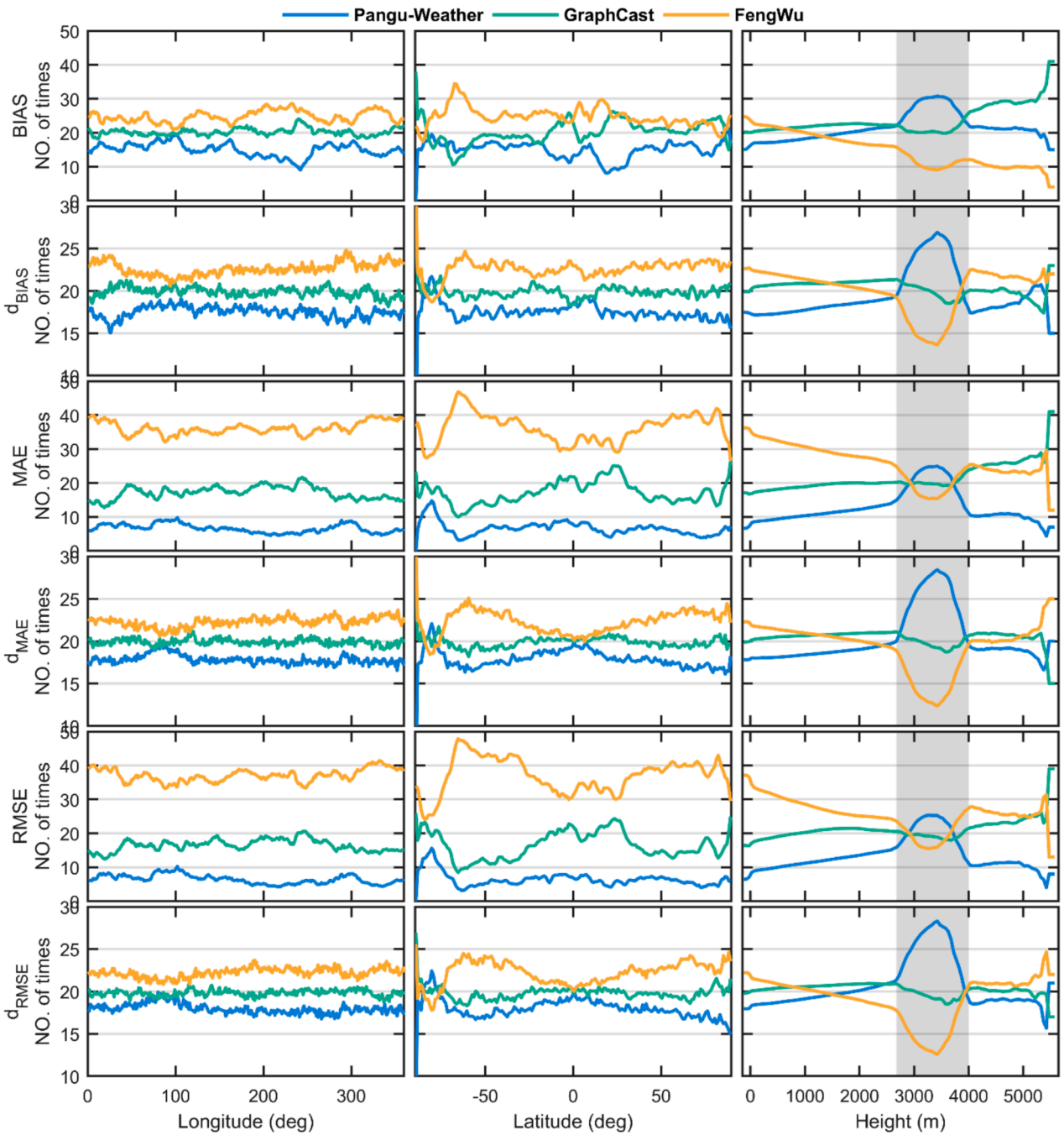


Fig. 12. The total number of times when the three foundation models, Pangu-Weather, GraphCast and FengWu performed best in 60 steps (times) of prediction using BIAS, d_{BIAS}, MAE, d_{MAE}, RMSE, and d_{RMSE} of ZWD as references, at different longitude, latitude and height (the higher the better).

other two but also stands out particularly in the low-latitude region. Regarding the variation with height, a notable inflection point emerges at 2,800 m. Beyond this inflection point, the number of steps where Pangu-Weather and FengWu predominate in most metrics begins to increase, while that of GraphCast continues to decline. This phenomenon is primarily due to the results in Antarctica, and it also indicates that Pangu-Weather performs more effectively in Antarctica.

From the ZWD results (Fig. 12), the performance of the three models is also not significantly correlated with longitude, and in the variation with latitude, FengWu shows a slight decrease in the low latitude region, while Pangu-Weather shows a slight increase. What is more interesting is the variation with height, Pangu-Weather exhibits an upward bump in the 2,800 m to 4,000 m height region, while FengWu exhibits a downward depression. This is again caused by the location of the Antarctic center region (Figs. S19–S30). It is clear that Pangu-Weather has significantly better water vapor forecasts in Antarctica than FengWu. To further establish this, we present the results for the Antarctica region exclusively in Fig. 13. From the results, it is clear that the areas with outstanding values correspond very well with the high-altitude regions, and it is in these regions that Pangu-Weather outperforms the other two models for the majority of the forecast time.

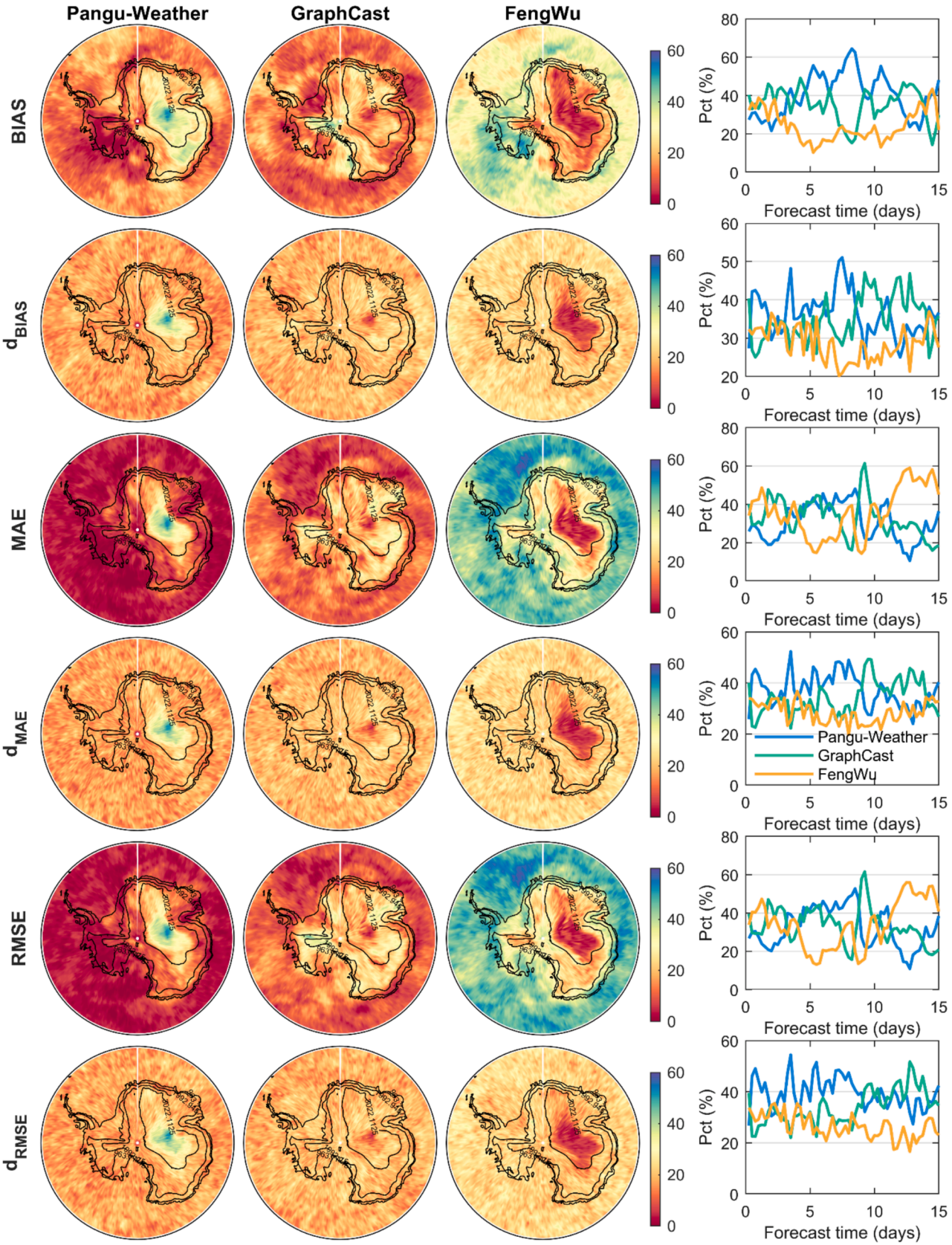


Fig. 13. Statistical results of the best performing model for ZWD in Antarctica. The curves in the left panels are topographic contours, and the right panels are variation in the percentage with forecast time of grid points with height greater than 2,800 m.

4. Conclusions and perspectives

In this contribution, we analyze the accuracy of tropospheric delay forecasts derived from AI weather forecasting foundation models represented by Pangu-Weather, GraphCast, and FengWu from the perspective of GNSS, as well as the rate of accuracy variation during multi-step predictions. We investigate the spatiotemporal inhomogeneity of AI model accuracy degradation and discuss the differences between various models and the potential causes. Drawing from this, we provide actionable insights and recommendations targeted at GNSS users and developers of AI foundational models alike. The experimental results can be summarized as follows:

1) The deterioration of the AI model accuracy with increasing forecast time is mainly manifested in the gradually increasing negative bias, i.e., the forecasts are underestimated, and the degree of underestimation increases with the number of forecast steps. The accuracy of the model shows accelerated deterioration on days 0–7 and decelerated deterioration on days 7–15 on ZHD, while it consistently shows decelerated deterioration on ZWD and ZTD.
2) The atmospheric variable total precipitation in the AI model contributes to accurate short-term forecasts of water vapor. Due to the addition of total precipitation, particularly in middle and low latitude regions, GraphCast outperforms both FengWu and Pangu-Weather in first-step inference. But FengWu outperforms the other two in the multi-step inference, suggesting that FengWu has an algorithmic advantage over the other two.
3) Taking into account the variations in terrain may enhance the accuracy of forecasts in high-altitude areas. In the high-altitude zone, Pangu-Weather is beginning to close the gap and surpass the other two models, suggesting that it is more proficient at predicting regions with complicated topography. The observed performance advantage of Pangu-Weather in high-altitude regions (e.g., Antarctica) may partially result from its 3DEST algorithm that explicitly encodes topographic information. However, it should be acknowledged that model performance differences likely stem from a combination of factors, such as data quality, model architecture, training strategy, and hyperparameter settings. Future research will explore these factors' contributions to better understand the performance variations of AI models in complex terrain regions.

Overall, the AI weather forecast foundation models have performed quite well, but they still possess significant potential for further development, and their ultimate performance is far beyond what is currently demonstrated. We anticipate that with advancements in AI technology and increased computational power, scholars will be able to incorporate more atmospheric variables, and even oceanic variables, into the training of foundation models. Additionally, terrain changes, vegetation types, geological conditions, and other factors can also be integrated into the algorithms. We hope that the research presented in this paper will make a positive contribution to advancing this process.

CRediT authorship contribution statement

Junsheng Ding: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Wu Chen:** Writing – review & editing, Funding acquisition. **Junping Chen:** Writing – review & editing, Funding acquisition. **Jungang Wang:** Writing – review & editing, Validation, Methodology. **Yize Zhang:** Writing – review & editing, Formal analysis. **Lei Bai:** Writing – review & editing, Resources. **Yuyan Wang:** Writing – review & editing. **Xiaolong Mi:** Writing – review & editing. **Tong Liu:** Writing – review & editing. **Duojie Weng:** Writing – review & editing.

Funding

This research was funded by the General Research Fund of Hong Kong (Grant No. 15229622), the Innovation and Technology Fund of Hong Kong (Grant No. ITP/019/22LP), the National Natural Science Foundation of China (No. 42474034). Jungang Wang is financially supported by the DFG COCAT (Nr. 490990195) Project.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Technische Universität Wien for the open-source ray-tracing software RADIATE (https://github.com/TUW-VieVS/RADIATE, *Hofmeister 2016*) and ecmwf-lab for open-source AI models command-line management tool ai-models (https://github.com/ecmwf-lab/ai-models).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2025.104473.

Data availability

This research processed data from more than 28 TB. The ERA5 data on pressure levels (Hersbach et al., 2018a) and ERA5 data on single levels (Hersbach et al., 2018b) are available at Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form and https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form (last accessed 8 Oct, 2023). The NGL GNSS tropospheric delay products are available at http://geodesy.unr.edu/gps_timeseries/trop/ (Blewitt and Hammond, 2018). The foundation model Pangu-Weather is available at https://github.com/198808xc/Pangu-Weather (*Bi et al., 2023*). The foundation model GraphCast is available at https://github.com/google-deepmind/graphcast (*Lam et al., 2023*). The foundation model FengWu is available at https://github.com/OpenEarthLab/FengWu (Chen et al., 2023).

References

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619 (7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3.

Blewitt, G., Hammond, W., 2018. Harnessing the GPS data explosion for interdisciplinary science. *Eos* 99. https://doi.org/10.1029/2018EO104623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al., 2021. On the opportunities and risks of foundation models. *arXiv preprint* arXiv: 2108.07258. https://doi.org/10.48550/arXiv.2108.07258.

Bonavita, M., 2024. On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.* 51 (12), e2023GL107377. https://doi.org/10.1029/2023GL107377.

Charlton-Perez, A.J., Dacre, H.F., Driscoll, S., Gray, S.L., Harvey, B., Harvey, N.J., Volonté, A., 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 (1), 93. https://doi.org/10.1038/s41612-024-00638-w.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J. J., et al., 2023a. FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead. *arXiv preprint arXiv*:2304.02948. https://doi.org/10.48550/arXiv.2304.02948.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., Li, H., 2023. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. npj Clim. Atmos. Sci. 6 (1), 190. https://www.nature.com/articles/s41612-023-00512-1.

Ding, J., Chen, J., 2020. Assessment of empirical troposphere model GPT3 based on NGL's global troposphere products. *Sensors* 20 (13), 3631. https://doi.org/10.3390/s20133631.

Ding, J., Chen, J., Wang, J., Zhang, Y., 2023. Characteristic differences in tropospheric delay between Nevada Geodetic Laboratory products and NWM ray-tracing. *GPS Solutions* 27 (1), 47. https://doi.org/10.1007/s10291-022-01385-2.

Ding, J., Mi, X., Wu, C., Chen, J., Wang, D. J., Zhang, Y., et al., 2024a. Forecasting of Tropospheric Delay Using Ai Foundation Models in Support of Microwave Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63, Article Number: 5803019. https://doi.org/10.2139/ssrn.4743983.

Ding, J., Chen, W., Chen, J., Wang, J., Zhang, Y., Weng, D., et al., 2024b. AI Foundation Models Facilitate Real-time Global GNSS Precipitable Water Vapor Retrieval with Sub-millimeter Accuracy. *ESS Open Archive*. July 12, 2024. https://doi.org/10.22541/essoar.172081357.72805131/v1.

Feldmann, M., Beucler, T., Gomez, M., Martius, O., 2024a. Lightning-Fast Thunderstorm Warnings: Predicting Severe Convective Environments with Global Neural Weather Models. *arXiv preprint arXiv:2406.09474*. https://arxiv.org/abs/2406.09474.

Feldmann, M., Poulain-Auzeau, L., Gomez, M., Beucler, T., Martius, O., 2024b. Convective environments in AI-models-What have AI-models learned about atmospheric profiles? (No. EGU24-5373). *Copernicus Meetings*. https://doi.org/10.5194/egusphere-egu24-5373.

Feng, D., Qin, Y., Feng, W., Li, W., Shang, K., Ma, H., 2024. Survey of research on confidential computing. *IET Commun.* 18 (9), 535–556. https://doi.org/10.1049/cmu2.12759.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al., 2018a. ERA5 hourly data on pressure levels from 1979 to present. *Copernicus climate change service (c3s) climate data store (cds)*, 10(10.24381). doi: 10.24381/cds.bd0915c6.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al., 2018b. ERA5 hourly data on single levels from 1979 to present. *Copernicus climate change service (c3s) climate data store (cds)*, 10(10.24381). doi: 10.24381/cds.adbb2d47.

Hsu, K., Liu, C. C., Peng, M., Chen, D. S., Chang, P. L., Hsiao, L.F., et al., 2024. Performance of Five Machine Learning-based Global Weather Prediction Models in the East Asia Region. https://www.researchsquare.com/article/rs-4250353/v1.

Hofmeister, A., 2016. Determination of path delays in the atmosphere for geodetic VLBI by means of ray-tracing. http://hdl.handle.net/20.500.12708/136.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Battaglia, P., 2023. Learning skillful medium-range global weather forecasting. *Science*. https://doi.org/10.1126/science.adi2336.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., et al., 2024. AIFS-ECMWF's data-driven forecasting system. arXiv preprint arXiv:2406.01465. https://arxiv.org/abs/2406.01465.

Lavers, D.A., Simmons, A., Vamborg, F., Rodwell, M.J., 2022. An evaluation of ERA5 precipitation for climate monitoring. Q. J. R. Meteorolog. Soc. 148 (748), 3152–3165. https://doi.org/10.1002/qj.4351.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616 (7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A., 2023. ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343. https://arxiv.org/abs/2301.10343.

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. https://doi.org/10.5194/egusphere-2024-1042.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al., 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214. https://arxiv.org/abs/2202.11214.

Puladi, B., Gsaxner, C., Kleesiek, J., Hölzle, F., Röhrig, R., Egger, J., 2023. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. *Int. J. Oral Maxillofac. Surg.* https://doi.org/10.1016/j.ijom.2023.09.005.

Shu, H., Wang, Y., Song, W., Guo, H., Song, Z., 2024. Forecasting the Future with Future Technologies: Advancements in Large Meteorological Models. *arXiv preprint arXiv: 2404.06668*. https://arxiv.org/abs/2404.06668.

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. *available at Research Square*. https://doi.org/10.21203/rs.3.rs-4645037/v1.