

# Forecasting the eddying ocean with a deep neural network

Received: 7 September 2024

Accepted: 20 February 2025

Published online: 06 March 2025



Yingzhe Cui<sup>1,2,6</sup>, Ruohan Wu<sup>3,6</sup>, Xiang Zhang<sup>1,2</sup>, Ziqi Zhu<sup>3</sup>, Bo Liu<sup>3</sup>, Jun Shi<sup>3</sup>, Junshi Chen<sup>2,3,4</sup>, Hailong Liu<sup>2</sup>, Shenghui Zhou<sup>2</sup>, Liang Su<sup>5</sup>, Zhao Jing<sup>1,2</sup>✉, Hong An<sup>1,2,3,4</sup>✉ & Lixin Wu<sup>1,2</sup>✉

Mesoscale eddies with horizontal scales from tens to hundreds of kilometers are ubiquitous in the upper ocean, dominating the ocean variability from daily to weekly time scales. Their turbulent nature causes great scientific challenges and computational burdens in accurately forecasting the short-term evolution of the ocean states based on conventional physics-driven numerical models. Recently, artificial intelligence (AI)-based methods have achieved competitive forecast performance and greatly increased computational efficiency in weather forecasts, compared to numerical models. Yet, their application to ocean forecasts remains challenging due to the different dynamic characteristics of the atmosphere and the ocean. Here, we develop WenHai, a data-driven eddy-resolving global ocean forecast system (GOFS), by training a deep neural network (DNN). The bulk formulae on momentum, heat, and freshwater fluxes are incorporated into the DNN to improve the representation of air-sea interactions. Ocean dynamics is exploited in the DNN architecture design to preserve ocean mesoscale eddy variability. WenHai outperforms a state-of-the-art eddy-resolving numerical GOFS and AI-based GOFS for the temperature profile, salinity profile, sea surface temperature, sea level anomaly, and near-surface current forecasts led by 1 day to at least 10 days. Our results highlight expertise-guided deep learning as a promising pathway for enhancing the global ocean forecast capacity.

The ocean is turbulent, containing eddies with a wide range of sizes. Among them, the mesoscale eddies form the major fraction of the ocean kinetic energy reservoir<sup>1,2</sup>. These mesoscale eddies redistribute heat and materials in the ocean, contributing dominantly to the short-term (from daily to weekly time scales) variations of ocean thermal-haline structures<sup>3,4</sup>, regulating the primary productivity and further marine ecosystem<sup>5,6</sup>, and acting as a major driver of extreme events like marine heatwaves<sup>7</sup>. They also interact strongly with the overlaying atmosphere, influencing air temperature, humidity, winds, cloud

fraction, and rainfall within local marine atmospheric boundary layer<sup>8,9</sup> as well as atmospheric synoptic variability<sup>10,11</sup> and large-scale circulations<sup>12</sup>. Therefore, accurate eddy-resolving ocean forecasts are not only essential for supporting marine activities and managements but also necessary to improve weather forecast accuracy.

So far, ocean forecasts rely primarily on ocean general circulation models (OGCMs) that make forecasts by numerically discretizing and integrating the governing partial differential equations of the ocean. However, constructing an accurate eddy-resolving global ocean

<sup>1</sup>Frontiers Science Center for Deep Ocean Multispheres and Earth System and Key Laboratory of Ocean dynamics/Academy of Future Ocean, Ocean University of China, Qingdao, China. <sup>2</sup>Department of Ocean Big Data and Prediction, Laoshan Laboratory, Qingdao, China. <sup>3</sup>School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. <sup>4</sup>Joint Laboratory of Advanced Computing for Transparent Oceans between Laoshan Laboratory and University of Science and Technology of China, Hefei, China. <sup>5</sup>Qingdao Guoshi Technology Group Co., Ltd, Qingdao, China. <sup>6</sup>These authors contributed equally: Yingzhe Cui, Ruohan Wu. ✉e-mail: [jingzhao@ouc.edu.cn](mailto:jingzhao@ouc.edu.cn); [han@ustc.edu.cn](mailto:han@ustc.edu.cn); [lxwu@ouc.edu.cn](mailto:lxwu@ouc.edu.cn)

forecast system (GOFS) based on the OGCMs (i.e., the numerical GOFS) remains a challenging issue both computationally and scientifically. On the one hand, resolving mesoscale eddies requires a grid size of  $O(10\text{ km})$  or even finer<sup>13,14</sup>, imposing a massive computational burden for operating global-scale OGCMs especially for implementing advanced data assimilation<sup>15</sup> and making ensemble forecast<sup>16</sup>. In fact, the eddy-resolving numerical GOFSs did not emerge until the recent decade partially due to the enlarged computational resources. On the other hand, the chaotic nature of mesoscale eddies makes their forecast very sensitive to errors in initial conditions, boundary forcings, and OGCMs. In particular, despite development over about a half-century, there are still many uncertainties in OGCMs, including numerical errors caused by discretization, uncertainties in parameterizations of unresolved processes, and insufficient representation of interactions of the ocean with other components of the Earth system<sup>17</sup>.

Artificial intelligence (AI)-based methods provide a data-driven approach for making forecasts and have been successfully applied to the global medium-range weather forecasts<sup>18–25</sup>. Their success is achieved primarily from high-quality atmospheric reanalysis training datasets and customized deep neural network (DNN) designed to capture the hidden representations of atmospheric dynamics and alleviate accumulative error when rolling out multistep autoregressive forecasts<sup>18–25</sup>. These AI-based forecast systems show competitive forecast performance yet substantially reduce the computational burden compared to their numerical counterparts. The successful application of AI-based methods in medium-range weather forecasts has been inspiring their ocean-related usage, including the OGCM emulators, the short-term ocean forecast as well as the interannual-to-decadal ocean prediction<sup>26–29</sup>, although it has been well recognized that ocean reanalysis datasets are less robust compared to their atmospheric counterparts primarily due to the sparsity of ocean observations<sup>30</sup>.

It should be noted that the existing DNN architectures developed for the medium-range weather forecast may not be suitable for the eddy-resolving ocean forecast. Air-sea interactions play an important role in the short-term ocean variability, especially in the surface mixed layer<sup>31</sup>. However, these interactions have not been explicitly incorporated into the AI-based medium-range weather forecast systems except for providing the SST as an input. Furthermore, the existing AI-based methods tend to smooth out mesoscale weather phenomena<sup>32,33</sup> and are expected to dampen ocean mesoscale eddies in a similar way. Such a blurring effect is tolerable for the medium-range weather forecast as its variability is generally dominated by synoptic processes<sup>34</sup>, but not so for short-term ocean forecast, at which time scale mesoscale eddies make a major contribution to the variability<sup>35</sup> (Supplementary Fig. S1).

In this study, we present WenHai, an AI-based GOFS for short-term eddy-resolving forecast across the global upper ocean (0–643 m). WenHai explicitly incorporates atmospheric forcings into the DNN by exploiting the bulk formulae<sup>36</sup> on air-sea fluxes. Furthermore, the design of WenHai's architecture is guided by the characteristics of mesoscale eddies to better preserve their variabilities. As demonstrated below, these features make WenHai outperform state-of-the-art numerical and AI-based GOFSs in forecasting the eddying ocean.

## Results

### An AI-based eddy-resolving GOFS

Trained on a state-of-the-art eddy-resolving ( $1/12^\circ$ ) global ocean reanalysis dataset<sup>37</sup> (See 'GLORYS reanalysis' in Methods), WenHai forecasts  $1/12^\circ$  daily averaged sea surface height (SSH) and three-dimensional temperature, salinity, and horizontal current in the upper 643 m across the global ocean in an autoregressive way. WenHai utilizes the Swin-Transformer<sup>38</sup> as its backbone. The training process is decomposed into two stages. WenHai is first pre-trained to minimize

the loss for the one-day forecast. Then we adopt a finetune technique<sup>20</sup> to minimize the accumulative loss for a sequence of autoregressive forecasts over 5 days, which improves WenHai's performance at long forecast lead times. The architecture of WenHai (Fig. 1) and its training details are elaborated in Methods (See 'WenHai model' in Methods).

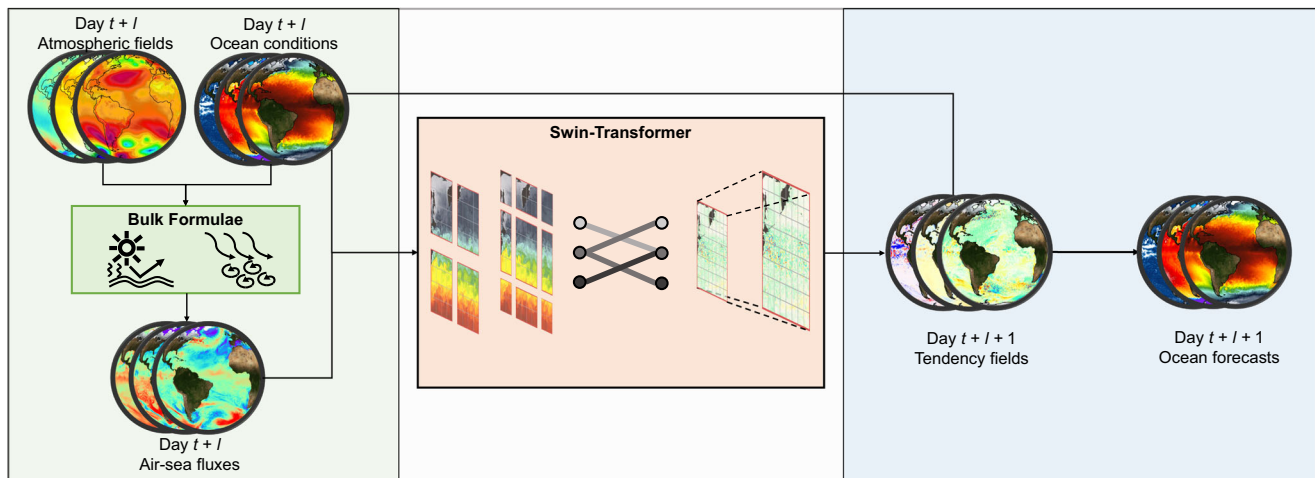
We exploit domain knowledge in the air-sea interactions and ocean dynamics to guide WenHai's architecture design, which enhances its capacity to forecast the eddying ocean. First, to explicitly represent the atmospheric forcings, we implement a specialized block for computing air-sea momentum, heat and freshwater fluxes from the surface atmosphere variables (e.g., air temperature, winds, etc.) based on the bulk formulae (Supplementary Note 1 and Supplementary Fig. S2). Second, the forecast output is chosen as the temporal tendency of an ocean variable between the two consecutive days rather than the ocean variable on the following day. This is essential to preserve the mesoscale eddy variabilities, as mesoscale eddies dominate the day-to-day variations not the absolute values of ocean variables (Supplementary Fig. S3). Similarly, we put more weight on the loss function in the upper part of the ocean, as mesoscale eddies are generally near-surface intensified<sup>39</sup>. Finally, the land region is masked to make WenHai focus on forecasting the ocean variability.

Supplementary Fig. S4 provides a visualization of sea surface kinetic energy (KE), sea surface height (SSH), and sea surface temperature (SST) forecast in the Kuroshio Extension region by WenHai. Here, the GLORYS reanalysis is approximated as the ground truth to test the capacity of WenHai to forecast the eddying ocean, with caution about the fidelity of the GLORYS reanalysis in representing reality. WenHai does capture the temporal evolution of KE, SSH, and SST reasonably well, outperforming the persistent forecast that assumes the future ocean state would be the same as the initial condition. In the following sections, we will provide more systematic and quantified assessments of the forecast performance of WenHai and compare it with a state-of-the-art numerical GOFS and AI-based GOFS.

### Outperformance of WenHai against a state-of-the-art numerical GOFS

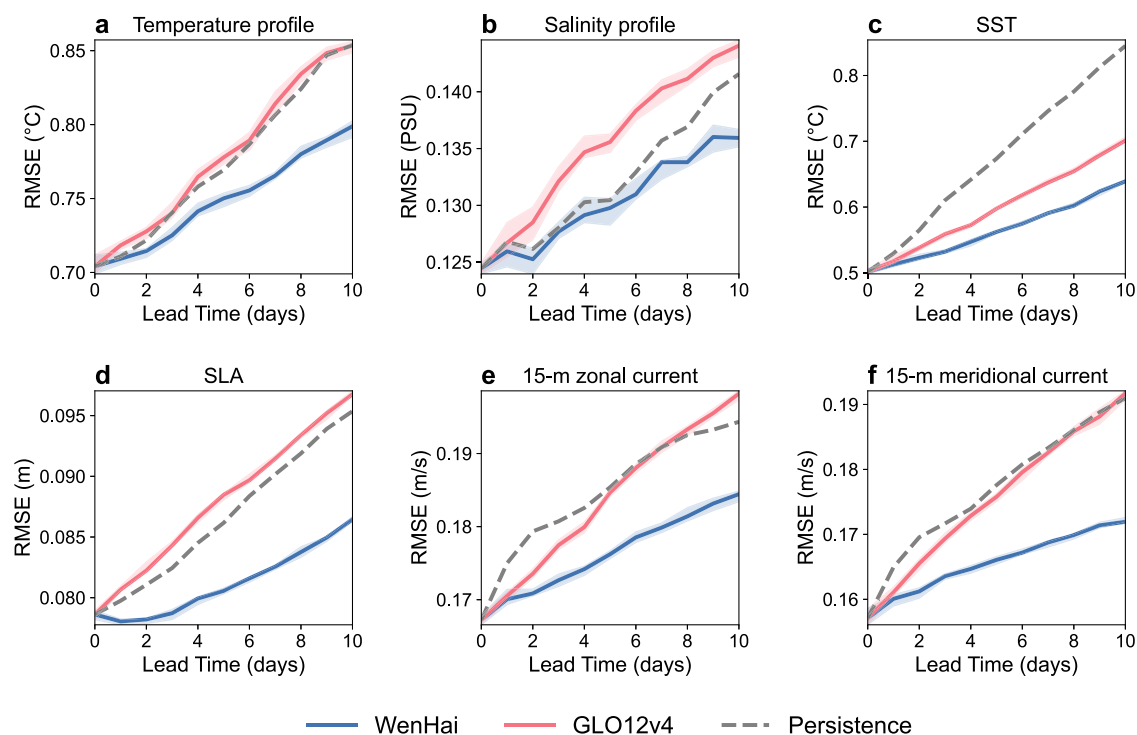
To evaluate the forecast performance of WenHai, we compare it with a state-of-the-art eddy-resolving numerical GOFS, i.e., the  $1/12^\circ$  operational ocean analysis and forecasting systems (GLO12v4<sup>40,41</sup>) from the France Mercator Océan (See 'GLO12v4' in Methods). WenHai is initialized from the same initial condition and forced by the same atmospheric forecast product as GLO12v4 during April–November 2024. We remark that using the same initial condition is essential to make a fair comparison between different forecast systems, as errors in the initial condition could have substantial influences on the forecast performance (Supplementary Note 2 and Supplementary Fig. S5). The ocean variables involved in the comparison are selected following the GODAE Ocean-View Inter-comparison and Validation Task Team (IV-TT) Class 4 framework<sup>42</sup>, including the SST and 15-m current measured by drifting buoys, temperature and salinity profiles measured by Argo profiling floats and level-3 along-track sea level anomaly (SLA) measured by satellite altimeters (See 'Observational datasets' in Methods).

Two metrics are used to quantify the forecast performance of WenHai and GLO12v4 (See 'Verification metrics' in Methods). The first is the root mean square error (RMSE), a conventional point-to-point verification metric adopted by the GODAE Ocean-View IV-TT Class 4 framework. However, it has been recognized that verification on a point-to-point basis is less appropriate for assessing the performance of high-resolution (eddy-resolving) forecast systems because of the double-penalty issue<sup>43</sup>, i.e., features correctly forecast but misplaced in space are penalized twice: once for not occurring at the observational site, and secondly for occurring at the forecast site, where they are not observed. To overcome this problem, we adopt a neighborhood-based verification metric that has been routinely applied to evaluating high-resolution oceanic and atmospheric forecast systems<sup>44,45</sup>. Specifically,



**Fig. 1 | A schematic of WenHai's architecture.** The surface atmosphere and ocean variables on the day  $t + l$  are first combined to compute the air-sea momentum, heat, and freshwater fluxes based on the bulk formulae (green block), where  $t$  is an arbitrary date index and  $l$  is the forecast lead time (indexing in 1-day intervals). Then, the ocean variables and air-sea fluxes are sent to a deep neural network (red

block) that forecasts temporal tendency between ocean variables on the days  $t + l$  and  $t + l + 1$ . The tendency field is added to the ocean variables on the day  $t + l$  to yield the ocean variable forecast on the day  $t + l + 1$  (blue block). Finally, the above processes are iterated to generate a sequence of forecasts. Maps created with Cartopy<sup>65</sup>. Background land image provided by NASA Earth Observatory.



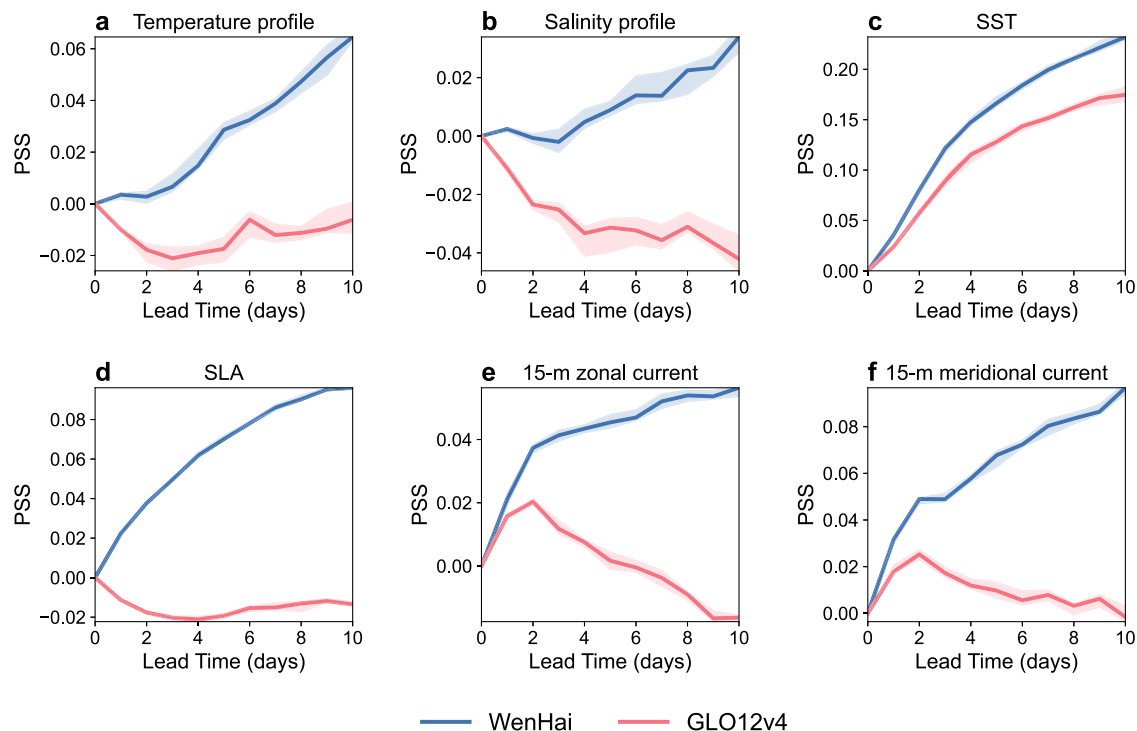
**Fig. 2 | Comparison of root mean square error (RMSE) between WenHai and a state-of-the-art numerical global ocean forecast system (GOFS) as a function of forecast lead time.** Globally averaged RMSE (the lower, the better) of the forecast temperature profile (a), salinity profile (b), sea surface temperature (SST) (c), sea level anomaly (SLA) (d), 15-m zonal current (e) and 15-m meridional current (f) as a

function of forecast lead time. The zero-lead time represents the initial conditions. The blue, red, and grey lines correspond to WenHai, GLO12v4 and persistent forecast, respectively. For the temperature and salinity profile forecast, the RMSE is vertically averaged over the upper 643 m. The shading corresponds to the 50% confidence interval computed from a bootstrap method.

forecasts within the neighborhood centered on an observational location are collected to generate a pseudo ensemble which can then be compared to the observed value using typical ensemble and probabilistic forecast verification metrics such as the continuous ranked probability score (CRPS)<sup>46</sup>. Here the neighborhood is chosen as  $1^\circ \times 1^\circ$ . On the one hand, this neighborhood is large enough to contain many ensemble members for robust estimates of the probability. On the other hand, it is sufficiently small so that the observation is not only

representative of its precise location but also has characteristics of the entire neighborhood.

**Comparison based on RMSE.** WenHai achieves lower RMSE for all the ocean variables compared to GLO12v4 (Fig. 2; Supplementary Fig. S6a, b). The superiority of WenHai over GLO12v4 becomes more evident as the forecast lead time increases. When forecasting 10 days in advance, WenHai has an RMSE for SST, SLA, 15-m zonal current,



**Fig. 3 | Comparison of persistence forecast skill (PSS) between WenHai and a state-of-the-art numerical global ocean forecast system (GOFS) as a function of forecast lead time.** Globally averaged PSS (the higher, the better) for the forecast temperature profile (a), salinity profile (b), sea surface temperature (SST) (c), sea level anomaly (SLA) (d), 15-m zonal current (e) and 15-m meridional current (f) as a

function of forecast lead time. The zero-lead time represents the initial conditions. The blue and red lines correspond to WenHai and GLO12v4, respectively. For the temperature and salinity profile forecast, the PSS is vertically averaged over the upper 643 m. The shading corresponds to the 50% confidence interval computed from a bootstrap method.

and 15-m meridional current 8.94%, 10.67%, 6.95% and 10.33% lower than GLO12v4. As to the 10-day-led temperature profile forecast, the RMSE of WenHai is lower than that of GLO12v4 throughout the upper 643 m (Supplementary Fig. S6a). So is the case for the salinity profile forecast (Supplementary Fig. S6b). The vertical mean RMSE for the temperature (salinity) profile forecast by WenHai is 6.02% (5.64%) lower than that by GLO12v4. We further assess the RMSE in different regions (Supplementary Fig. S7). The forecast performances of WenHai and GLO12v4 are region-dependent, with their RMSE varying from place to another. Nevertheless, the RMSE of WenHai is lower than that of GLO12v4 for most of the regions. Finally, the RMSE of WenHai is universally lower than that of GLO12v4 regardless of the time periods (Supplementary Fig. S8). This provides some support that WenHai should likely outperform GLO12v4 in boreal winter during which period their comparison is not conducted in this study due to the data limitation.

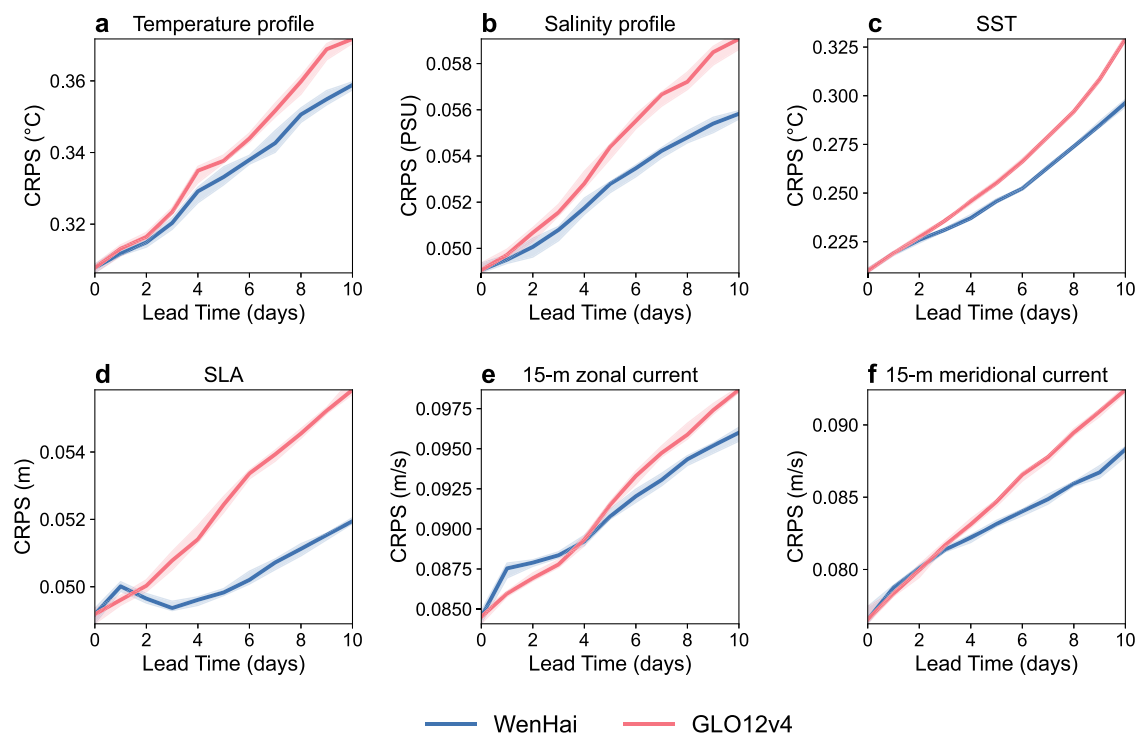
The RMSE quantifies the closeness of the forecast to the observation but doesn't by itself indicate the skill in the forecast. The usefulness of a forecast is usually measured by comparing its RMSE to that of some readily available reference forecasts. Here we chose the reference forecast as the persistent forecast. Accordingly, the persistence skill score (PSS) (See 'Verification metrics' in Methods) is used to measure the forecast skills of WenHai and GLO12v4 relative to the persistent forecast. A positive PSS indicates superiority over the persistent forecast, while a negative PSS indicates the opposite. GLO12v4 does not beat the persistent forecast except for the SST, whereas WenHai is superior to the persistent forecast for all the ocean variables at most of the forecast lead times (Figs. 2 and 3). In particular, the PSS of WenHai shows a positive trend for all the ocean variables as the forecast lead time increases, suggesting that WenHai can capture the temporal variations of the eddying ocean reasonably well. In contrast, the PSS of GLO12v4 shows a positive trend only for the SST, yet its slope is smaller than that of WenHai.

**Comparison based on CRPS.** Verification based on the CRPS yields the same conclusion as verification based on the RMSE, further suggesting the outperformance of WenHai against GLO12v4 (Fig. 4; Supplementary Fig. S6c, d). The CRPS of WenHai is lower than that of GLO12v4 for all the ocean variables and their difference is generally enlarged with the increasing forecast lead time. For the forecasts led by 10 days, the CRPS of WenHai is 3.5% lower than that of GLO12v4 for the temperature profile, 5.5% lower for the salinity profile, 10.0% lower for the SST, 7.0% lower for the SLA, and 2.7% (4.5%) lower for the 15-m zonal (meridional) current, respectively. We remark that the lower CRPS of WenHai than GLO12v4 is not sensitive to the choice of neighborhood. Varying the area of the neighborhood from  $0.3^\circ \times 0.3^\circ$  to  $2^\circ \times 2^\circ$  leads to qualitatively the same conclusions (Supplementary Fig. S9).

### Outperformance of WenHai against the latest AI-based GOFS

In this subsection, we compare WenHai with the latest AI-based GOFS, i.e., XiHe<sup>28</sup>. Both WenHai and XiHe are trained in the GLORYS and ERA reanalysis and aimed to provide a global  $1/12^\circ$  ocean forecast in the upper 643 m, so that a fair comparison between their forecast performance can be made. We note that although WenHai preserves the ocean mesoscale eddy variabilities reasonably well, XiHe leads to severe underestimation (See 'Computation of mesoscale eddy variabilities' in Methods; Table 1; Supplementary Fig. S10). For instance, the variance of mesoscale temperature anomalies forecast by WenHai differs from that in the GLORYS reanalysis by 6.1% (Table 1). In contrast, the variance of mesoscale temperature anomalies forecast by XiHe is 30.0% lower than that in the GLORYS reanalysis (Table 1). Similar conclusions also hold for the variance of mesoscale salinity anomalies and eddy kinetic energy. The strong damping of mesoscale eddy variabilities in XiHe is likely to result from the blurring effect intrinsic to the AI-based forecast<sup>20,25,32,33</sup>. Such a blurring effect is largely alleviated in WenHai due to its customized design to preserve mesoscale eddy variabilities.





**Fig. 4 | Comparison of the continuous ranked probability score (CRPS) between WenHai and a state-of-the-art numerical global ocean forecast system (GOFS) as a function of forecast lead time.** Globally averaged CRPS (the lower, the better) of the forecast temperature profile (a), salinity profile (b), sea surface temperature (SST) (c), sea level anomaly (SLA) (d), 15-m zonal current (e) and 15-m meridional

current (f) as a function of forecast lead time. The zero-lead time represents the initial conditions. The blue and red lines correspond to WenHai and GLO12v4, respectively. For the temperature and salinity profile forecast, the CRPS is vertically averaged over the upper 643 m. The shading corresponds to the 50% confidence interval computed from a bootstrap method.

The above analysis suggests that WenHai has a higher effective resolution than XiHe. Accordingly, point-to-point verification metrics such as the RMSE are not suitable for comparing their forecast performance as these metrics tend to make higher-resolution forecast systems verify worse superficially, although they may provide more realistic forecasts<sup>44,47</sup> (Supplementary Note 3 and Supplementary Fig. S11, 12). This issue is confirmed by the systematically reduced RMSE of WenHai by using the spatially smoothed initial condition to make forecast (Supplementary Fig. S11). For this reason, we only use the CRPS to quantify the forecast performance of WenHai and XiHe.

WenHai has much lower CRPS than XiHe for all the ocean variables at all the forecast lead times (Fig. 5). In terms of the temperature profile, salinity profile, SST, SLA, 15-m zonal current and 15-m meridional current forecast, the CRPS in WenHai is 6.8%–10.8%, 4.3%–10.0%, 24.5%–31.4%, 1.8%–21.0%, 1.4%–3.8%, and 0.02%–2.4% lower than that in XiHe, respectively, depending on the forecast lead times. In addition,

the superiority of WenHai over XiHe is more evident in the upper 100 m, suggesting the advantage of explicit incorporation of air-sea fluxes into the DNN (Supplementary Fig. S13). Again, the lower CRPS in WenHai than XiHe holds for various choices of the area of neighborhood (Supplementary Fig. S9). In summary, WenHai has better skills in forecasting the eddying ocean than XiHe.

## Discussion

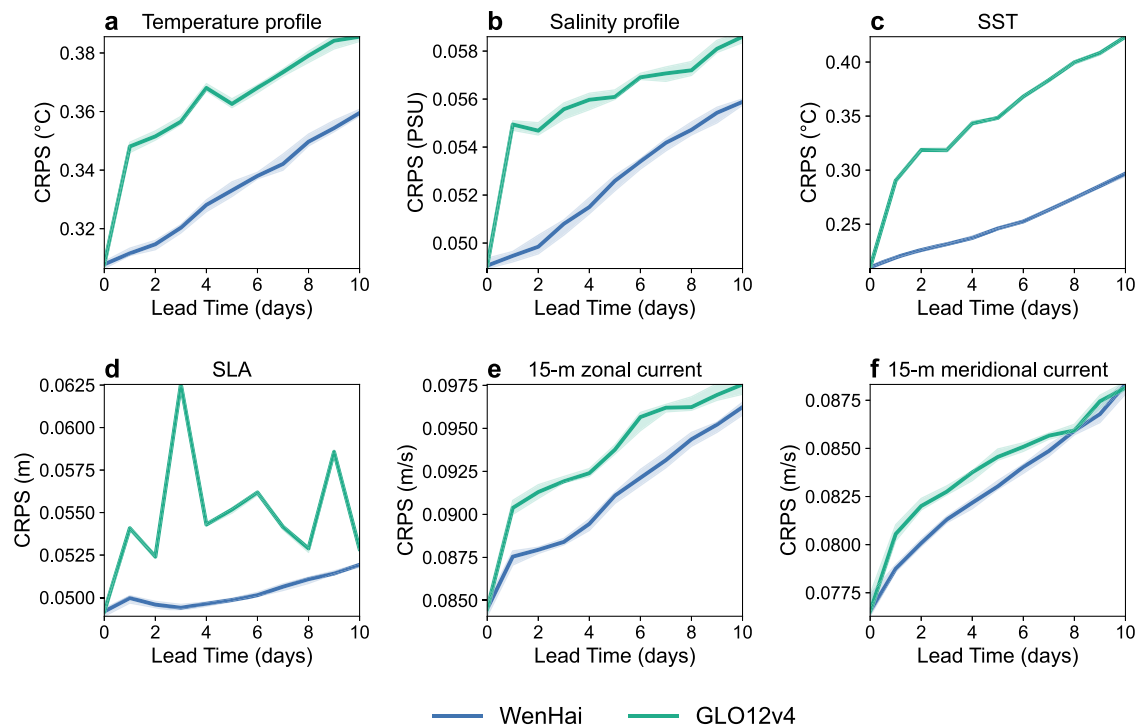
In this study, we present WenHai, an AI-based eddy-resolving GOFS whose design is guided by the domain knowledge in the air-sea interactions and ocean dynamics. It has a better representation of atmospheric forcing effects on the ocean by incorporating the bulk formulae of the air-sea fluxes into the DNN and shows a much-improved capacity to preserve ocean mesoscale variability compared to the latest AI-based GOFS through customized model architecture design guided by characteristics of mesoscale eddies. When initialized from the same initial condition and forced by the same atmospheric forecast product, WenHai surpasses the state-of-the-art eddy-resolving numerical GOFS for forecasting all the ocean variables led by one day to at least ten days. Nevertheless, it does not mean that AI-based GOFSs can substitute numerical GOFSs. In fact, the AI-based GOFSs, including WenHai, are trained based on high-quality ocean reanalysis datasets produced via numerical GOFSs in combination with data assimilation<sup>37,48</sup>. It is more appropriate to think that AI-based GOFSs stand on the shoulders of numerical GOFSs.

Despite the good forecast performance of WenHai, it has some limitations. First, WenHai provides daily average forecast that filters out the diurnal cycle and cannot resolve inertial oscillations or internal waves<sup>49</sup> over most parts of the global ocean. The daily average forecast may also be deficient in representing upwelling/downwelling on the continental shelf and coastal trapped waves. Second, the maximum depth of WenHai is chosen as 643 m that is not a genuine lower boundary of the ocean. This may degrade the forecast performance

**Table 1 | Preservation of ocean mesoscale variabilities by artificial intelligence (AI)-based global ocean forecast systems (GOFSs)**

	WenHai	XiHe	GLORYS
eddy kinetic energy $10^{15} \text{ m}^5 \text{ s}^{-2}$	2.21	1.96	2.39
temperature variance $10^{16} \text{ }^{\circ}\text{C}^2 \text{ m}^3$	8.43	6.29	8.98
salinity variance $10^{15} \text{ PSU}^2 \text{ m}^3$	3.38	2.35	3.54

Globally integrated eddy kinetic energy, variance of mesoscale temperature anomaly, and variance of mesoscale salinity anomaly in the upper 643 m averaged over the forecast lead times from 1 day to 10 days. Here the forecasts are initialized from the GLORYS reanalysis and forced by the ERA5 reanalysis during 2020. The ocean mesoscale eddy variabilities in the GLORYS reanalysis are treated as the ground truth.



**Fig. 5 | Comparison of the continuous ranked probability score (CRPS) between WenHai and the latest AI-based global ocean forecast system (GLO12v4) as a function of forecast lead time.** Globally averaged CRPS (the lower, the better) of the forecast temperature profile (a), salinity profile (b), sea surface temperature (SST) (c), sea level anomaly (SLA) (d), 15-m zonal current (e) and 15-m meridional

current (f) as a function of forecast lead time. The zero-lead time represents the initial conditions. The blue and red lines correspond to WenHai and XiHe, respectively. For the temperature and salinity profile forecast, the CRPS is vertically averaged over the upper 643 m. The shading corresponds to the 50% confidence interval computed from a bootstrap method.

near 643 m. Third, WenHai does not take into consideration the effects of river runoff and sea ice on the ocean and may have some deficiencies in coastal and polar regions. Finally, there is still a slight yet noticeable damping of mesoscale eddy variabilities in WenHai (Table 1) despite its customized design to alleviate this problem. This is partially because WenHai adopts a deterministic pointwise loss function that nudges WenHai to provide smoother forecasts at longer forecast lead times due to the double penalty issue. We envision that WenHai can be further improved by training a hierarchy of DNNs representing oceanic processes of different spatio-temporal scales, adopting probabilistic loss function or adding physical conservation constraints to the loss function, using hybrid model architectures combining deep learning and numerical solver, and utilizing higher-quality ocean reanalysis datasets for training. Achieving these improvements requires the expertise in the marine science and more oceanographers to get involved in the development of AI-based GOFSSs.

## Methods

### GLORYS reanalysis

The Copernicus Global 1/12° Oceanic and Sea Ice GLORYS12 Reanalysis (GLORYS reanalysis for short), developed by Mercator Océan, provides a realistic representation of key oceanic quantities such as sea level, water mass properties, mesoscale activity or sea ice extent. The ocean and sea ice components are based on the NEMO platform<sup>50</sup>. It has a quasi-isotropic horizontal grid with a 1/12° resolution and 50 vertical levels with vertical spacing increasing progressively with depth. The ocean model is driven by the ERA-Interim atmospheric reanalysis<sup>51</sup> provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Observations are assimilated using a reduced-order Kalman filter derived from a singular evolutive extended Kalman (SEEK) filter<sup>52</sup> with a three-dimensional multivariate background error covariance matrix and a 7-day assimilation cycle<sup>53</sup>. Assimilated observations include the satellite-based SLA, SST, and sea ice concentration

and in situ temperature and salinity profiles. The reanalysis dataset covers 1993–2020 with daily resolution.

It should be noted that ocean reanalysis datasets including the GLORYS reanalysis are less robust compared to their atmospheric counterparts primarily due to the sparsity of ocean observations<sup>30</sup>. In fact, some deficiencies of the GLORYS reanalysis have been reported<sup>37</sup>. First, a few zonal currents such as the Antarctic circumpolar current and Western Pacific south equatorial current are overly strong. In addition, inter-basins volume exchanges are larger compared to observations and other reanalyses. Finally, there is also unexpected behavior in the Tropical Indian and North East Atlantic basins.

### WenHai model

**Training and validation datasets.** WenHai is trained on the GLORYS reanalysis and ERA5 reanalysis<sup>54</sup> during 1993–2018 and validated based on the datasets during 2019.

**Input and output details.** WenHai is aimed to forecast a sequence of tendency of daily mean ocean variables in the upper ocean  $\{\Delta \mathbf{O}_t^l = \mathbf{O}_t^l - \mathbf{O}_{t-1}^l, l \in [1, N]\}$  in an autoregressive way, given the initial condition  $\mathbf{O}_0^l$  and the surface atmospheric variables  $\{\mathbf{A}_t^l, l \in [0, N-1]\}$ , where  $t$  is an arbitrary date index and  $l$  is the forecast lead time (indexing in 1-day intervals). We use bold characters here to emphasize that  $\mathbf{O}_t^l$  and  $\mathbf{A}_t^l$  are tensors consisting of multiple variables at multiple grid points. The ocean variables forecast by WenHai are the same as the numerical GOFSS, including zonal velocity, meridional velocity, temperature, salinity, and SSH. WenHai shares the same horizontal grid points as the GLORYS reanalysis and has 23 depth levels selected from those of the GLORYS reanalysis in the upper 643 m (Supplementary Fig. 14). The different ocean variables are stacked along the vertical direction. Accordingly,  $\mathbf{O}_t^l$  has a dimension size of  $N_O \times N_{lat} \times N_{lon}$  with  $N_O = 4 \times 23 + 1 = 93$ ,  $N_{lat} = 2041$  and  $N_{lon} = 4320$ .

Here  $N_O$ ,  $N_{lat}$  and  $N_{lon}$  are analogous to the channel number, height and width of an image in the field of computer vision.

The surface atmosphere variables are 10-m zonal wind, 10-m meridional wind, mean sea level pressure, 2-m temperature, and 2-m dewpoint temperature, which are routinely forecast by numerical weather forecast systems<sup>55</sup>. It should be noted that the ocean is not directly forced by these surface atmosphere variables. Instead, it is directly forced by the net momentum, heat, and freshwater fluxes at the air-sea interface. For this reason, we implement a block to compute the fluxes based on the surface values of  $\mathbf{O}_t^l$  and  $\mathbf{A}_t^l$  using bulk formulae. Specifically, the surface atmosphere variables are first horizontally interpolated to the oceanic grids, and then the bulk formulae are used to transfer the surface atmosphere and ocean variables into the air-sea fluxes, including the surface latent heat flux, surface sensible heat flux, surface upward long-wave radiation flux, zonal wind stress, meridional wind stress, and evaporation rate. Finally, the surface net short-wave radiation flux, surface downward long-wave radiation flux, and precipitation rate from the numerical weather forecasts are included, and the surface downward and upward long-wave radiation fluxes are combined to yield the surface net long-wave radiation flux. This results in a total of eight air-sea flux variables denoted as  $\mathbf{F}_t^l$  whose dimension size is  $N_F \times N_{lat} \times N_{lon}$  with  $N_F = 8$ . The  $\mathbf{F}_t^l$  and  $\mathbf{O}_t^l$  are fed into the DNN to forecast  $\Delta\mathbf{O}_t^l$ . Once  $\Delta\mathbf{O}_t^l$  is obtained, it is added on  $\mathbf{O}_t^l$  to obtain  $\mathbf{O}_t^{l+1}$ . Then  $\mathbf{F}_t^{l+1}$  is computed based on the surface values of  $\mathbf{O}_t^{l+1}$  and  $\mathbf{A}_t^{l+1}$ . The above processes are repeated to make a sequence of forecasts.

**Architecture overview.** For each channel, the values of  $\mathbf{O}_t^l$  and  $\mathbf{F}_t^l$  are normalized to range from 0 to 1. The normalized variables first go through two independent pre-processing modules designed to reduce dimensionality. The ocean and atmosphere outputs of pre-processing modules with a dimension of  $259 \times 546 \times H$  are then reshaped to  $141414 \times H$  and added together before being fed into 10 homogeneous Swin-Transformer layers whose hidden dimension size  $H$  is 768 and window size is 7. Following this, a post-processing module is employed to restore the outputs of Swin-Transformer layers to the original dimension of ocean variables.

**Pre- and post-processing.** The pre-processing module consists of a patch embedding block and a down-sampling block. The patch embedding block is a convolution layer whose kernel size and stride (also known as the patch size) is chosen as 4 in this study. As  $N_{lat}$  and  $N_{lon}$  need to be divisible by both the window size 7 and patch size 4 to feed  $\mathbf{O}_t^l$  and  $\mathbf{F}_t^l$  into the patch embedding block, a zero-value padding is applied to change  $N_{lat} \times N_{lon}$  of  $\mathbf{O}_t^l$  and  $\mathbf{F}_t^l$  to  $2072 \times 4368$ . The output of the patch embedding block is the embedded  $\mathbf{O}_t^l$  and  $\mathbf{F}_t^l$  with the same dimension of  $H \times 518 \times 1092$  (i.e.,  $H \times 2072/4 \times 4368/4$ ). It should be noted that using a larger patch size leads to more reduction of the dimensionality of  $\mathbf{O}_t^l$  and potentially makes WenHai more difficult to preserve the mesoscale variabilities, which is confirmed by the stronger damping of EKE with the increasing patch size (Supplementary Fig. S15). The down-block uses a convolution layer with kernel size 3 and stride 2 to further reduce the data dimension to  $H \times 259 \times 546$ , followed by permuting the dimension to  $259 \times 546 \times H$ . The post-processing module consists of an up-sampling block and a linear layer. The up-sampling block is a 2-D transposed convolution layer with kernel size 2 and stride 2, changing the dimension of reshaped output from the Swin-Transformer layers from  $H \times 259 \times 546$  to  $H \times 518 \times 1092$ , followed by permuting the dimension to  $518 \times 1092 \times H$ . The linear layer changes the data dimension from  $518 \times 1092 \times H$  to  $518 \times 1092 \times 1488$ . Then the data are rearranged to have a dimension of  $93 \times 2072 \times 4368$ . Finally, the outputs are restored to the original dimension of  $\mathbf{O}_t^l$  (i.e.,  $93 \times 2041 \times 4320$ ) by the pad-back module to remove redundant marginal zero values.

**Window-based Transformer.** After being processed by the pre-processing module, inputs are transformed into numerous patch tokens with a sequence length of 141414. As the memory consumption of the self-attention layer in the Vision Transformer (ViT) increases quadratically with the sequence length, we turn to using Swin-Transformer V2 with the window-attention mechanism. The long token sequences are divided into 2886 windows, and each window calculates self-attention with only 49 tokens, which greatly reduces the amount of computation and memory consumption. It should be noted that the down-sampling operations in Swin-Transformer V2 are not used in our model to ensure that the shapes of inputs and outputs are unchanged, making it convenient to increase the number of transformer layers.

**Land-sea mask.** A land-sea mask is adopted to handle the missing values of ocean variables on the land points. Specifically, the values on the land points are always replaced with 0 at any step of the processing. Accordingly, WenHai puts no attention on the land points, making it focus on the ocean forecast.

**Two-stage training strategy.** The training process is divided into two stages. The first stage is to pretrain a base model by optimizing a weighted mean absolute error (MAE) for  $\Delta\mathbf{O}_t^l$  integrated over the global ocean, i.e.,

$$\min \iiint \|\mathbf{w} \cdot (\Delta\mathbf{O}_t^l - \Delta\mathbf{O}_t^1)\|_1 dV \approx \min \sum_i \|\mathbf{w}(i) \cdot (\Delta\mathbf{O}_{t,i}^l - \Delta\mathbf{O}_{t,i}^1)\|_1 \delta V(i) \quad (1)$$

where  $\Delta\mathbf{O}_{t,i}^1$  represents the forecast tendency by WenHai at the  $i$ -th grid cell,  $\Delta\mathbf{O}_{t,i}^l$  is the ground truth obtained from the GLORYS reanalysis,  $\delta V(i)$  is the volume of the  $i$ -th ocean grid cell,  $\mathbf{w}(i)$  is a weight vector for different ocean variables, and  $\|\cdot\|_1$  is the L1 norm. The  $\mathbf{w}(i)$  is chosen as:

$$\mathbf{w}(i) = \frac{\mathbf{s}}{\delta h(i)} \quad (2)$$

where  $\delta h(i)$  is the layer thickness of the  $i$ -th ocean grid cell. As  $\delta h(i)$  becomes larger as the depth increases, the presence of  $\delta h(i)$  in the denominator makes WenHai put more attention on the shallower regions where mesoscale eddies are stronger<sup>56</sup>, enhancing its capacity to forecast mesoscale eddy variabilities. The value of  $\mathbf{s}$  is chosen in a way so that the tendencies of different normalized ocean variables have similar variances. We find that the variances of tendency of the normalized variables have comparable magnitudes except the tendency of normalized salinity which is much smaller than the others. Accordingly, the component of  $\mathbf{s}$  associated with salinity is set as 4.5 and components associated with the other variables as 1 to reduce the difference of variances among different normalized variables. It should be noted that the above choice of  $\mathbf{w}$  is unlikely optimal. Nevertheless, it leads to better performance on the validation dataset than choosing  $\mathbf{w}(i) = \mathbf{1}/\delta h(i)$  or  $\mathbf{w}(i) = \mathbf{s}$ .

During the second training stage, the pretrained-based model is finetuned to improve its performance at longer forecast lead times. This is done by optimizing the accumulated MAE over a sequence of autoregressive forecasts, i.e.,

$$\min \sum_{i=1}^M \sum_t \|\mathbf{w}(i) \cdot (\Delta\mathbf{O}_{t,i}^l - \Delta\mathbf{O}_{t,i}^l)\|_1 \delta V(i) \quad (3)$$

We try different values of  $M$  and find  $M = 5$  gives rise to the overall best forecast performance on the validation dataset. Applying the finetune technique reduces the forecast RMSE for all the ocean variables (Supplementary Fig. S16).

**Distributed training.** During the second training stage, iterations on high-resolution inputs with patch size 4 consume considerable GPU memory beyond 80 GB. To address this issue, the model is strategically divided into several parts to share the overhead on different devices. Specifically, the 10 Transformer layers are segmented into five stages: 1, 3, 3, 3, and 0 Transformer layers, respectively, for optimal memory load balancing. The initial stage incorporates the model's patch embedding and down-sampling block, while the final stage encompasses the up-sampling block and patch recovery module. Iterative finetune deviates from conventional pipeline model parallelism by creating a feedback loop, where the output of one iteration becomes the input for the next. The process begins by loading initial data into the first stage for forward computation. Results are then sequentially passed through subsequent stages. Upon reaching the final stage, the loss for the current iteration is calculated, and these results are transmitted back to the first device. Each iteration's output serves as the input for the following iteration. This cycle continues until the 5-day computation is complete. The final stage aggregates losses from all iterations. By employing this pipeline model parallelism, the finetuning capacity has been significantly expanded from a single day to five days.

### GLO12v4

GLO12v4 is the latest version of the Operational Mercator global ocean analysis and forecast system, providing global ocean forecasts in the next 10 days, updated daily. It includes an ocean component (NEMO3.6) with a horizontal resolution of  $1/12^\circ$  and 50 vertical levels and a sea ice component (LIM3 Multi-categories sea ice model)<sup>57</sup>. A reduced-order Kalman filter derived from a SEEK filter with a 3-D multivariate modal decomposition of the forecast error and a 7-day assimilation cycle is applied for assimilating observations including satellite-based SLA, SST and sea ice concentration and in situ temperature and salinity profiles.

The forecast is initialized from the hindcast simulation of GLO12v4 spanning the last 24 h leading to real-time (i.e., -24 to 0 h) without data assimilation. The atmosphere variables during the forecast period are obtained from the ECMWF IFS HRES forecast product<sup>58</sup>. We collect GLO12v4 forecasts along with the initial conditions and atmospheric forecast product during April, 2024–November, 2024 and use these data to assess the forecast performance.

### Observational datasets

For temperature and salinity profiles, we use quality controlled (QC) near-real-time Argo<sup>59</sup> profiles from E.U. Copernicus Marine Service Information<sup>60</sup>. For SLA, we use an along-track Level-3 near-real-time product from AVISO. The Ssalto/Duacs altimeter products were produced and distributed by the Copernicus Marine and Environment Monitoring Service (CMEMS) (<http://www.marine.copernicus.eu>). The product contains measurements of satellite altimeters including Sentinel-6A-HR, Jason-3 interleaved, Sentinel-3A, Sentinel-3B, SARAL-DP/AltiKa, Cryosat-2, HaiYang-2B, and SWOT-nadir. The filtered version of the product is chosen to suppress the effect of measurement noises. The SST and 15-m current are obtained from the near-real-time drifter observations provided by the Global Drifter Program<sup>61</sup>. The hourly drifter records are bin-averaged to produce the daily mean values. For the observations of all the ocean variables, only records passing all the real-time QC tests are retained.

It should be noted that GOFSSs forecast SSH that differs from SLA by a reference sea level<sup>62</sup>. In the observation, the reference sea level is computed as the time mean SSH during 1993–2012. For the GOFSSs, the reference sea level is computed as the time-mean SSH of the GLORYS reanalysis during the same period. There is likely a bias of SLA in GOFSSs due to the difference between the observation and GLORYS reanalysis. To alleviate this bias, SLA in GOFSSs is subtracted by the difference of SLA between the GLORYS reanalysis and observation averaged over the global ocean during 1993–2020.

### Verification metrics

**Point-to-point verification metrics.** The forecasts of GOFSSs are partitioned according to their initial time  $t$  and forecast lead time  $l$ , and compared to observations at time  $t+l$ . As the observations are distributed irregularly in space and not located on the grid points of GOFSSs, forecasts are interpolated to individual sites of observations to form point-to-point match-ups between the forecasts and observations, so that the forecast errors can be computed, i.e.,

$$\varepsilon_{t,l}^i = \hat{V}_{t,l}^i - V_{t+l}^i \quad (4)$$

where  $V_{t+l}^i$  is the observation and  $\hat{V}_{t,l}^i$  is the interpolated forecast for the  $i$ -th match-up. To make a fair comparison among GOFSSs, we only retain the match-ups for those having a forecast counterpart from every GOFSS (Note that different GOFSSs differ in their grids and land-sea masks).

Specifically, the match-ups of SST and SLA are obtained by horizontally interpolating the forecast values to the observational sites using the bilinear interpolation. The match-ups of 15-m zonal and meridional currents are obtained by horizontally interpolating the forecast values to the observational sites using the bilinear interpolation and vertically to 15 m using the linear interpolation. As to the match-ups of vertical profiles of temperature and salinity, the forecast values are horizontally interpolated to the sites of Argo profiles using the bilinear interpolation and vertically to the depths of Argo profiles using the linear interpolation. For each match-up of the observational and forecast profiles, we vertically bin-average the squared forecast errors in the prescribed 42 vertical bins containing approximately equal numbers of Argo measurements (Supplementary Fig. S17).

Two point-to-point verification metrics are used to measure the forecast performance of GOFSSs, including the root mean square error (RMSE) and persistence skill scores (PSS). The RMSE is defined as<sup>42</sup>:

$$\text{RMSE}_{t,l} = \sqrt{\frac{\sum_{i=1}^N (\varepsilon_{t,l}^i)^2}{N}} \quad (5)$$

where  $N$  is the number of total match-ups at the initial time  $t$  and forecast lead time  $l$ . For the temperature and salinity profiles,  $(\varepsilon_{t,l}^i)^2$  corresponds to the vertically bin-averaged squared forecast errors in the prescribed 42 vertical bins. In this case, the value of  $N$  differs among different vertical bins and  $\text{RMSE}_{t,l}$  is expressed as a function of center depth of these vertical bins.

The PSS is defined as<sup>42</sup>:

$$\text{PSS}_{t,l} = 1 - \frac{\text{RMSE}_{t,l}}{\text{RMSE}_{t,l}^p} \quad (6)$$

where  $\text{RMSE}_{t,l}^p$  is the RMSE of a persistent forecast.

Once  $\text{RMSE}_{t,l}$  and  $\text{PSS}_{t,l}$  are obtained, their median values among different initial time  $t$  are computed and expressed as a function of forecast lead time  $l$  (Figs. 2 and 3).

**Neighborhood-based verification metrics.** The continuous ranked probability core (CRPS) of GOFSSs evaluated based on some observation  $V_{t+l}^i$  is defined as<sup>46</sup>:

$$\text{CRPS}_{t,l}^i = \int_{-\infty}^{+\infty} [F(x) - H(x \geq V_{t+l}^i)]^2 dx \quad (7)$$

where  $H$  represents the Heaviside function and  $F(x)$  denotes the cumulative distribution function of the pseudo ensemble forecast for  $V_{t+l}^i$ . The individual members of pseudo ensemble forecast of GOFSSs are collected over the  $1^\circ \times 1^\circ$  box centered on the grid point closest to the site of  $V_{t+l}^i$ . For the temperature and salinity profiles, the collected forecast values are first interpolated to the depths of Argo profiles to



compute the CRPS at each depth, and these CRPS values are then vertically bin-averaged in the prescribed 42 vertical bins.

The globally averaged CRPS at some initial time  $t$  and forecast lead time  $l$  (denoted as  $\text{CRPS}_{t,l}^i$ ) is obtained by averaging  $\text{CRPS}_{t,l}^i$  over all the observational sites at the time  $t + l$ . Once  $\text{CRPS}_{t,l}^i$  is obtained, its median values among different initial time  $t$  are computed and expressed as a function of forecast lead time  $l$  (Figs. 4 and 5).

### Computation of mesoscale eddy variabilities

In this study, the mesoscale eddies are loosely defined as the processes with a horizontal scale ranging from tens to hundreds of kilometers, including fronts, filaments and coherent vortices<sup>1</sup>. To isolate anomalies induced by mesoscale eddies, we subtract a  $4^\circ \times 4^\circ$  spatially running mean from the original variables<sup>63</sup>. The eddy kinetic energy is computed as  $\langle \frac{1}{2}(u'^2 + v'^2) \rangle$  where  $u$  and  $v$  are zonal and meridional velocity, respectively, the primes denote the mesoscale anomalies and the brackets denote the spatio-temporal average. The variance of mesoscale temperature and salinity anomalies are computed as  $\langle T'^2 \rangle$  and  $\langle S'^2 \rangle$  where  $T$  and  $S$  denote the temperature and salinity, respectively. To evaluate the preservation of ocean mesoscale variabilities, eddy kinetic energy and variances of mesoscale temperature and salinity anomalies are integrated in the upper 643 m.

### Data availability

The GLORYS reanalysis datasets are obtained from [https://data.marine.copernicus.eu/product/GLOBAL\\_MULTIYEAR\\_PHY\\_001\\_030/description](https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/description). The ERA5 reanalysis datasets are obtained from <https://doi.org/10.24381/cds.adbb2d47>. The GLO12v4 datasets are obtained from [https://data.marine.copernicus.eu/product/GLOBAL\\_ANALYSIS\\_FORECAST\\_PHY\\_001\\_024/description](https://data.marine.copernicus.eu/product/GLOBAL_ANALYSIS_FORECAST_PHY_001_024/description). The ECMWF HRES forecasts are obtained from <https://doi.org/10.21957/open-data>. The Argo observations are obtained from [https://data.marine.copernicus.eu/product/INSITU\\_GLO\\_PHYBGCWAV\\_DISCRETE\\_MYNRT\\_013\\_030/description](https://data.marine.copernicus.eu/product/INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030/description). The drifter observations are obtained from [https://data.marine.copernicus.eu/product/INSITU\\_GLO\\_PHY\\_UV\\_DISCRETE\\_NRT\\_013\\_048/description](https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_UV_DISCRETE_NRT_013_048/description). The SLA observations are obtained from [https://data.marine.copernicus.eu/product/SEALEVEL\\_GLO\\_PHY\\_L3\\_NRT\\_008\\_044/description](https://data.marine.copernicus.eu/product/SEALEVEL_GLO_PHY_L3_NRT_008_044/description). Data used to produce the figures in the main text and the supplementary information can be found at <https://github.com/Cuiyingzhe/WenHai>.

### Code availability

The inference code and model weights of WenHai can be found at <https://github.com/Cuiyingzhe/WenHai><sup>64</sup>. The inference codes and model weights of XiHe can be found at <https://github.com/Ocean-Intelligent-Forecasting/XiHe-GlobalOceanForecasting>. The bulk formulae code can be found at <https://github.com/xgcm/aerobulk-python>. The code for CRPS calculation can be found at <https://github.com/proper scoring/proper scoring>.

### References

- Stammer, D. Global characteristics of ocean variability estimated from regional TOPEX/POSEIDON altimeter measurements. *J. Phys. Oceanogr.* **27**, 1743–1769 (1997).
- Ferrari, R. & Wunsch, C. Ocean circulation kinetic energy: reservoirs, sources, and sinks. *Annu. Rev. Fluid Mech.* **41**, 253–282 (2009).
- Smith, K. S. & Ferrari, R. The production and dissipation of compensated thermohaline variance by mesoscale stirring. *J. Phys. Oceanogr.* **39**, 2477–2501 (2009).
- He, Q. et al. Enhancing impacts of mesoscale eddies on Southern Ocean temperature variability and extremes. *Proc. Natl Acad. Sci.* **120**, e2302292120 (2023).
- Chelton, D. B., Gaube, P., Schlax, M. G., Early, J. J. & Samelson, R. M. The Influence of Nonlinear Mesoscale Eddies on Near-Surface Oceanic Chlorophyll. *Science* **334**, 328–332 (2011).
- McGillicuddy, D. J. Mechanisms of physical-biological-biogeochemical interaction at the oceanic mesoscale. *Annu. Rev. Mar. Sci.* **8**, 125–159 (2016).
- Bian, C. et al. Oceanic mesoscale eddies as crucial drivers of global marine heatwaves. *Nat. Commun.* **14**, 2970 (2023).
- Frenger, I., Gruber, N., Knutti, R. & Münnich, M. Imprint of Southern Ocean eddies on winds, clouds and rainfall. *Nat. Geosci.* **6**, 608–612 (2013).
- Yuan, M., Li, F., Ma, X. & Yang, P. Spatio-temporal variability of surface turbulent heat flux feedback for mesoscale sea surface temperature anomaly in the global ocean. *Front. Mar. Sci.* **9**, 957796 (2022).
- Ma, X. et al. Distant Influence of Kuroshio Eddies on North Pacific Weather Patterns? *Sci. Rep.* **5**, 17785 (2015).
- Ma, X. et al. Importance of resolving Kuroshio Front and eddy influence in simulating the North Pacific Storm Track. *J. Clim.* **30**, 1861–1880 (2017).
- Seo, H. et al. Ocean mesoscale and frontal-scale ocean–atmosphere interactions and influence on large-scale climate: a review. *J. Clim.* **36**, 1981–2013 (2023).
- Hurlburt, H. E. The potential for ocean prediction and the role of altimeter data. *Mar. Geod.* **8**, 17–66 (1984).
- Chassignet, E. P. et al. Impact of horizontal resolution on global ocean–sea ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2). *Geosci. Model Dev.* **13**, 4595–4637 (2020).
- Moore, A. M. et al. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. *Prog. Oceanogr.* **91**, 34–49 (2011).
- Thoppil, P. G. et al. Ensemble forecasting greatly expands the prediction horizon for ocean mesoscale variability. *Commun. Earth Environ.* **2**, 89 (2021).
- Fox-Kemper, B. et al. Challenges and prospects in ocean circulation models. *Front. Mar. Sci.* **6**, 65 (2019).
- Kurth, T. et al. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proc. Platform for Advanced Scientific Computing Conference*, 1–11 (ACM, Davos Switzerland, 2023).
- Chen, K. et al. FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead. Preprint at <http://arxiv.org/abs/2304.02948> (2023).
- Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
- Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
- Chen, L. et al. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 190 (2023).
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: A foundation model for weather and climate. In *Proc. 40th International Conference on Machine Learning (ICML)* (2023).
- Bodnar, C. et al. Aurora: A foundation model of the atmosphere. Preprint at <http://arxiv.org/abs/2405.13063> (2024).
- Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* <https://doi.org/10.1038/s41586-024-08252-9> (2024).
- Subel, A. & Zanna, L. Building ocean climate emulators. Preprint at <http://arxiv.org/abs/2402.04342> (2024).
- Xiong, W. et al. AI-GOMS: Large AI-driven global ocean modeling system. Preprint at <http://arxiv.org/abs/2308.03152> (2023).
- Wang, X. et al. XiHe: A data-driven model for global ocean eddy-resolving forecasting. Preprint at <http://arxiv.org/abs/2402.02995> (2024).
- Guo, Z. et al. ORCA: A global ocean emulator for multi-year to decadal predictions. Preprint at <http://arxiv.org/abs/2405.15412> (2024).
- Storto, A. et al. Ocean reanalyses: recent advances and unsolved challenges. *Front. Mar. Sci.* **6**, 418 (2019).

31. Marchuk, G. I., Kochergin, V. P., Klimok, V. I. & Sukhorukov, V. A. On the dynamics of the ocean surface mixed layer. *J. Phys. Oceanogr.* **7**, 865–875 (1977).
32. Charlton-Perez, A. J. et al. Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Clim. Atmos. Sci.* **7**, 93 (2024).
33. Bonavita, M. On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.* **51**, e2023GL107377 (2024).
34. Hollingsworth, A. & Lönnberg, P. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A* **38A**, 111–136 (1986).
35. Storer, B. A., Buzzicotti, M., Khatir, H., Griffies, S. M. & Aluie, H. Global energy spectrum of the general oceanic circulation. *Nat. Commun.* **13**, 5314 (2022).
36. Large, W. G. & Yeager, S. G. The global climatology of an inter-annually varying air–sea flux data set. *Clim. Dyn.* **33**, 341–364 (2009).
37. Jean-Michel, L. et al. The Copernicus Global 1/12° Oceanic and Sea Ice GLORYS12 Reanalysis. *Front. Earth Sci.* **9**, 698876 (2021).
38. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021).
39. Müller, V. et al. Variability of eddy kinetic energy in the eurasian basin of the Arctic Ocean inferred from a model simulation at 1-km resolution. *J. Geophys. Res.: Oceans* **129**, e2023JC020139 (2024).
40. Global Ocean Physics Analysis and Forecast. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). <https://doi.org/10.48670/moi-00016>.
41. Lellouche, J.-M. et al. Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system. *Ocean Sci.* **14**, 1093–1126 (2018).
42. Ryan, A. G. et al. GODAE OceanView Class 4 forecast verification framework: global ocean inter-comparison. *J. Oper. Oceanogr.* **8**, s98–s111 (2015).
43. Ebert, E. E. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorol. Appl.* **15**, 51–64 (2008).
44. Crocker, R., Maksymczuk, J., Mittermaier, M., Tonani, M. & Pequignot, C. An approach to the verification of high-resolution ocean models using spatial methods. *Ocean Sci.* **16**, 831–845 (2020).
45. Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B. & Ebert, E. E. Intercomparison of spatial forecast verification methods. *Weather Forecast.* **24**, 1416–1430 (2009).
46. Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000).
47. Barbosa Aguiar, A. et al. The Met Office Forecast Ocean Assimilation Model (FOAM) using a 1/12-degree grid for global forecasts. *Q. J. R. Meteorol. Soc.* **150**, 3827–3852 (2024).
48. Chassignet, E. P. et al. The HYCOM (HYbrid Coordinate Ocean Model) data assimilative system. *J. Mar. Syst.* **65**, 60–83 (2007).
49. Garrett, C. & Munk, W. Internal Waves in the Ocean. *Annu. Rev. Fluid Mech.* **11**, 339–369 (1979).
50. Madec, G. *NEMO Ocean Engine* (2008).
51. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
52. Brasseur, P. & Verron, J. The SEEK filter method for data assimilation in oceanography: a synthesis. *Ocean Dyn.* **56**, 650–661 (2006).
53. Lellouche, J.-M. et al. Evaluation of global monitoring and forecasting systems at Mercator Océan. *Ocean Sci.* **9**, 57–81 (2013).
54. Hersbach, H. et al. The ERA5 global reanalysis. *Quart. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
55. Bougeault, P. et al. The THORPEX interactive grand global ensemble. *Bull. Am. Meteorol. Soc.* **91**, 1059–1072 (2010).
56. Ni, Q., Zhai, X., LaCasce, J. H., Chen, D. & Marshall, D. P. Full-depth eddy kinetic energy in the global ocean estimated from altimeter and argo observations. *Geophys. Res. Lett.* **50**, e2023GL103114 (2023).
57. Rousset, C. et al. The Louvain-La-Neuve sea ice model LIM3.6: global and regional capabilities. *Geosci. Model Dev.* **8**, 2991–3005 (2015).
58. Lang, S. et al. IFS upgrade brings many improvements and unifies medium-range resolutions. *ECMWF Newsl.* **176**, (2023).
59. Wong, A. P. S. et al. Argo data 1999–2019: two million temperature–salinity profiles and subsurface velocity observations from a global array of profiling floats. *Front. Mar. Sci.* **7**, 700 (2020).
60. Global Ocean - In-Situ Near-Real-Time Observations. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). <https://doi.org/10.48670/moi-00036>.
61. Elipot, S., Sykulski, A., Lumpkin, R., Centurioni, L. & Pazos, M. A dataset of hourly sea surface temperature from drifting buoys. *Sci. Data* **9**, 567 (2022).
62. Rio, M. H. & Hernandez, F. A mean dynamic topography computed over the world ocean from altimetry, in situ measurements, and a geoid model. *J. Geophys. Res.* **109**, 2003JC002226 (2004).
63. Wang, S. et al. A more quiescent deep ocean under global warming. *Nat. Clim. Change* **14**, 961–967 (2024).
64. Cui et al., Forecasting the Eddydy Ocean with a Deep Neural Network, Cuiyingzhe/WenHai: Release WenHai, <https://doi.org/10.5281/zenodo.14881628>, 2025.
65. P. Elson et al. SciTools/cartopy: REL: v0.24.1. Zenodo <https://doi.org/10.5281/ZENODO.13905945> (2024).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (42325601 and 92358303 to Z.J.). Computational resources were supported by Laoshan Laboratory (No. LSKJ202300302 and No. LSKJ202300305). We thank the Mercator Océan and E.U. Copernicus Marine Service Information for providing the GLORYS reanalysis (<https://doi.org/10.48670/moi-00021>), GLO12v4 analysis and forecast (<https://doi.org/10.48670/moi-00016>) and observational data (<https://doi.org/10.48670/moi-00036>, <https://doi.org/10.48670/moi-00041>, <https://doi.org/10.48670/moi-00147>). We thank the ECMWF for providing ERA5 reanalysis (<https://doi.org/10.24381/cds.adbb2d47>) and IFS HRES forecasts (<https://doi.org/10.21957/open-data>). We thank XiHe team for making the model weights and inference code of XiHe available (<https://github.com/Ocean-Intelligent-Forecasting/XiHe-GlobalOceanForecasting>). The Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it (<https://argo.ucsd.edu>, <https://www.ocean-ops.org>). The Argo Program is part of the Global Ocean Observing System. The altimeter products were produced by Ssalto/Duacs and distributed by AVISO+, with support from CNES (<https://www.aviso.altimetry.fr>). The surface drifter data are produced by the Global Drifter Program (<https://www.aoml.noaa.gov/phod/gdp/data.php>).

## Author contributions

Y.C. and R.W. contributed equally to this manuscript. Y.C. and R.W. processed the data and developed pre-trained base model. X.Z. assessed the forecast performance of GOFs. Z.Z. and B.L. conducted the finetune procedure. J.C. and J.S. provided technical support during the training processes. H.L. assisted in the forecast assessment of numerical GOFs and provided guidance on interpreting the assessment results. S.Z. assisted in the data downloading and preprocessing. L.S. assisted in the model training. Z.J. and H.A. instructed the development of WenHai and data analysis. Z.J., H.A., and L.W. conceived the project and led the research.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57389-2>.

**Correspondence** and requests for materials should be addressed to Zhao Jing, Hong An or Lixin Wu.

**Peer review information** *Nature Communications* thanks Gary Brassington, Adam Subel, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025