Journal of the Meteorological Society of Japan

EARLY ONLINE RELEASE

The DOI for this manuscript is
DOI:10.2151/jmsj.2025-031
J-STAGE Advance published date: June 26, 2025

DeepMedcast: A Deep Learning Method for Generating Intermediate Weather Forecasts among Multiple NWP Models

Atsushi KUDO[1]

Numerical Prediction Division
Japan Meteorological Agency, Tokyo, Japan

November 15, 2024

---
1) Corresponding author: Atsushi Kudo, Numerical Prediction Division, Japan Meteorological Agency, 3-6-9, Toranomon, Minato-ku, Tokyo 105-8431 JAPAN
Email: atsushi.kudou@met.kishou.go.jp
Tel: +81-3-6758-3900

Abstract

Numerical weather prediction (NWP) centers around the world operate a variety of NWP models. In addition, recent advances in AI-driven NWP models have further increased the availability of NWP outputs. While this expansion holds the potential to improve forecast accuracy, it raises a critical question: which prediction is the most plausible? If the NWP models have comparable accuracy, it is impossible to determine in advance which one is the best. Traditional approaches, such as ensemble or weighted averaging, combine multiple NWP outputs to produce a single forecast with improved accuracy. However, they often result in meteorologically unrealistic and uninterpretable outputs, such as the splitting of tropical cyclone centers or frontal boundaries into multiple distinct systems.

To address this issue, we propose DeepMedcast, a deep learning method that generates intermediate forecasts between two or more NWP outputs. Unlike averaging, DeepMedcast provides predictions in which meteorologically significant features—such as the locations of tropical cyclones, extratropical cyclones, fronts, and shear lines—approximately align with the arithmetic mean of the corresponding features predicted by the input NWP models, without distorting meteorological structures. We demonstrate the capability of DeepMedcast through case studies and verification results, showing that it produces realistic and interpretable forecasts with higher accuracy than the input NWP

models. By providing plausible intermediate forecasts, DeepMedcast can significantly contribute to the efficiency and standardization of operational forecasting tasks, including general, marine, and aviation forecasts.

Keywords deep neural network; intermediate forecast; numerical weather prediction

1. Introduction

In recent decades, numerical weather predictions (NWPs) and their post-processing have played a central role in issuing weather forecasts, warnings, and advisories (WMO 2013; Vannitsem 2021). NWP centers around the world have developed and are operating a variety of NWP models for accurate weather predictions. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) operates the Integrated Forecasting System (IFS) and its ensemble prediction system (ECMWF 2024); the UK Met Office operates the Unified Model and the Met Office Global and Regional Ensemble Prediction System (Brown et al. 2012; Hagelin et al. 2017; Inverarity et al. 2023). The National Centers for Environmental Prediction (NCEP) at the National Oceanic and Atmospheric Administration (NOAA) operates the Global Forecast System (NCEP 2016), the High-Resolution Rapid Refresh (Dowell et al. 2022), and the Hurricane Weather Research and Forecasting model (Gopalakrishnan et al. 2011). The Japan Meteorological Agency (JMA) operates three deterministic NWP models and two ensemble prediction systems for short-range to weekly forecasts: the Global Spectrum Model (GSM), the Meso-Scale Model (MSM), the Local Forecast Model, the Global Ensemble Prediction System, and the Mesoscale Ensemble Prediction System (JMA 2024). These models cover different areas with varying resolutions and processes.

In addition to traditional physics-based NWP models, recent advancements in artificial intelligence (AI) have introduced new methods for producing weather predictions. AI-driven

NWP models, such as FourCastNet (Pathak et al. 2022; Bonev et al. 2023), GraphCast (Lam et al. 2022), Pangu-Weather (Bi et al. 2022; Bi et al. 2023), FengWu (Chen et al. 2023; Han et al. 2024), Aurora (Bodnar et al. 2024), GenCast (Price et al. 2023), and AIFS (Lang et al. 2024), have demonstrated the ability to enhance both the speed and accuracy of weather predictions by leveraging deep learning techniques to model complex atmospheric systems.

At present, forecasters are able to use multiple NWP models including AI-driven NWP models, which provide a range of possible atmospheric states, allowing them to select the most plausible prediction from available NWPs. However, this raises a critical question: Which prediction is the most plausible? If the models have comparable accuracy, it is impossible to determine in advance which one is the best. One practical and widely used solution is to average the results from multiple NWP models or their post-processed outputs, as this can reduce random errors inherent in NWP models and can lead to higher accuracy than individual models (Vislocky and Fritsch 1997; JMA 2018). The National Hurricane Center and the Joint Typhoon Warning Center in the United States use consensus forecasts (e.g., Simon et al. 2018; Cangialosi et al. 2023), which are weighted averages, extensively for both tropical cyclone (TC) track and intensity predictions. JMA employs consensus forecasting for TC track predictions by averaging positions of TC centers from multiple NWP outputs to improve forecast accuracy (Nishimura and Fukuda 2019; JMA 2022). The UK Met Office operates the IMPROVER system (Roberts et al. 2023), which applies a weighted

95 average of post-processing and nowcasts based on multiple NWP outputs. The National
96 Weather Service (NWS) at NOAA operates the National Blend of Models (NBM), which
97 provides statistically post-processed multi-model ensemble guidance (Hamill et al. 2017).
98 The German Meteorological Service uses MOSMIX and ModelMIX (Primo et al. 2024),
99 which are weighted averages of post-processing based on IFS, their global model, and their
100 regional ensemble model. Additionally, the World Area Forecast Centre, comprising centers
101 in London and Washington, operates harmonized forecasts, including mean, maximum, and
102 minimum forecasts, from both NWP outputs for aviation hazards such as cumulonimbus
103 clouds, turbulence, and in-flight icing (ICAO 2016).

104 It is straightforward to average the central position of TCs, extra tropical cyclones, or the
105 location of fronts because averaging does not degrade their clarity. However, averaging
106 atmospheric fields such as pressure or wind speed around these systems is not appropriate.
107 This is because averaging can smooth out or distort these fields, weakening the central
108 pressure or wind speeds around cyclones and fronts, and even causing TCs or fronts to split
109 into two, resulting in predictions that are meteorologically unrealistic and difficult to interpret.
110 Forecasters must then choose between two options: using a single model that is realistic
111 and interpretable but potentially less accurate or using an averaged prediction that is
112 unrealistic and uninterpretable but may be more accurate.

113 Beyond these challenges, weather forecasting faces additional difficulties in operational
114 practice. In JMA's forecasting and warning issuance process, TC track forecasts, which are

115 based on consensus forecasts derived by averaging the positions of TC centers from
116 multiple NWP models, take precedence. As a result, forecasters responsible for general,
117 marine, and aviation forecasts must ensure that their forecasts align with TC track forecasts.
118 However, since no NWP model inherently conforms to the TC track forecasts, forecasters
119 need to adjust the existing NWP outputs in their minds to construct forecast scenarios that
120 align with them. This process requires significant time and effort and can pose a major
121 obstacle to the standardization of forecasting workflows, leading to inefficiencies in
122 operational forecasting.

123 In addition to these operational challenges, machine learning-based post-processing
124 presents its own set of difficulties, particularly regarding data requirements. In conventional
125 model output statistics (MOS), obtaining long-term, homogeneous datasets is particularly
126 difficult because the input NWP model is periodically updated, causing changes in its
127 systematic errors. Consequently, the statistical relationships learned from past data may no
128 longer hold after a model update.

129 The objective of this study is to propose DeepMedcast, a method that uses deep
130 learning to generate a realistic and interpretable "intermediate forecast" between two or
131 more NWP models. In this study, we do not attempt to define intermediate forecasts in a
132 physical or mathematical sense. Instead, we adopt a pragmatic definition: an intermediate
133 forecast is a predicted meteorological field in which meteorologically significant features—
134 such as the center positions of TCs or extratropical cyclones, frontal boundaries, and shear

lines—are located at the arithmetic mean of the corresponding features predicted by input NWP models, and it simultaneously exhibits higher forecast accuracy against observations than the input models.

Unlike averaging, DeepMedcast can produce atmospheric fields around cyclones and fronts without smoothing out or disturbing their distributions. This capability is crucial in operational forecasts, where accurate and interpretable predictions are needed for issuing reliable warnings and advisories—particularly when a TC is approaching—and also contributes to the standardization of forecasting workflows by reducing reliance on manual adjustments by individual forecasters. While mathematical frameworks such as displacement interpolation and barycenters in optimal transport theory (McCann 1997; Agueh and Carlier 2011; Peyré and Cuturi, 2020) provide rigorously defined intermediate states and have been widely used in machine learning fields such as image processing, applying them directly to forecasts from multiple NWP models remains challenging. Le Coz et al. (2023) proposed a barycenter-based method using optimal transport to combine forecasts from multiple NWP models in the context of subseasonal prediction, offering a mathematically principled approach to multi-model ensemble forecasts. Duc and Sawada (2024) pointed out that traditional arithmetic ensemble means tend to excessively smooth rainfall structures, and proposed a Gaussian–Hellinger barycenter based on unbalanced optimal transport theory to derive more realistic and structurally coherent ensemble means. Their method is particularly effective in representing heavy precipitation and may contribute

155 to future advances in ensemble post-processing and spatial verification techniques. To the
156 best of the authors' knowledge, however, no prior study has successfully generated two-
157 dimensional intermediate forecasts for mean sea-level pressure or surface wind vectors at
158 high temporal resolution and in areas strongly influenced by topography. In contrast,
159 DeepMedcast is not intended to generate physically or mathematically consistent
160 intermediate forecasts but to provide forecasters with operationally practical solutions,
161 addressing a critical challenge in operational forecasting. In many national meteorological
162 centers, including JMA, machine learning-based post-processing methods, referred to as
163 forecast guidance, are operationally employed. Although forecast guidance does not
164 preserve physical consistency, it enhances forecast accuracy by reducing biases inherent
165 in NWP models and significantly improves the efficiency of operational forecasting.
166 DeepMedcast is a kind of post-processing method that is not designed to ensure physical
167 consistency but to provide practical support for forecasters.

168 This paper is structured as follows: Section 2 presents the methodology and data used
169 for DeepMedcast, detailing the deep learning architecture and training process. Section 3
170 discusses the results of applying DeepMedcast to multiple NWP models with case studies
171 and verification results, and Section 4 offers a discussion of contributions to operational
172 forecasting and the key features of DeepMedcast's architecture. Finally, Section 5
173 concludes with a summary of the findings and future research directions.

174

2. Method and data

2.1 The framework of DeepMedcast

The main idea behind DeepMedcast lies in its original approach to generating intermediate forecasts between two NWP outputs. Figure 1 illustrates the framework of DeepMedcast. During the training phase, instead of using two different NWP outputs intended for creating an intermediate forecast, DeepMedcast utilizes data at two forecast lead times (FT), FT = t - Δt and FT = t + Δt, from a single NWP model as input variables for the deep neural network (DNN) (Fig. 1a). The output from the DNN is then compared to the forecast from the same NWP at the intermediate lead time (FT = t) to calculate the loss for the backpropagation process. This approach enables the network to generate intermediate forecasts while reducing the blurring effect often seen in machine learning (ML)-based post-processing, as the input variables are not inherently affected by errors relative to observed values, and the predictions at FT = t are expected to lie between those at FT = t ± Δt. During the inference phase, two different NWP outputs at the same forecast lead time are used to generate an intermediate forecast for the projection time (Fig. 1b).

Fig. 1

Fig. 2

DeepMedcast is primarily designed to generate intermediate forecasts between two NWP models. However, the same DNN model can be applied recursively to generate intermediate forecasts between more than two NWP models. For instance, by taking intermediate forecasts between two pairs of NWP models, DeepMedcast can generate an intermediate result between four NWP models (Fig.2). This recursive approach could be

195 extended further to create intermediate forecasts between 8, 16, or even more NWP models.

196

197 *2.2 Data used for the study*

198 The NWP model used for training is GSM, which is operated by JMA four times a day
199 (at 00, 06, 12, and 18 UTC as initial times). The training period spans nine years, from
200 January 2013 to December 2021, while the validation period covers one year, from January
201 to December 2022. GSM had a horizontal resolution of approximately 20 km until March
202 2023, after which it was upgraded to about 13 km (JMA 2024). The GSM data used in this
203 study is stored at JMA, where it is trimmed and linearly interpolated onto a 121 × 151 grid
204 with a resolution of 0.25 degrees by 0.2 degrees around Japan (Fig. 3). Hereafter, we refer Fig. 3
205 to this as the target grid domain.

206 The forecast variables include wind components (U, V), temperature (T), and relative
207 humidity (RH) at both the surface and the 700 hPa level, as well as mean sea-level pressure
208 at the surface (P_{sea}). To reduce computational cost and execution time, each variable is
209 used individually to train separate networks, with each network dedicated to a single variable.
210 That means, each DNN model always takes two input channels (at FT = t - Δt and FT = t +
211 Δt) and outputs one channel (at FT = t) for DeepMedcast. Both input channels are utilized
212 by swapping their order, i.e., both FT = t - Δt and FT = t + Δt, as well as FT = t + Δt and FT
213 = t - Δt, are employed to preserve symmetry. This strategy encourages the network to learn
214 symmetric representations, so that meteorologically significant features—such as the center

positions of TCs or extratropical cyclones, frontal boundaries, and shear lines—approximately align with the arithmetic mean of the corresponding features from the two input channels. The networks trained with 700 hPa data are employed to generate intermediate forecasts for the upper atmosphere (e.g., 850 hPa, 700 hPa, and 500 hPa) to reduce computational costs. The forecast lead times used in this study are t = 9, 10, 11, 12, 13, and 14 hours, with $\Delta t = \pm 3$ and ± 6 hours, corresponding to t and Δt in Fig. 1, determined by taking both computational costs and accuracy into account.

For the case studies in Section 3, MSM, IFS, GraphCast, and Pangu-Weather are used along with GSM for inference. MSM is operated by JMA eight times a day (at 00, 03, ..., and 21 UTC as initial times) with a 5 km horizontal resolution, providing forecasts up to FT = 78 hours for 00 and 12 UTC initial times and up to FT = 39 hours for other initial times. IFS data, provided by ECMWF for the World Meteorological Organization (WMO) members, has a horizontal resolution of 0.5 degrees and is initialized four times daily at 00, 06, 12, and 18 UTC. Both GraphCast and Pangu-Weather have a horizontal resolution of 0.25 degrees, with data initialized at 00, 06, 12, and 18 UTC. These NWP outputs are linearly interpolated to the target grid domain for inference.

2.3 DNN model architecture

In this study, a U-Net architecture (Ronneberger et al. 2015) is applied as the DNN model. The structure of the network is illustrated in Fig. 4. The encoder part of the U-Net

Fig. 4

utilizes a convolutional network with kernel size = 3, stride = 1, and padding = 1 for convolution operations. To progressively reduce the image size, MaxPooling layers with kernel size = 2 and stride = 2 are employed. This downsampling process continues until the image size is reduced to 1/8 of the original dimensions, at which point the channel count reaches 2048, starting from an initial 2 channels that are expanded to 256 channels. In each downsampling stage, the image size is halved while the number of channels doubles. In the decoder part, transposed convolutional layers with kernel size = 2 and stride = 2 are applied to upsample the feature maps, restoring the image to its original size while reducing the channel count by half at each stage. By the final stage, the image is returned to its original dimensions with 256 channels, which are then reduced to 1 channel in the output layer. The activation functions used in this network include rectified linear units (ReLU, Nair and Hinton 2010) for all layers except the output layer, which uses a sigmoid function to ensure output values are scaled between 0 and 1.

During the training phase, the two input channels and one ground truth channel, each consisting of 121 × 151 grids, are normalized to a value range of 0 to 1 using the maximum and minimum values across all three channels. Specifically, for T, RH, and P_{sea}, the normalization is applied as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where x' is the normalized value, x is the input value, and $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values, respectively. For wind components U and V, $x_{\max}$ and

$x_{\min}$ are defined as:

$$x_{\max} = \max(|x_{\max}|,|x_{\min}|)$$

$$x_{\min} = -x_{\max}$$

and the same normalization is applied.

After normalization, the values are extended to 128 × 158 grids by copying the last column and row to adjust to the network structure. The output values are compared with the normalized and extended ground truth values using the mean square error (MSE) as the loss function. We employ Adam (Kingma and Ba 2014) as optimization.

During the inference phase, the input values are normalized to the 0 to 1 range using the maximum and minimum values of the two input channels. The output values are then denormalized using the same maximum and minimum values, and resized back to 121 × 151 grids by trimming the extended columns and rows, providing predictions at the target grid domain.

3. Results

In this section, we demonstrate the capability of DeepMedcast through four case studies and verification results. The forecast data used here is from a period beginning in January 2023, which is independent of the DNN's training and validation periods. The case studies compare the atmospheric fields generated by DeepMedcast with those obtained via arithmetic mean of the NWP outputs.

3.1 Case 1: Position discrepancy in a typhoon forecast between GSM and MSM

The first case study examines a typhoon forecast where there is a positional discrepancy between GSM and MSM. Figure 5 shows the predictions at FT = 51 hours based on the initial time of 12 UTC on 12 August 2023. This case focuses on Typhoon LAN which was moving northwest over the ocean south of Japan. At FT = 51 hours, GSM predicted the typhoon's position at 33.3°N, 137.1°E, while MSM placed it southwest at 32.7°N, 135.7°E. Both models predicted a central pressure of 960 hPa, with the maximum wind speed of 79 kt (1 kt $\simeq$ 0.514 m s^{-1}) (GSM) and 68 kt (MSM) (Figs. 5a and 5b).

Fig. 5

When the mean sea-level pressure and surface wind components from GSM and MSM were averaged arithmetically, the typhoon's center split into two, aligning with the predicted positions from each model (Fig. 5c). Such a result is evidently unrealistic and lacks interpretability. The central pressure weakened to 974 hPa, and the maximum wind speed reduced to 58 kt, which made the forecast meteorologically unnatural, with the typhoon taking on an elongated structure and weakening wind speeds near the center. This resulted in a forecast that was difficult to explain and potentially misleading.

In contrast, DeepMedcast generated a plausible forecast, placing the typhoon at 33.0°N, 136.4°E, halfway between GSM and MSM predictions (Fig. 5d). The typhoon maintained a single, natural, and interpretable shape with a central pressure of 960 hPa and the maximum wind speed of 69 kt, representing an intermediate forecast between the two NWP models.

Figure 6 compares the DeepMedcast outputs with different input orders. Figure 6a shows the result when GSM and MSM are used as inputs in that order (identical to Fig. 5d), while Fig. 6b shows the result when the input order is reversed (MSM-GSM). As expected, the predicted structure—including the typhoon center position, central pressure, and surrounding wind field—remains qualitatively consistent, despite minor differences due to the asymmetry of the trained neural network. This indicates that while DeepMedcast is not strictly order-invariant, the resulting intermediate forecasts are robust to changes in input order.

Fig. 6

3.2 Case 2: Discrepancy in a front position forecast between GSM and MSM

The second case study examines a forecast where there was a positional discrepancy in the predicted location of a front between GSM and MSM. Figure 7 shows the predictions at FT = 30 hours based on the initial time of 00 UTC on 17 June 2024. At the initial time, a stationary front was located south of Japan (not shown), and by FT = 30 hours, the front was predicted to move northward toward Tokyo (indicated by the blue circles in the figure).

Fig. 7

GSM predicted the front to the south of Tokyo, indicated by the blue dashed line, with a clear wind direction and speed shear, which corresponds well with the 21°C isotherm around Tokyo (Fig. 7a). In contrast, MSM placed the front north of Tokyo (green dashed line), also with a clear wind direction and speed shear aligned with the 21°C isotherm (Fig. 7b), resulting in a positional discrepancy between the two NWP models. Consequently, GSM

315 predicted a northerly to northeasterly wind and cooler temperatures around Tokyo, while
316 MSM predicted southerly to southwesterly winds and warmer temperatures, leading to
317 significant differences in the forecast for Tokyo.

318 The arithmetic mean of the GSM and MSM predictions (Fig. 7c) results in a split structure
319 for the front, with wind shear corresponding to the locations predicted by GSM and MSM
320 (shown by the purple lines), while the 21°C isotherm is predicted between the two fronts.
321 This demonstrates that when there is a discrepancy in the predicted front position, simple
322 averaging of the atmospheric fields leads to an unnatural and uninterpretable forecast that
323 cannot maintain the original front structure.

324 In contrast, DeepMedcast (Fig. 7d) generates a clear wind direction and speed shear at
325 the intermediate position between the GSM and MSM predictions (indicated by the brown
326 dashed line), which corresponds well with the 21°C isotherm. DeepMedcast successfully
327 produces a realistic and interpretable intermediate forecast while preserving the structure of
328 the original front.

329

330 *3.3 Case 3: Significant difference in low-pressure system position between GSM and MSM*

331 The third case study highlights a situation where there was a large difference in the Fig. 8
332 predicted position of a low-pressure system between GSM and MSM. Figure 8 shows
333 surface wind and mean sea-level pressure, along with temperature and dew-point
334 depression at 850 hPa, at FT = 75 hours based on the initial time of 12 UTC on 28 August

335 2024. At the initial time, Typhoon SHANSHAN was located south of Kyushu (see Fig. 3,
336 location 4) at 30.6°N, 130.2°E, slowly moving northward (not shown). By 15 UTC on 31
337 August (FT = 75 hours), the system, which had either remained a tropical storm or
338 transitioned into a low-pressure system, was predicted by GSM to be south of the Kanto
339 region (see Fig. 3, location 3) at 35.4°N, 139.8°E (Fig. 8a), while MSM placed it east of
340 Hokkaido (see Fig. 3, location 1) at 42.4°N, 147.3°E (Fig. 8b).

341 This case highlights a large positional difference of about 1000 km between the GSM
342 and MSM predictions. When the arithmetic mean of these is taken (Fig. 8c), it results in two
343 distinct low-pressure systems at the positions predicted by each model, creating an
344 uninterpretable forecast. In contrast, DeepMedcast predicted a single low-pressure system
345 located between the two forecasts, at 39.5°N, 143.0°E, off the Pacific coast of Tohoku (see
346 Fig. 3, location 2; Fig. 8d). Additionally, when examining the moisture area at 850 hPa (dew-
347 point depression < 3°C), the arithmetic mean shows moist areas surrounding both the GSM
348 and MSM low-pressure systems, with a relatively dry region in between. On the other hand,
349 DeepMedcast represents a moist area around its low-pressure system, corresponding well
350 with the surface pressure field, providing a realistic and interpretable forecast.

351

352 *3.4 Case 4: Intermediate forecast between four NWP models for typhoon KHANUN*

353 The fourth case study presents an intermediate forecast between four NWP models: Fig. 9
354 GSM, IFS, GraphCast, and Pangu-Weather. Figure 9 shows surface wind and mean sea- Table.1

355 level pressure at FT = 108 hours, based on the initial time of 12 UTC on 2 August 2023. At
356 the initial time, Typhoon KHANUN was located west of Okinawa (see Fig. 3, location 5) at
357 26.2°N, 125.6°E, slowly moving westward (not shown). By 00 UTC on 7 August (FT = 108
358 hours), the central position was predicted by the four models to be at 31.4°N, 131.1°E (GSM),
359 28.9°N, 133.0°E (IFS), 28.0°N, 131.2°E (GraphCast), and 28.8°N, 130.7°E (Pangu-
360 Weather). The typhoon central position, central pressure, and maximum wind speed at FT
361 = 108 hours for each model are summarized in Table 1.

362 When the arithmetic mean of the four models is taken, the center splits into two, with a
363 weakened central pressure of 979 hPa and the maximum wind speed of 36 kt (Fig. 9e), both
364 the same or weaker than the predictions of any individual model. In contrast, DeepMedcast
365 predicted a single center at 29.3°N, 131.4°E, with the central pressure of 964 hPa and the
366 maximum wind speed of 42 kt (Fig. 9f), representing an intermediate intensity forecast
367 between the four NWP models. The average central latitude, longitude, and pressure of the
368 four NWP models were 29.3°N, 131.5°E, and 964 hPa, respectively, matching
369 DeepMedcast's prediction. The average maximum wind speed of the four models was 49
370 kt, meaning DeepMedcast’s forecast was slightly weaker than the average.

371 Figure 10 illustrates the effect of changing the order in which intermediate forecasts are Fig. 10
372 taken when combining the four NWP models. Figure 10a is identical to Fig. 9f and shows
373 the result when intermediate forecasts are first generated between GSM and IFS, and
374 between GraphCast and Pangu-Weather, followed by taking an intermediate forecast

between those two results. Figure 10b shows the case where the intermediate forecasts are first taken between GSM and GraphCast, and between IFS and Pangu-Weather, then combined. Figure 10c presents the result when intermediate forecasts are first taken between GSM and Pangu-Weather, and between IFS and GraphCast. As in the two-model case discussed in Section 3.1, the outputs differ slightly due to the inherent asymmetry of the trained network and the recursive nature of the procedure. However, the predicted typhoon structure, including its central position, pressure, and wind field, remains qualitatively consistent across all three cases. This suggests that although DeepMedcast is not strictly order-invariant, it produces robust intermediate forecasts in practice.

3.5 Statistical evaluation of DeepMedcast using surface wind observations

This section presents the verification results of DeepMedcast generated from GSM and MSM. Surface wind predictions from DeepMedcast, along with the input GSM and MSM forecasts, were verified against observations from the Automated Meteorological Data Acquisition System (AMeDAS), an automated observation network operated by JMA. The verification metric is the root mean square error (RMSE), defined as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{n=1}^{N}(F_{nt} - O_{nt})^2}$$

where T and N are the numbers of forecast times and stations used for verification, and F_{nt} and O_{nt} represent the forecast and observed winds at station n and time t,

respectively. Predictions from DeepMedcast, GSM, and MSM were linearly interpolated to each AMeDAS station from the four surrounding grid points.

Figure 11 shows the RMSE of wind speed (Fig. 11a) and wind direction (Fig. 11b) by forecast lead time, ranging from 3 to 39 hours. The verification was conducted over one year, from January to December 2023, using forecasts initialized four times daily (00, 06, 12, and 18 UTC), independent of the training and validation periods. In both panels, the red, blue, and green lines represent DeepMedcast, GSM, and MSM, respectively. As shown in Fig. 11, DeepMedcast achieves lower RMSEs for both wind speed and direction across all forecast lead times compared to its input models. The RMSE for wind direction in Fig. 11b shows a 6-hourly fluctuation pattern, reflecting the four-times-daily initialization and diurnal variation.

Fig. 11

4. Discussion

4.1 Contributions to operational forecasting

As demonstrated by the case studies and verification in Section 3, DeepMedcast is capable of generating plausible and interpretable intermediate forecasts. The ability to generate intermediate forecasts between multiple models is expected to significantly contribute to operational forecasting. As mentioned in the Introduction, TC track forecasts, which are based on consensus from multiple NWP models, specifically the average position of the TC center predicted by these models, serve as primary reference in JMA's operational

forecasting. Consequently, forecasters responsible for general, marine, and aviation forecasts must ensure that their forecasts align with the TC track forecasts. However, since no NWP model inherently conforms to the TC track forecasts, forecasters must adjust the existing NWP outputs in their minds to construct forecast scenarios that follow the TC track forecasts. DeepMedcast has a capability to provide two-dimensional wind and pressure fields that align with TC track forecasts, which could greatly improve the efficiency and standardization of tasks for operational forecasting.

Additionally, DeepMedcast can be effectively utilized in operational forecasting when there are discrepancies in the predicted positions of low-pressure systems or fronts among multiple models. When significant differences exist among NWP models, forecasters need to choose between two options: either using one model as the main scenario while treating others as alternative scenarios, or applying averaging methods. By generating an intermediate state between two or more NWP models, DeepMedcast provides a forecast scenario that is more plausible than individual NWP models. However, it should be noted that it does not explicitly represent the variability among the original NWP models. This issue is not unique to DeepMedcast but is also present when using arithmetic or weighted averaging of multiple NWP models or post-processing techniques. Since variability is an indicator of forecast uncertainty, especially for longer lead times, it is important to take this information into account in daily operational forecasting. One practical approach is to incorporate the values of the original NWP models and their spread alongside the

intermediate forecast. This framework can further enhance operational forecasting by offering a practical way to take model variability into account.

4.2 Key Features of DeepMedcast's Architecture

There are two important features associated with DeepMedcast's architecture. The first is its flexibility in increasing the amount of training data. As mentioned in Section 2, this study used t = 9–14 and Δt = ±3, ±6, and in our experience, increasing t and Δt leads to better forecast representation. One common issue in training DNN models is a lack of sufficient training data (e.g., Deng 2009; LeCun 2015). However, in the case of DeepMedcast, more training data can easily be generated by increasing t and Δt or by adding additional NWP models. It is important to note, though, that increasing t and Δt requires more memory and computational time, which should be considered when expanding the dataset.

The second is DeepMedcast's maintainability. Despite being trained solely on 20-km resolution GSM data, DeepMedcast works effectively not only with 13-km resolution GSM data but also with other NWP models such as MSM, IFS, GraphCast, and Pangu-Weather. This is significant because most AI or ML methods in meteorology learn the relationship between input and target data, and when the characteristics of the input data change due to NWP model updates, retraining, fine-tuning, and/or online learning are usually required. While this is an unavoidable task for most AI or ML methods in meteorology, it is a time-consuming yet essential task that operational centers have traditionally managed. However,

DeepMedcast can be applied to various NWP models without updating the DNN model since it is not designed to correct NWP model biases, which significantly reduces the maintenance costs for operational centers.

5. Summary

In this paper, we introduced DeepMedcast, a novel deep learning-based approach for producing intermediate forecasts between two or more NWP models. DeepMedcast was developed to generate plausible and interpretable intermediate forecast, bridging the gap between NWP model outputs.

A key advantage of DeepMedcast is its applicability to various NWP outputs without the need for retraining or fine-tuning the DNN. By providing plausible intermediate forecasts, DeepMedcast can significantly enhance the efficiency and standardization of operational forecasting tasks, including general, marine, and aviation forecasts.

Although DeepMedcast introduces some advancements, further research and development are needed to address several challenges. In this study, U-Net was employed as the DNN architecture; however, advanced methods such as Transformers (Vaswani et al. 2017; Dosovitskiy et al. 2020) and Diffusion models (Song and Ermon 2019; Ho et al. 2020) may further enhance DeepMedcast's representational capabilities. As shown in the case study in Section 3.4, the current method tends to slightly underestimate the maximum wind speed near TCs. Enhancing the DNN could help resolve this issue. Additionally, while

this study trained separate networks for each physical variable to reduce computational cost, incorporating multiple physical variables as input could potentially enhance forecast accuracy. By developing post-processing methods that use DeepMedcast as input, it would be possible to provide even more accurate predictions. This study focused on generating intermediate forecasts using a 1:1 weighting ratio, meaning that the two input models were given equal weight. Future work should explore methods for generating intermediate forecasts with other weighting ratios, such as 1:2. This would enable the application of DeepMedcast to cases involving several models that is not a power of two, such as finding an intermediate forecast among three models. Furthermore, while this study demonstrated intermediate forecasts using two or four NWP models, DeepMedcast could be extended to 8, 16, or more inputs, enabling the use of multiple NWP and ensemble models. Lastly, this study did not address intermediate precipitation forecasts. Since precipitation is one of the most critical variables in weather forecasting, future work will focus on developing intermediate precipitation forecasts.

Data Availability Statement

The datasets used in this study are available from the following sources: The GSM and MSM data are operationally produced by JMA and can be accessed through the Japan Meteorological Business Support Center (http://www.jmbsc.or.jp/en/index-e.html). The IFS data are accessible to WMO members via the ECMWF website

(https://www.ecmwf.int/en/forecasts/datasets/wmo-additional). The GraphCast and Pangu-Weather source code and plugins are available under open-source licenses in the ECMWF GitHub repository (https://github.com/ecmwf-lab/ai-models). Pre-trained models of Pangu-Weather and GraphCast, used without modification to generate forecast data, are specifically accessible at https://github.com/ecmwf-lab/ai-models-panguweather and https://github.com/ecmwf-lab/ai-models-graphcast.

Acknowledgments

We acknowledge ECMWF for providing IFS data available to WMO members through their website (https://www.ecmwf.int/en/forecasts/datasets/wmo-additional). Additionally, we are grateful for the availability of GraphCast and Pangu-Weather and extend our thanks to their respective developers. GraphCast and Pangu-Weather were used without modification to generate forecast data and are accessible via the ECMWF AI GitHub repository (https://github.com/ecmwf-lab/ai-models), supported by ECMWF. The author declares no conflicts of interest associated with this manuscript.

References

Agueh, M. and G. Carlier, 2011: Barycenters in the Wasserstein space SIAM Journal on Mathematical Analysis, 43(2), 904–924. [Available at https://hal.science/hal-00637399v1/document]

514 Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-Weather: A 3D High-
515 Resolution System for Fast and Accurate Global Weather Forecast. *arXiv preprint*
516 *arXiv:2211.02556.*
517 Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tina, 2023: Accurate medium-range
518 global weather forecasting with 3D neural networks. *Nature,* **619**, 533–538.
519 https://doi.org/10.1038/s41586-023-06185-3.
520 Bodnar, C., W. P. Bruinsma, A. Lucic, M. Stanley, J. Brandstetter, P. Garvan, M. Riechert,
521 J. Weyn, H. Dong, A. Vaughan, J. K. Gupta, K. Tambiratnam, A. Archibald, E. Heider,
522 M. Welling, R. E. Turner, P. Perdikaris. 2024: Aurora: A Foundation Model of the
523 Atmosphere. *arXiv preprint arXiv:2405.13063*.
524 Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, A. Anandkumar, 2023:
525 Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere. *arXiv*
526 *preprint arXiv:2306.03838.*
527 Brown, A., S. Milton, M. Cullen, B. M. J. Golding, and A. Shelly, 2012: Unified modeling
528 and prediction of weather and climate: a 25 year journey, *Bull. Amer. Meteor. Soc.*, **93**,
529 1865–1877.
530 Cangialosi, J., B.J. Reinhart, and J. Martinez, 2023: *National Hurricane Center verification*
531 *report, 2023 Hurricane Season*. Natinal Hurricane Center, 81pp. [Available at
532 https://www.nhc.noaa.gov/verification/pdfs/Verification_2023.pdf.]
533 Chen, K., T. Han, J. Gong, L. Bai, F. Ling, J. Luo, X. Chen, L. Ma, T. Zhang, R. Su, Y. Ci,

B. Li, X. Yang, W. Ouyang, 2023: FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead. *arXiv preprint arXiv:2304.02948.*

Deng, J., W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, 2009: ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248–255.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, 2020: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929.*

Dowell, D. C., C. R. Alexander, E. P. James, S. S. Weygandt, S. G. Benjamin, G. S. Manikin, B. T. Blake, J. M. Brown, J. B. Olson, M. Hu, T. G. Smirnova, T. Ladwig, J. S. Kenyon, R. Ahmadov, D. D. Turner, J. D. Duda, T. I. Alcott, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Duc, L. and Y. Sawada, 2024: Geometry of rainfall ensemble means: from arithmetic averages to Gaussian-Hellinger barycenters in unbalanced optimal transport. *J. Meteor. Soc. Japan*, **102**, 35-47, https://doi.org/10.2151/jmsj.2024-003.

ECMWF, 2024: IFS Documentation. European Centre for Medium-Range Weather Forecasts. [Available at https://www.ecmwf.int/en/publications/ifs-documentation.]

Gopalakrishnan, S. G., F. Marks Jr., X. Zhang, J.-W. Bao, K.-S. Yeh, and R. Atlas, 2011: The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.*, **139**, 1762–1784.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861.

Hamill, T. M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for Statistical Postprocessing of Probability of Precipitation and Deterministic Precipitation Amount. *Mon. Wea. Rev.*, **145**, 3441–3463.

Han T., S. Guo, F. Ling, K. Chen, J. Gong, J. Luo, J. Gu, K. Dai, W. Ouyang, L. Bai, 2024: FengWu-GHR: Learning the Kilometer-scale Medium-range Global Weather Forecasting. *arXiv preprint* arXiv:2402.00059.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239.*

ICAO, 2016: Guidance on the harmonized WAFS grids for Cumulonimbus cloud, icing and turbulence forecasts (version 2.6). International Civil Aviation Organization, 16pp. [Available at https://www.icao.int/airnavigation/METP/MOG%20WAFS%20Reference%20Documents/WAFS_HazardGridUserGuide.pdf.]

574 Inverarity, G. W., W. J. Tennant, L. Anton, N. E. Bowler, A. M. Clayton, M. Jardak, A. C.
575 Lorence, F. Rawlins, S. A. Thompson, M. S. Thurlow, D. N. Walters, and M. A. Wlasak,
576 2023: Met Office MOGREPS-G initialisation using an ensemble of hybrid four-
577 dimensional ensemble variational (En-4DEnVar) data assimilations. *Quart. J. Roy.*
578 *Meteor. Soc.*, **149**, 1138–1164.
579 JMA, 2018: Instruction for guidance. *Report of Numerical Prediction Division*. **64**, Japan
580 Meteorological Agency, 248 pp (in Japanese). [Available at
581 https://www.jma.go.jp/jma/kishou/books/nwpreport/64/No64_all.pdf.]
582 JMA, 2022: *Annual Report on the Activities of the RSMC Tokyo - Typhoon Center 2022.*
583 Japan Meteorological Agency, 143pp. [Available at https://www.jma.go.jp/jma/jma-
584 eng/jma-center/rsmc-hp-pub-eg/AnnualReport/2022/Text/Text2022.pdf.]
585 JMA, 2024: *Outline of the operational numerical weather prediction at the Japan*
586 *Meteorological Agency.* Japan Meteorological Agency, 262pp. [Available at
587 https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-latest-nwp/index.htm.]
588 Kingma, D. P., and J. Ba, 2014: Adam: A Method for Stochastic Optimization. *arXiv*
589 *preprint arXiv:1412.6980.*
590 Lam, R., A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S.
591 Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals,
592 J. Stott, A. Pritzel, S. Mohamed and P. Battaglia, 2022: GraphCast: Learning skillful
593 medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794.*

Lang, S., M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M. C. A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. B. Bouallègue, A. P. Nemesio, P. D. Dueben, A. Brown, F. Pappenberger, F. Rabier, 2024: AIFS — ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465.*

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444.

Le Coz, C., A. Tantet, R. Flamary, and R. Plougonven, 2023: A barycenter-based approach for the multi-model ensembling of subseasonal forecasts. *arXiv preprint arXiv:2310:17933.*

McCann, R. J., 1997: A Convexity Principle for Interacting Gases. *Advances in Mathematics*, **128(1)**, 153–179.

Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10), Haifa, Israel, 807–814.

NCEP, 2016: *Global Forecast System - Global Spectral Model (GSM) - v13.0.2*. [Available at https://vlab.noaa.gov/web/gfs/documentation.]

Nishimura, S., and J. Fukuda, 2019: Advancement of Tropical Cyclone Track Forecasts. *Yohou Gijutsu Kenshu Text*, **24**, 114–141 (in Japanese). [Available at https://www.jma.go.jp/jma/kishou/books/yohkens/24/all.pdf.]

Pathak, J., S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A.

614 Anandkumar, 2022: FourCastNet: A Global Data-driven High-resolution Weather Model
615 using Adaptive Fourier Neural Operators. *arXiv preprint arXiv:2202.11214.*
616 Peyré, G., and M. Cuturi, 2020: Computational Optimal Transport. *arXiv preprint arXiv:*
617 *1803.00567*
618 Price, I., A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T.
619 Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, M. Willson, 2023: GenCast:
620 Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint*
621 *arXiv:2312.15796.*
622 Primo, C., B. Schulz, S. Lerch, and R. Hess, 2024: Comparison of Model Output Statistics
623 and Neural Networks to Postprocess Wind Gusts. *arXiv preprint arXiv:2401.11896.*
624 Roberts, N., A. Benjamin, E. Gavin, M. Stephen, R. Fiona, S. Caroline, T. Tomasz, A.
625 Paul, B, Laurence. C. Neil, F. Ben, F. Jonathan, G. Tom, H. Leigh, H. Aaron, H.
626 Katharine, J. Simon, J. Caroline, M. Ken, S. Christopher, S. Michael, W. Bruce, B.
627 Simon, B. Mark, B. Daniel, B. Anna, B. Clare, C. Robert, C. Sean, C. Ric, H. Roger, H.
628 Kathryn, H. Teresa, M. Marion, P. Jon, P. Tim, S. Victria, S. Eleanor, and W. Mark,
629 2023: IMPROVER: The New Probabilistic Postprocessing System at the Met Office. *Bull*
630 *Amer. Meteor. Soc.*, **104**, E680–E697.
631 Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for
632 Biomedical Image Segmentation. *arXiv preprint arXiv:1505.04597.*
633 Simon, A., A. B. Penny, M. DeMaria, J. L. Franklin, R. J. Pasch, E. N. Rappaport, and D.

A. Zelinsky, 2018: A description of the real-time HFIP Corrected Consensus Approach (HCCA) for Tropical Cyclone Track and Intensity Guidance. *Wea. Forecasting*, **33**, 37–57, https://doi.org/10.1175/WAF-D-17-0068.1.

Song, Y. and S. Ermon, 2019: Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv preprint arXiv:1907.05600.*

Vannitsem, S., J. B. Bremnes, J. Demaeyer, G. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts, S. Theis, A. Atencia, Z. B. Bouallègue J. Bhend, M. Dabernig, L. D. Cruz, L. Hieta, O. Mestre, L. Moret, I. O. Plenkovic, M. Schmeits, J. Ylhäisi, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bull Amer. Meteor. Soc.,* **102**, E681–E699.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin 2017: Attention Is All You Need, *arXiv preprint arXiv:1706.03762.*

Vislocky, R. L., and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*, **78**, 2851–2857.

WMO, 2013: Cascading Process to Improve Forecasting and Warning Services. *Bulletin nº,* **62**, 11–15. [Available at https://public.wmo.int/en/resources/bulletin/cascading-process-improve-forecasting-and-warning-services.]

654 List of Figures

655 Fig. 1 DeepMedcast framework for training and inference. (a) During the training phase,
656 two forecast lead times from the same NWP model (NWP1 at FT = t - Δt and FT = t + Δt)
657 are used as input, and the output from the DNN is compared with the same NWP
658 model's forecast at FT = t as the ground truth to train the network. (b) During the
659 inference phase, predictions from two different NWP models (NWP1 and NWP2) at the
660 same lead time (FT = t) are used as input to generate an intermediate forecast between
661 the two models at FT = t.

662

663 Fig. 2 The recursive application of DeepMedcast, where intermediate forecasts are first
664 generated between two NWP models (NWP1 and NWP2, NWP3 and NWP4), followed
665 by the creation of an additional intermediate forecast between the outputs of the first two
666 pairs.

667

668 Fig. 3 The target grid domain for this study. 121 × 151 grids with 0.25-degree × 0.2-
669 degree resolution around Japan. The dots on the map represent these grid points. The
670 numbers and region names indicated in the figure are used in the case studies in
671 Section 3.

672

673 Fig. 4 The DNN architecture used in DeepMedcast. The model takes two input channels

and outputs a single channel. Input data is normalized using the maximum and minimum values, and during inference, the same values are applied for the denormalization process.

Fig. 5 Comparison of Typhoon LAN predictions by (a) GSM, (b) MSM, (c) the arithmetic mean, and (d) DeepMedcast. The forecasts are based on the initial time of 12 UTC on 12 August 2023 with a forecast lead time of 51 hours. The black contours indicate mean sea-level pressure and wind barbs (units in kt) show surface winds.

Fig. 6 Comparison of DeepMedcast outputs with different input orders for the case in Fig. 5. (a) Result when GSM and MSM are provided in that order (same as Fig. 5d). (b) Result when the input order is reversed (MSM-GSM). While slight differences are present due to network asymmetry, the outputs remain qualitatively identical.

Fig. 7 Comparison of predicted front positions by (a) GSM, (b) MSM, (c) the arithmetic mean, and (d) DeepMedcast. The forecasts are based on the initial time of 00 UTC on 17 June 2024 with a forecast lead time of 30 hours. The black contours indicate mean sea-level pressure, the red contours represent surface temperature, and wind barbs (units in kt) show surface winds. The blue, green, purple, and brown dashed lines represent the predicted front positions by GSM, MSM, the arithmetic mean, and

694 DeepMedcast, respectively. Blue circles indicate the location of Tokyo.

695

696 Fig. 8 Comparison of predicted low-pressure systems by (a) GSM, (b) MSM, (c) the
697 arithmetic mean, and (d) DeepMedcast. The forecasts are based on the initial time of 12
698 UTC on 28 August 2024 with a forecast lead time of 75 hours. The black contours
699 indicate mean sea-level pressure, the wind barbs (units in kt) represent surface winds,
700 the red contours show 850 hPa temperature, and the shaded regions in green and
701 yellow highlight areas where the dew-point depression at 850 hPa is below 3°C and
702 above 15°C, respectively.

703

704 Fig. 9 Comparison of Typhoon KHANUM predictions by (a) GSM, (b) IFS, (c) GraphCast,
705 (d) Pangu-Weather, (e) the arithmetic mean, and (f) DeepMedcast. The predictions are
706 based on the initial time of 12 UTC on 2 August 2023 with a forecast lead time of 108
707 hours. The black contours indicate mean sea-level pressure and wind barbs (units in kt)
708 show surface winds.

709

710 Fig. 10 DeepMedcast forecasts based on the initial time of 12 UTC on 2 August 2023
711 with a forecast lead time of 108 hours for Typhoon KHANUN using different orders of
712 intermediate forecast generation from the four NWP models: GSM, IFS, GraphCast, and
713 Pangu-Weather. (a) Intermediate forecasts are first taken between GSM and IFS, and

714 between GraphCast and Pangu-Weather, then combined. (b) GSM–GraphCast and
715 IFS–Pangu-Weather. (c) GSM–Pangu-Weather and IFS–GraphCast.
716
717 Fig. 11 Root mean square error (RMSE) of (a) surface wind speed and (b) surface wind
718 direction for DeepMedcast (red), input GSM (blue), and input MSM (green) forecasts,
719 verified against AMeDAS observations. The verification period spans one year, from
720 January to December 2023, with forecasts initialized at 00, 06, 12, and 18 UTC.
721
722

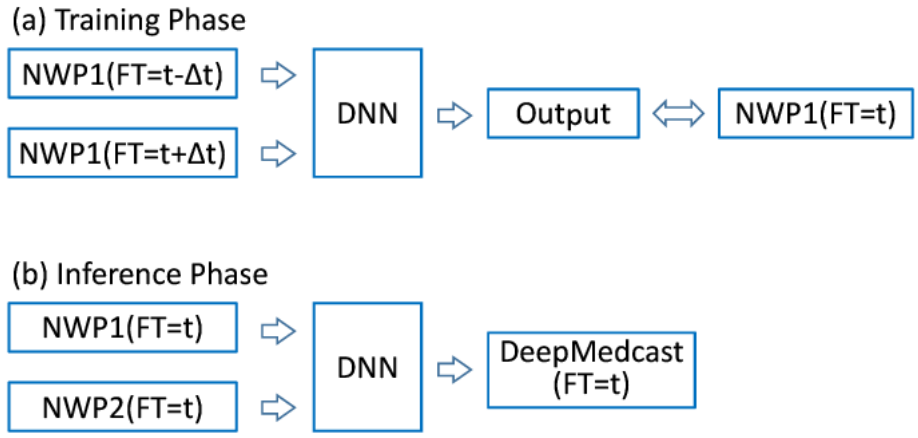


Fig. 1 DeepMedcast framework for training and inference. (a) During the training phase, two forecast lead times from the same NWP model (NWP1 at FT = t - Δt and FT = t + Δt) are used as input, and the output from the DNN is compared with the same NWP model's forecast at FT = t as the ground truth to train the network. (b) During the inference phase, predictions from two different NWP models (NWP1 and NWP2) at the same lead time (FT = t) are used as input to generate an intermediate forecast between the two models at FT = t.

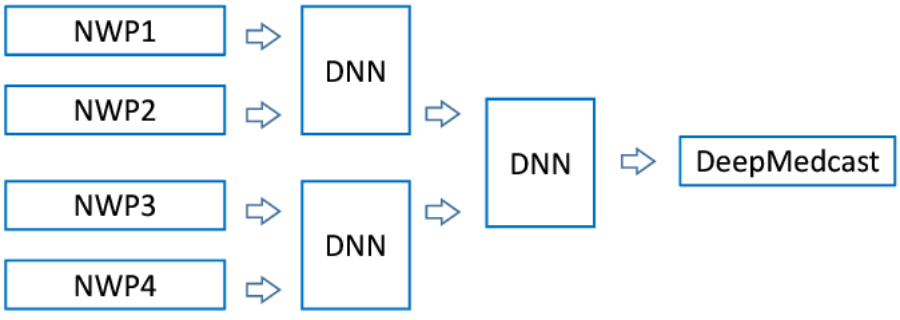
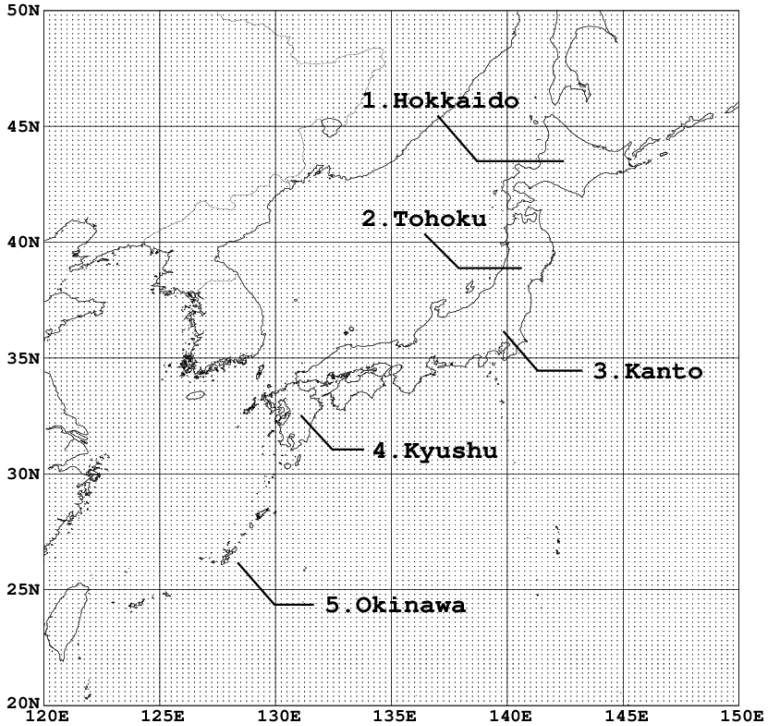


Fig. 2 The recursive application of DeepMedcast, where intermediate forecasts are first generated between two NWP models (NWP1 and NWP2, NWP3 and NWP4), followed by the creation of an additional intermediate forecast between the outputs of the first two

737 pairs.



738

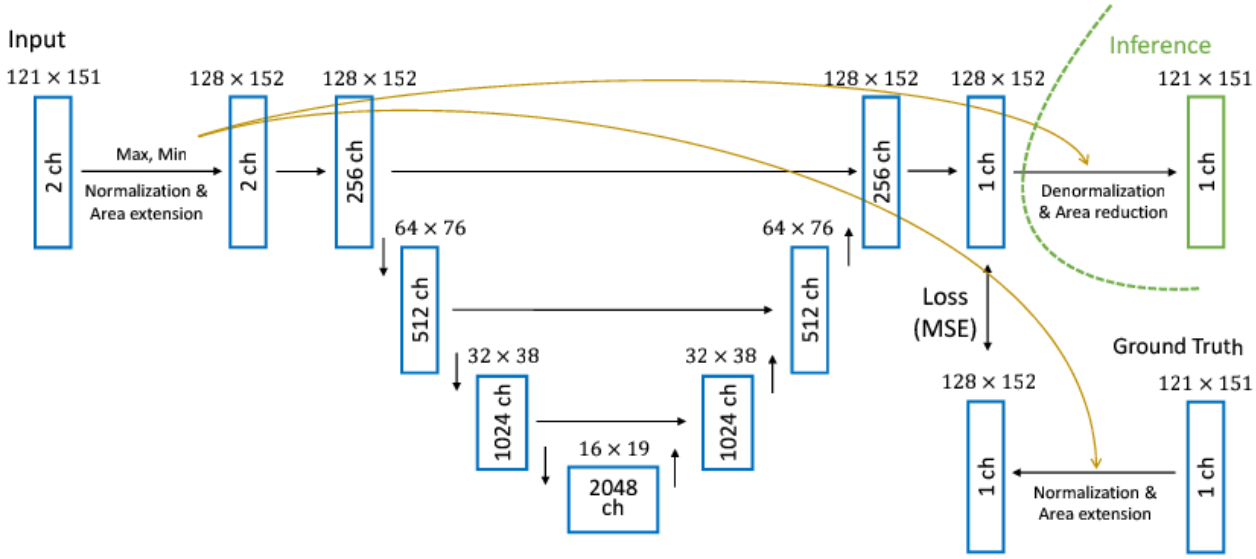
739 Fig. 3 The target grid domain for this study. 121 × 151 grids with 0.25-degree × 0.2-

740 degree resolution around Japan. The dots on the map represent these grid points. The

741 numbers and region names indicated in the figure are used in the case studies in

742 Section 3.

743



744

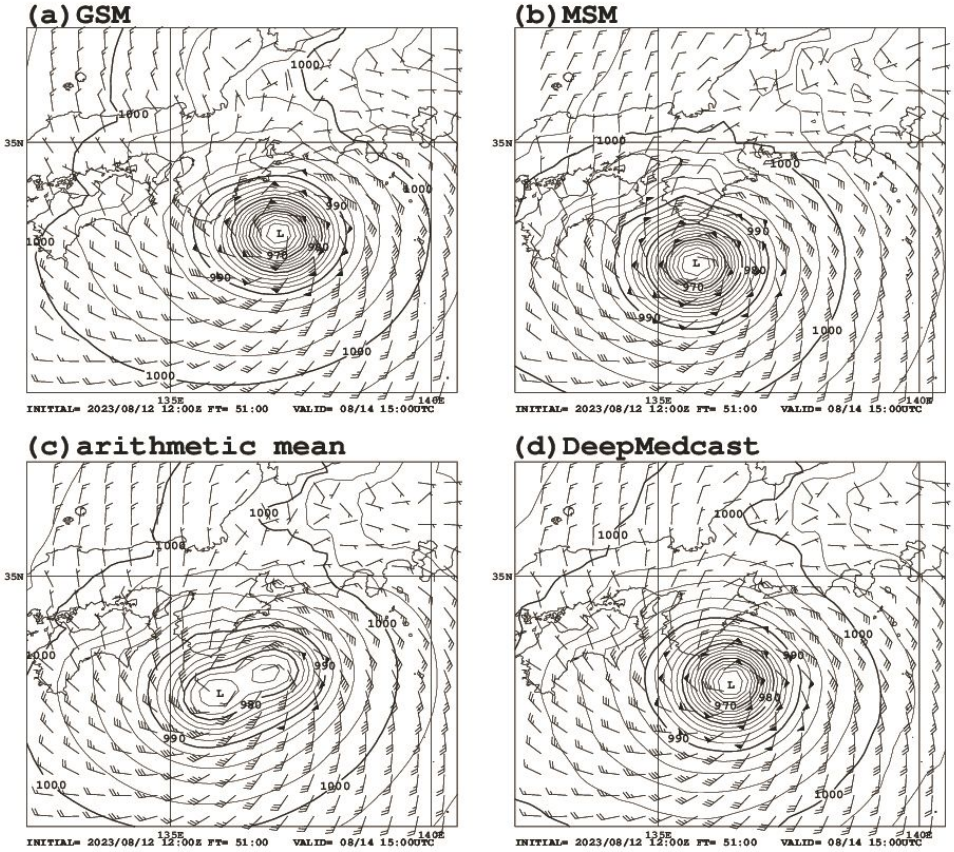
745 Fig. 4 The DNN architecture used in DeepMedcast. The model takes two input channels

746 and outputs a single channel. Input data is normalized using the maximum and minimum

747 values, and during inference, the same values are applied for the denormalization

748 process.

749



750

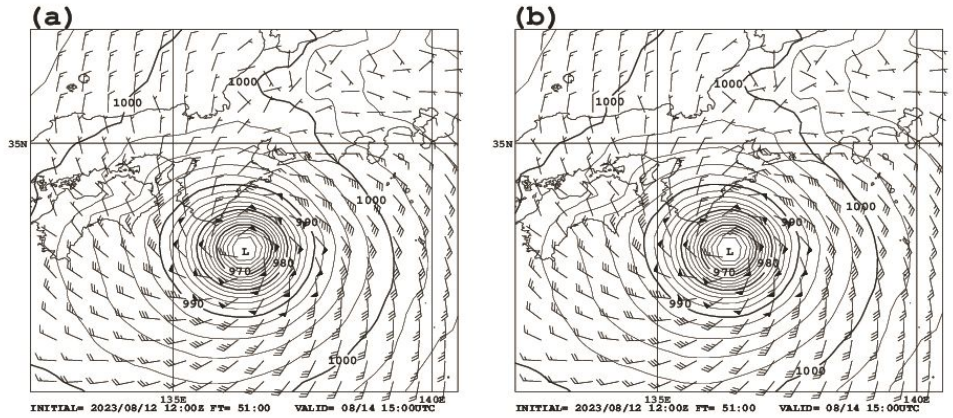
751 Fig. 5 Comparison of Typhoon LAN predictions by (a) GSM, (b) MSM, (c) the arithmetic

752 mean, and (d) DeepMedcast. The forecasts are based on the initial time of 12 UTC on

753 12 August 2023 with a forecast lead time of 51 hours. The black contours indicate mean

754 sea-level pressure and wind barbs (units in kt) show surface winds.

755



756

Fig. 6 Comparison of DeepMedcast outputs with different input orders for the case in Fig. 5. (a) Result when GSM and MSM are provided in that order (same as Fig. 5d). (b) Result when the input order is reversed (MSM-GSM). While slight differences are present due to network asymmetry, the outputs remain qualitatively identical.

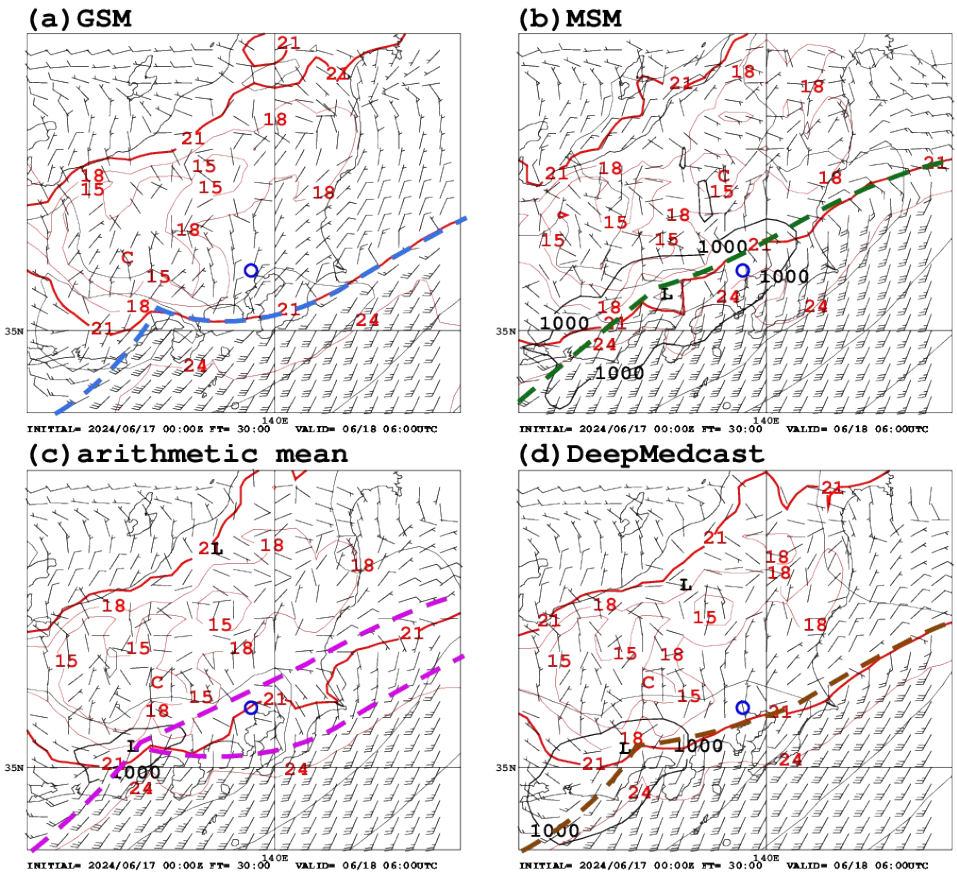


Fig. 7 Comparison of predicted front positions by (a) GSM, (b) MSM, (c) the arithmetic mean, and (d) DeepMedcast. The forecasts are based on the initial time of 00 UTC on 17 June 2024 with a forecast lead time of 30 hours. The black contours indicate mean sea-level pressure, the red contours represent surface temperature, and wind barbs (units in kt) show surface winds. The blue, green, purple, and brown dashed lines

represent the predicted front positions by GSM, MSM, the arithmetic mean, and DeepMedcast, respectively. Blue circles indicate the location of Tokyo.

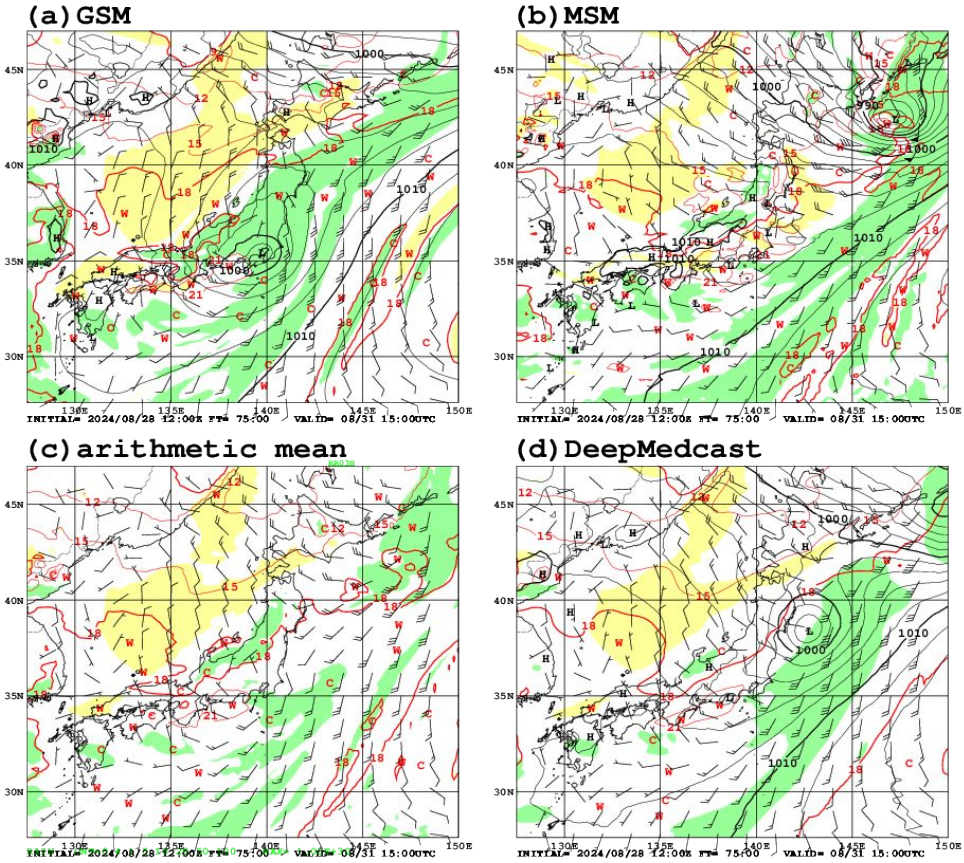


Fig. 8 Comparison of predicted low-pressure systems by (a) GSM, (b) MSM, (c) the arithmetic mean, and (d) DeepMedcast. The forecasts are based on the initial time of 12 UTC on 28 August 2024 with a forecast lead time of 75 hours. The black contours indicate mean sea-level pressure, the wind barbs (units in kt) represent surface winds, the red contours show 850 hPa temperature, and the shaded regions in green and yellow highlight areas where the dew-point depression at 850 hPa is below 3°C and above 15°C, respectively.

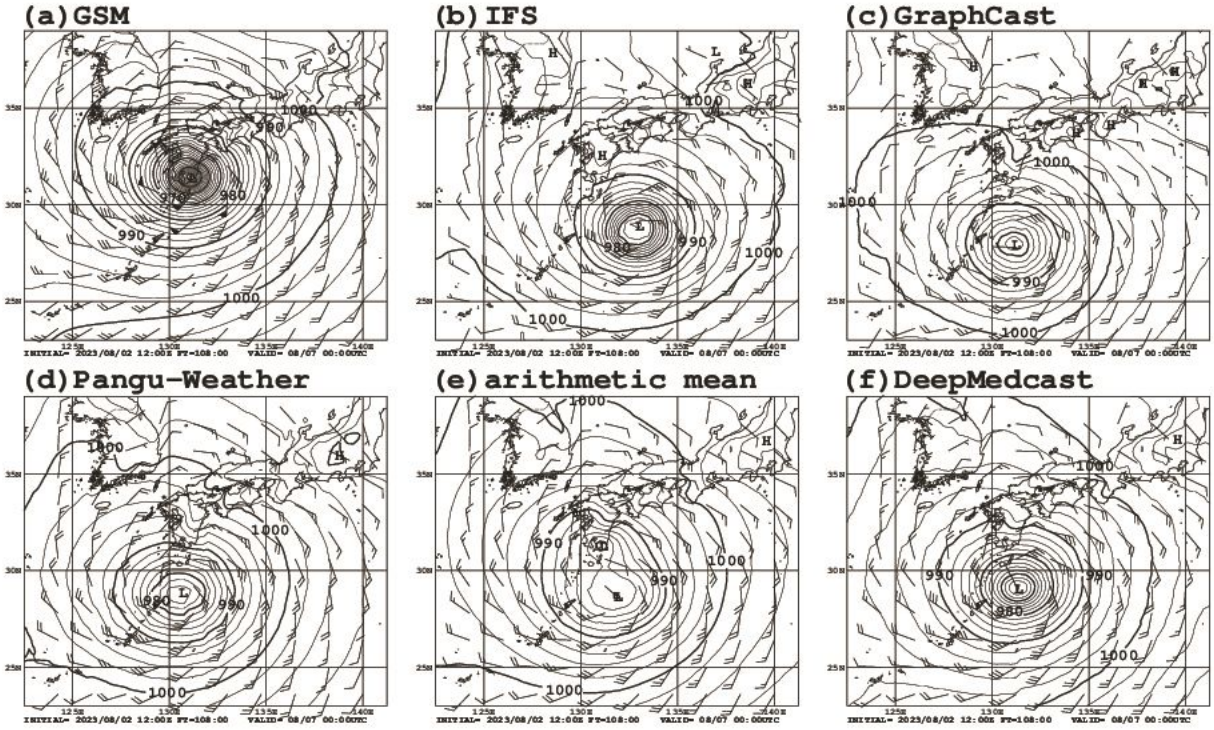
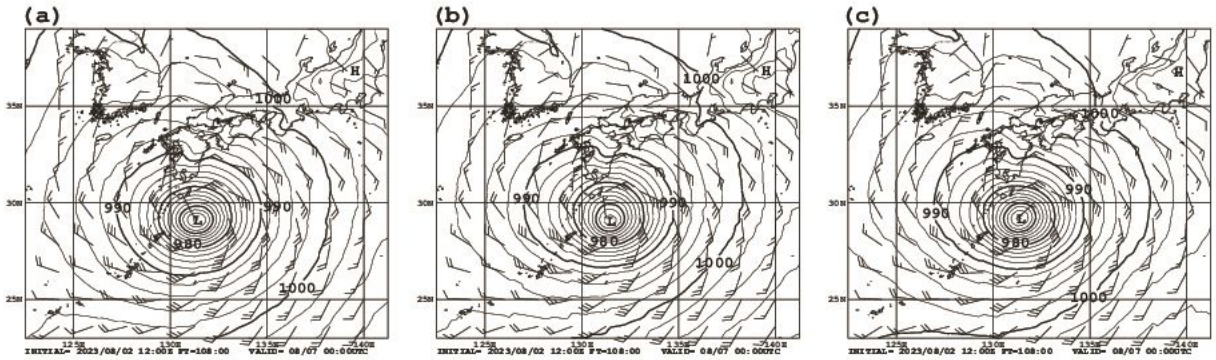
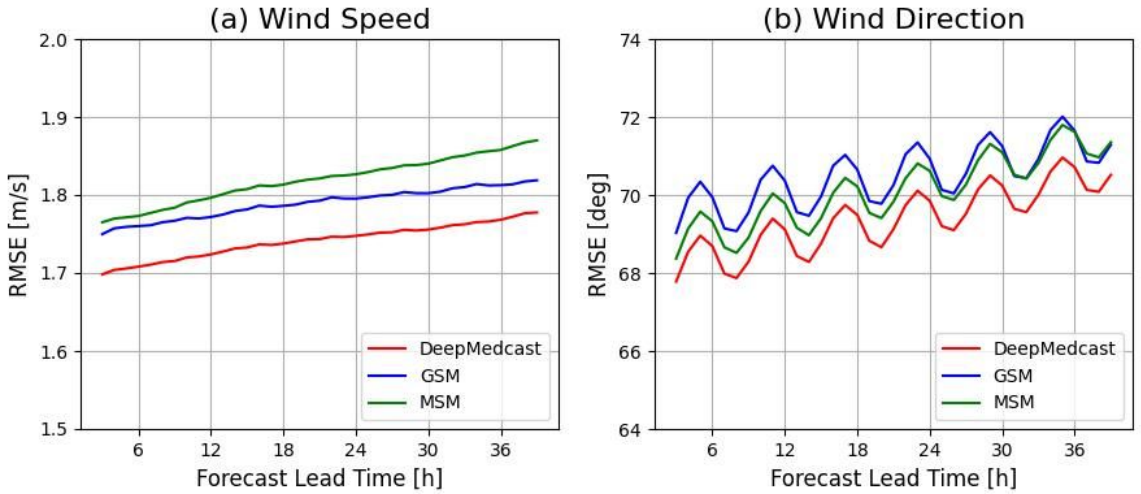


Fig. 9 Comparison of Typhoon KHANUM predictions by (a) GSM, (b) IFS, (c) GraphCast, (d) Pangu-Weather, (e) the arithmetic mean, and (f) DeepMedcast. The predictions are based on the initial time of 12 UTC on 2 August 2023 with a forecast lead time of 108 hours. The black contours indicate mean sea-level pressure and wind barbs (units in kt) show surface winds.



789 Fig. 10 DeepMedcast forecasts based on the initial time of 12 UTC on 2 August 2023
790 with a forecast lead time of 108 hours for Typhoon KHANUN using different orders of
791 intermediate forecast generation from the four NWP models: GSM, IFS, GraphCast, and
792 Pangu-Weather. (a) Intermediate forecasts are first taken between GSM and IFS, and
793 between GraphCast and Pangu-Weather, then combined. (b) GSM–GraphCast and
794 IFS–Pangu-Weather. (c) GSM–Pangu-Weather and IFS–GraphCast.

795



796

797 Fig. 11 Root mean square error (RMSE) of (a) surface wind speed and (b) surface wind
798 direction for DeepMedcast (red), input GSM (blue), and input MSM (green) forecasts,
799 verified against AMeDAS observations. The verification period spans one year, from
800 January to December 2023, with forecasts initialized at 00, 06, 12, and 18 UTC.

801

802 List of Tables

803 Table 1 Comparison of Typhoon KHANUN predictions from four NWP models, their
804 arithmetic mean, and DeepMedcast. The table presents the predicted central position,

805 central pressure, and maximum wind speed based on the initial time of 12 UTC on 2
806 August 2023 with a forecast lead time of 108 hours.
807
808 Table 1 Comparison of Typhoon KHANUN predictions from four NWP models, their
809 arithmetic mean, and DeepMedcast. The table presents the predicted central position,
810 central pressure, and maximum wind speed based on the initial time of 12 UTC on 2
811 August 2023 with a forecast lead time of 108 hours.

	Model	Central position	Central pressure	Maximum wind speed
(a)	GSM	31.4°N, 131.1°E	938 hPa	68 kt
(b)	IFS	28.9°N, 133.0°E	966 hPa	51 kt
(c)	GraphCast	28.0°N, 131.2°E	977 hPa	36 kt
(d)	Pangu-Weather	28.8°N, 130.7°E	975 hPa	39 kt
(e)	arithmetic mean	---	979 hPa	36 kt
(f)	DeepMedcast	29.3°N, 131.4°E	964 hPa	42 kt

812
813