FuXi-Ocean: A Global Ocean Forecasting System with Sub-Daily Resolution

Qiusheng Huang[1,2†], **Yuan Niu**[3†], **Xiaohui Zhong**[1], **Anboyu Guo**[1], **Lei Chen**[1], **Dianjun Zhang**[3], **Xuefeng Zhang**[3*], **Hao Li**[1,2*]
[1]Artificial Intelligence Innovation and Incubation Institute, Fudan University
[2]Shanghai Innovation Institute
[3]School of Marine Science and Technology,Tianjin University

Abstract

Accurate, high-resolution ocean forecasting is crucial for maritime operations and environmental monitoring. While traditional numerical models are capable of producing sub-daily, eddy-resolving forecasts, they are computationally intensive and face challenges in maintaining accuracy at fine spatial and temporal scales. In contrast, recent data-driven approaches offer improved computational efficiency and emerging potential, yet typically operate at daily resolution and struggle with sub-daily predictions due to error accumulation over time. We introduce FuXi-Ocean, the first data-driven global ocean forecasting model achieving six-hourly predictions at eddy-resolving 1/12° spatial resolution, reaching depths of up to 1500 meters. The model architecture integrates a context-aware feature extraction module with a predictive network employing stacked attention blocks. The core innovation is the Mixture-of-Time (MoT) module, which adaptively integrates predictions from multiple temporal contexts by learning variable-specific reliability , mitigating cumulative errors in sequential forecasting. Through comprehensive experimental evaluation, FuXi-Ocean demonstrates superior skill in predicting key variables, including temperature, salinity, and currents, across multiple depths.

1 Introduction

Ocean forecasting systems play a vital role in maritime operations, providing critical information for navigation, search and rescue, fisheries management, and offshore energy production. These applications require increasingly accurate predictions at finer temporal and spatial resolution to capture rapidly evolving oceanic phenomena. For instance, high-resolution forecasts of ocean currents are indispensable for maritime search and rescue and oil spill tracking, where accurate backtracking and source identification are critical [41]. Despite notable advances in numerical modeling and computational capacity, achieving both high temporal resolution and global coverage remains challenging [16, 13].

Operational ocean forecasting traditionally relies on physics-based numerical models that solve the governing equations of ocean dynamics [15]. While based on well-established physical laws, these models are computationally expensive, particularly when resolving mesoscale eddies, which require grid resolutions of 1/12° or finer. Even after decades of development, ocean general circulation models (OGCMs) still exhibit substantial uncertainties, due to numerical discretization errors [1] and uncertainties in parameterizing unresolved subgrid processes [34].

[*] Corresponding author.

[†] These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recently, deep learning approaches emerge as promising alternatives to traditional OGCMs. Studies [45, 43, 46, 2, 11] demonstrate that such models can achieve comparable or even superior accuracy at daily temporal resolution compared to operational numerical systems, and their operational efficiency can be improved by a thousand times. However, extending these models to sub-daily resolutions presents unique challenges. Ocean variables exhibit diverse temporal dynamics that vary significantly across depths and spatial regions. Surface variables, such as sea surface temperature, often follow pronounced diurnal cycles similar to atmospheric variables, whereas deeper ocean currents evolve on considerably slower timescales. Unlike atmospheric forecasting, which is dominated by high-frequency variability, ocean prediction must accommodate a broad spectrum of temporal behaviors, from fast-changing surface processes to slowly evolving deep-ocean processes. A key limitation of existing deep learning-based ocean forecasting models lies in their predominant focus on daily forecasts, which restricts their ability to adaptively model variable-specific temporal dynamics at sub-daily intervals. Lacking mechanisms to adaptively process temporal information across multiple timescales, these models often struggle with the complexity of high-frequency sub-daily predictions. Moreover, designing effective sub-daily models requires architectures capable of learning efficiently from limited historical data while avoiding overfitting to spurious correlations.

In this work, we present FuXi-Ocean, the first deep learning-based global ocean forecasting model to achieve six-hour temporal resolution at eddy-resolving 1/12° spatial resolution. Our approach explicitly addresses the challenges of sub-daily ocean prediction through three key innovations: First, we design an autoregressive architecture specifically tailored to capture multi-scale temporal dependencies of different oceanic variables across an unprecedented depth range (0-1500 m). Unlike previous models that treat all variables uniformly, our model adaptively learns temporal context appropriate for each variable and region, effectively distinguishing between fast-evolving surface processes (e.g., diurnal warming) and slowly varying deep-ocean dynamics. Second, we introduce the Mixture-of-Time (MoT) module that adaptively integrates predictions from multiple temporal windows based on their empirical reliability for each variable. This mechanism enables the model to select the most informative temporal context, thereby mitigating the accumulation of forecast errors typically encountered in sequential prediction tasks. Third, we demonstrate remarkable data efficiency, achieving state-of-the-art performance with only 9 years of training datasignificantly less than required by comparable models. This efficiency stems from our architecture's ability to effectively leverage physical constraints and spatial coherence in the learning process. Our key contributions are as follows:

- We propose FuXi-Ocean, the first data-driven global ocean forecasting model to achieve six-hour temporal resolution, 1/12° spatial resolution, and 0-1500 m depth coverage.
- We introduce the Mixture-of-Time (MoT) module, which adaptively integrates variable-specific temporal dependencies to reduce cumulative errors in sequential prediction.
- We validate FuXi-Ocean with reanalysis and observational datasets, demonstrating superior performance over traditional numerical forecasting models at sub-daily intervals.

2 Related Work

2.1 Numerical Models

Ocean forecasting traditionally relies on OGCMs that solve fundamental physical equations governing ocean dynamics. State-of-the-art operational systems, such as the HYbrid Coordinate Ocean Model (HYCOM) [8, 3], Ocean Physical System (PSY4) [27], Global Ice Ocean Prediction System (GIOPS) [42], Forecast Ocean Assimilation Model (FOAM) [5], BLUElinK OceanMAPS (BLK) [40], Nucleus for European Modelling of the Ocean (NEMO) [19], Modular Ocean Model (MOM) [17], Real-Time Ocean Forecast System (RTOFS) [14] and GLORY12 [22], make significant advances in global ocean prediction capabilities.

Despite their solid theoretical foundations, these models face persistent challenges. The computational cost increases dramatically with resolution, making global simulations at eddy-resolving scales particularly expensive [36]. Additionally, uncertainties in parameterizing unresolved processes, such as vertical mixing and air-sea interactions, introduce systematic model biases [44, 21]. The inherent chaotic nature of ocean systems further complicates forecasting, as small errors in initial conditions can rapidly amplify through nonlinear interactions. These limitations become particularly acute for high-frequency predictions, where sub-daily forecasts reveal the full spectrum of model deficiencies that might otherwise be obscured in daily averages.

2.2 Data-Driven Deep Learning Models

Recent advances in deep learning introduce promising new approaches for ocean forecasting and successfully applied to the global medium-range forecasts [24, 9, 26, 4, 33, 6, 37].At present, deep learning based ocean forecasting technology is developing rapidly and can predict ocean variables such as satellite sea surface temperature, sea surface height, wave height, temperature, salinity, U and V [47, 38], as well as seasonal or interannual ocean phenomena such as Indian Ocean Dipole (IOD) and ElNiño Southern Oscillation (ENSO) [20, 48, 49]. Wang et al. [43] developed "XiHe", a global forecasting model based on the Swin-Transformer architecture, incorporating specialized ocean-land mask mechanisms. Yang et al. [46] introduced "Langya", which mitigates cumulative forecast errors by employing a Time Embedding module, enabling forecasts for up to 7 days without iterative steps. Both models achieve state-of-the-art performance for daily predictions at 1/12° resolution. In addition, Cui et al. [11] and Xiong et al. [45] also experimented with autoregressive forecasting architectures and achieved good performance.

Despite these advances, existing data-driven ocean forecasting approaches face three key limitations that constrain their practical utility. First, temporal resolution remains limited to daily intervals. Ocean parameters evolve on markedly different timescales, with surface temperatures exhibiting pronounced diurnal cycles while deep currents evolving more slowly over extended time periods. Without mechanisms to adaptively learn these multiscale temporal dependencies, current models are fundamentally limited in their ability to capture the full spectrum of ocean dynamics at sub-daily resolutions. Second, vertical coverage in existing models remains severely constrained, typically not extending beyond 700 m depth. The thermocline structure and deep water mass properties below this depth play crucial roles in energy transfer and climate dynamics, yet remain largely unaddressed in current data-driven frameworks. Third, most current approaches rely heavily on atmospheric forcing as input variables, creating additional computational overhead and external data dependencies.

3 Method

This section details our methodology for high-frequency ocean forecasting. We first describe our data preparation approach, then present the architecture of FuXi-Ocean, including our novel Mixture-of-Time module for adaptive temporal modeling, and finally outline our training strategy.

3.1 Problem Formulation

Ocean forecasting requires processing high-dimensional spatiotemporal data that captures the complex dynamics of global marine systems. FuXi-Ocean addresses the task of predicting future oceanic states from sequential historical observations with high spatial and temporal precision. Formally, given a sequence of historical observations spanning N consecutive time points $\{\mathrm{X}^{t-N+1}, \mathrm{X}^{t-N+2}, \ldots, \mathrm{X}^{t}\} \in \mathbb{R}^{N \times C \times H \times W}$, we aim to predict the state at the subsequent time step $\mathrm{X}_{t+1} \in \mathbb{R}^{C \times H \times W}$. Here, $H = 2160$ and $W = 4320$ represent the spatial dimensions of our global grid at 1/12° resolution, providing eddy-resolving capability essential for capturing mesoscale ocean dynamics.

We focus on five fundamental ocean state variables that collectively characterize the physical state of the global ocean: temperature (T), salinity (S), zonal and meridional components of ocean currents (U and V), and sea surface height (SSH). While SSH is inherently a surface variable, the remaining four variables exhibit substantial vertical variability that necessitates prediction across multiple depth levels. We discretize the water column into 20 strategically selected depth levels from the surface (0m) to the deeper ocean (1500 m): 0, 2, 4, 6, 10, 20, 30, 40, 50, 60, 70, 80, 100, 125, 150, 200, 300, 500, 1000, and 1500 m. This vertical resolution is particularly dense near the surface to accurately capture upper-ocean dynamics that strongly influence air-sea interactions. Consequently, our input and output tensors comprise $C = 81$ channels (4 variables $\times$ 20 depth levels + SSH).

The temporal resolution of our data is set at 6-hour intervals, aligning with operational oceanic forecasting standards and balancing computational feasibility with information density. We empirically determine that $N = 4$ consecutive time steps (spanning 24 hours) provides sufficient historical context to accurately capture diurnal cycles and inertial oscillations while maintaining computational efficiency. To preserve physical coherence in predictions, we apply a land-sea mask to all variables, ensuring the model only processes and predicts values for ocean grid points.

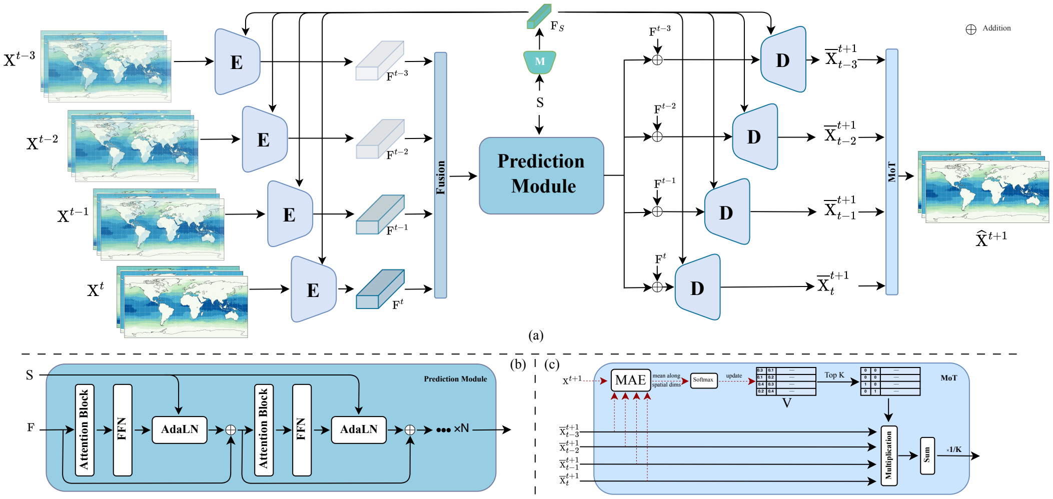

Figure 1: **Architecture of our ocean forecasting framework.** Our model processes sequential ocean states X^{t-3} through X^{t} to predict $\widehat{X}^{t+1}$. (a) The main pipeline consists of a shared encoder $\mathbf{E}$ that transforms input states into latent representations, modulated by spatiotemporal features F_S from network $\mathbf{M}$. The fused representations feed into the prediction module, whose outputs are processed by decoders $\mathbf{D}$ with skip connections. (b) The prediction module employs stacked attention blocks with adaptive layer normalization (AdaLN) and feed-forward networks (FFN), capturing complex temporal dynamics. (c) The Mixture-of-Time (MoT) module performs channel-wise selection across the four decoder outputs from different temporal skip connections. For each channel, MoT identifies the top-K temporal dependencies using matrix V (derived from spatially-averaged MAE metrics and softmax) and computes the optimal weighted average to synthesize the final prediction $\widehat{X}^{t+1}$.

3.2 Model Architecture

FuXi-Ocean employs an autoregressive architecture designed to capture the multiscale temporal dynamics of oceanic variables. The model comprises three primary components: a feature extraction module that encodes multi-temporal inputs, a prediction module that captures temporal evolution patterns, and a feature remapping module that reconstructs the target variables. Figure 1 provides an overview of the model architecture.

3.2.1 Feature Extraction

Ocean forecasting requires robust feature representations that capture both spatial dependencies and temporal correlations. We design a feature extraction pipeline that combines context-aware encoding with efficient feature fusion, shown in Fig. 1(a).

Our design employs a shared encoder $\mathbf{E}$ for patch embedding. This encoder uses convolutional layers with matched kernel size and stride for spatial downsampling, followed by layer normalization. This approach balances computational efficiency with representation power when processing global ocean data. To leverage spatiotemporal context, we implement a prior information network $\mathbf{M}$ that processes:

$$F_S = \mathbf{M}(S) \tag{1}$$

where S contains temporal information (diurnal, seasonal patterns) and spatial data (coordinates, bathymetry). The network $\mathbf{M}$ uses sinusoidal positional encodings for temporal components and learnable embeddings for spatial coordinates. We then modulate the encoder weights using these contextual features through:

$$F^t = \mathbf{E}(X^t, F_S) = \mathbf{Norm}(\mathbf{Conv}(X^t, W \odot F_S)) \tag{2}$$

where W represents the learnable convolution parameters and $\odot$ denotes element-wise multiplication. This modulation enhances the encoder's sensitivity to region-specific patterns, such as boundary currents or seasonal mixed layer dynamics. The encoder processes each input time step $t - i$ ($i \in \{0, 1, 2, 3\}$) to produce feature tensors $\mathrm{F}^{t-i} \in \mathbb{R}^{C' \times H' \times W'}$. Here, C' is the feature dimension while H' and W' represent the downsampled spatial dimensions.

Our feature fusion module concatenates these tensors along the channel dimension, preserving their temporal characteristics. We then apply 1×1 convolutions with normalization layers to integrate information while maintaining spatial coherence. This fusion approach enables the model to capture complex spatiotemporal patterns essential for accurate forecasting, without compromising computational efficiency.

3.2.2 Prediction Module

Building upon the extracted features, our prediction process involves a sequence of specialized components designed to model oceanic dynamics across multiple temporal scales, as illustrated in Figure 1(b).

The prediction module processes the fused representations to model the nonlinear evolution of oceanic states, which consists of stacked attention blocks[28] paired with feed-forward networks. Each attention block captures dependencies across spatiotemporal features, while adaptive layer normalization (AdaLN)[18] incorporates contextual information F_S to modulate normalization parameters. This design allows the model to focus selectively on relevant oceanic patterns while maintaining computational efficiency.

For feature reconstruction, we employ a shared decoder $\mathbf{D}$ that transforms latent representations back into the physical variable space. The decoder consists of transposed convolutional layers with normalization operations. We establish skip connections between encoder features and corresponding decoder layers, preserving fine-grained spatial details often lost during predictiona critical factor when modeling mesoscale ocean dynamics.

3.2.3 Mixture-of-Time Module

A key innovation in our framework is the Mixture-of-Time (MoT) module, illustrated in Figure 1(c), that adaptively integrates predictions from multiple temporal windows. This module addresses a fundamental challenge in ocean forecastingdifferent oceanic variables exhibit distinct temporal evolution characteristics, from fast-changing surface processes to the relatively slow and stable subsurface variability extending down to the twilight zone (1500 m). The MoT module performs channel-wise adaptive selection across predictions derived from different temporal windows. Specifically, we maintain a selection matrix $V \in \mathbb{R}^{C \times 4}$ that quantifies prediction reliability across temporal scales for each channel (variable and depth combination). For a given channel c and spatial location (h, w), the final prediction is computed as:

$$\widehat{\mathrm{X}}^{t+1}(c, h, w) = \frac{1}{K} \sum_{i=0}^{3} \mathrm{TopK}_{\mathrm{V}(c,i)} \cdot \overline{\mathrm{X}}_{t-i}^{t+1}(c, h, w) \tag{3}$$

Here, the TopK operator selects the K most reliable temporal windows (corresponding to smallest values in V) for each channel. Specifically, for each channel c, it assigns a value of 1 to the K smallest elements in the vector $\mathrm{V}(c, \cdot)$ and 0 to all others. $\overline{\mathrm{X}}_{t-i}^{t+1}$ represents the forecast for time $t + 1$ generated using historical context beginning at time $t - i$. This approach enables variable-specific temporal context selection without increasing model complexity. During training, we update the selection matrix based on prediction performance:

$$\mathrm{V}(c, i) = \alpha \cdot \mathrm{V}(c, i) + (1 - \alpha) \cdot \mathrm{softmax}(\mathrm{AvgPool}_{\mathrm{H,W}}(\mathrm{MAE}(\overline{\mathrm{X}}_{t-i}^{t+1}, \mathrm{X}^{t+1})))_c \tag{4}$$

where α controls update momentum, and MAE measures prediction error. The spatially averaged MAE captures the reliability of each temporal window for each variable, which is then normalized through softmax to ensure the selection weights sum to one. This adaptive approach enables FuXi-Ocean to dynamically adjust its reliance on different temporal windows for each variable and region. For example, deeper ocean current predictions might benefit from recent observations that capture immediate flow evolution, while surface temperature predictions might rely more on longer temporal contexts that effectively represent diurnal cycles and day-night transitions.

3.3 Training Strategy

Training deep learning models for ocean forecasting presents unique challenges due to Earth's spherical geometry. To address the varying grid cell areas across latitudes, we implement a latitude-weighted Charbonnier loss[7]:

$$\mathcal{L}_{\text{pred}} = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \alpha_i \sqrt{(\widehat{\mathrm{X}}_{c,i,j}^{t+1} - \mathrm{X}_{c,i,j}^{t+1})^2 + \epsilon^2} \tag{5}$$

where $\alpha_i = H \times \frac{\cos \Phi_i}{\sum_{i=1}^{H} \cos \Phi_i}$ is the latitude-dependent weighting factor at latitude Φ_i, and ϵ is a small constant ensuring differentiability. This approach prevents bias toward high-latitude regions in error calculations, which would otherwise be overrepresented in the rectangular grid projection.

Besides, we employ a two-stage training approach to enhance model stability and performance. In the first stage, we pre-train the model using a single-step prediction objective. In the second stage, we fine-tune the model with a multi-step loss that penalizes predictions across multiple consecutive time steps, mitigating error accumulation in autoregressive forecasting. This approach follows recent advances in autoregressive modeling [10, 11], which demonstrate the effectiveness of multi-step training for improving long-term forecast stability.

4 Experiment

4.1 Data

HYCOM-RD. We train and evaluate FuXi-Ocean using the HYCOM Reanalysis Data [8], the only publicly available ocean dataset with 6-hour temporal resolution. This dataset features 1/12° horizontal resolution with up to 40 vertical layers from sea surface to 5000m depth. We selecte approximately 8.5 years (January 2006 to June 2014) of data for training, a six-month period (July to December 2014) for validation, and one year (January 2015 to December 2015) for testing, focusing on 20 strategically chosen vertical layers down to 1500 m that capture essential ocean dynamics.
IV-TT Framework. For evaluation against real-world observations, we employ the GODAE Ocean View Intercomparison and Validation Task Team (IV-TT) Class 4 framework [39], which is obtained from publicly available sources. Details are in the appendix. This framework provides observational datasets from drifting buoys for 2022, alongside interpolated outputs from operational numerical forecasting systems. In addition, we apply a filter to remove anomalous points. Specifically, we compare HYCOM reanalysis data with observational data and remove outliers with high MAE from the observations to ensure fairness in the comparison process. However, after evaluation, the raw data errors for salinity and temperature were too large to be used directly. We primarily evaluate sea surface temperature (SST) forecasts, as this variable exhibits high variability and effectively demonstrates forecasting skill. To ensure consistent comparison, we transform the irregularly distributed observational data to a regular grid matching FuXi-Ocean's output format using nearest-neighbor interpolation. For operational evaluation, FuXi-Ocean is initialized with six-hourly analysis fields from HYCOM.

4.2 Implementation Details

The FuXi-Ocean employs the PyTorch framework [35] with the AdamW [23, 29] optimizer, configured with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a cosine annealing learning rate schedule [30] that decays from 2.5×10^{-4} to 10^{-8}. We train on a cluster of 4 NVIDIA H100 GPUs for 60,000 iterations with a batch size of 1 per GPU, requiring approximately 81 hours to complete. Similar to other autoregressive models [10, 11], FuXi adopts an autoregressive approach during inference, feeding the output back into the input to iteratively generate results for the next time step. For further details, please refer to the supplementary materials.

4.3 Metrics

Following standard practices in operational ocean forecast evaluation[11], we use latitude-weighted root mean square error (RMSE) as our primary evaluation metrics to account for the varying grid cell

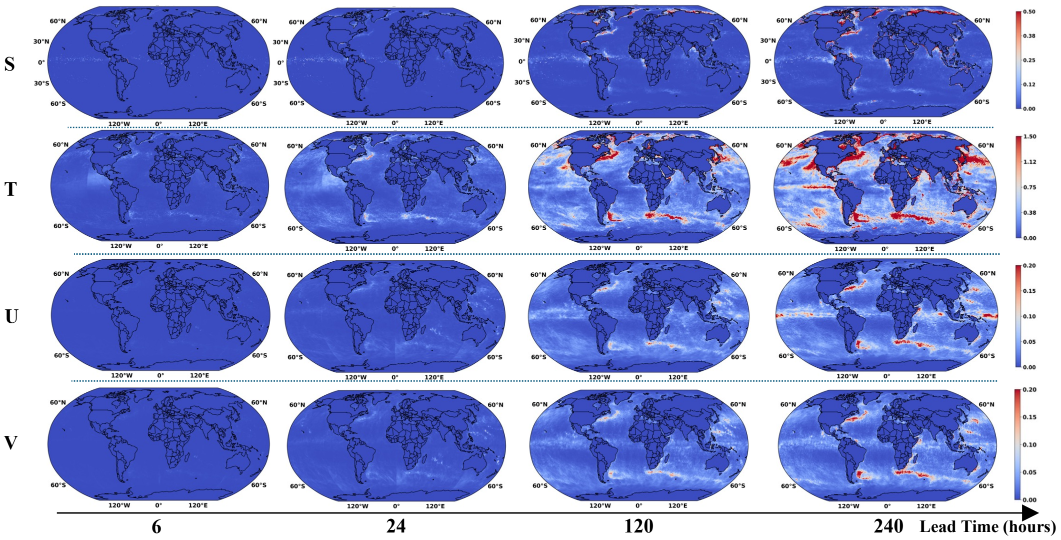

Figure 2: **Global RMSE distribution of sea surface.** From top to bottom, the results represent salinity (psu), temperature (řC), and ocean current U/V components (m/s), and from left to right correspond to different forecast lead times. Each subplot represents the average RMSE (lower is better) for the test set.

sizes across different latitudes, formulated as follows:

$$\mathrm{RMSE}(c,\tau) = \frac{1}{|\mathrm{D}|}\sum_{t_0 \in \mathrm{D}} \sqrt{\frac{1}{\mathrm{H}\times\mathrm{W}}\sum_{i=1}^{\mathrm{H}}\sum_{j=1}^{\mathrm{W}} a_i(\widehat{\mathbf{X}}_{c,i,j}^{t_0+\tau} - \mathbf{X}_{c,i,j}^{t_0+\tau})^2} \tag{6}$$

where t_0 denotes the forecast initialization time in the testing dataset D, and τ represents the lead time steps added to t_0.

5 Results

We first evaluate FuXi-Ocean on the 2015 HYCOM set derived from HYCOM reanalysis data to assess its ability to generate high-frequency (6-hourly) predictions. We then compare our model with operational forecasting systems using 2022 observational data through the GODAE Ocean View IV-TT framework. This dual evaluation approach allows us to comprehensively assess both the model's intrinsic predictive capabilities and its real-world operational performance relative to established systems.

5.1 Performance testing

Horizontal Spatial Error Analysis. Fig. 2 shows the global spatial distribution of RMSE for sea surface variables (temperature, salinity, and currents) at different lead times (6, 24, 120, and 240 hours). For temperature (T), the model maintains particularly low errors in the tropical and subtropical regions across all forecast horizons, with RMSE values typically below 0.3řC for 6-hour predictions. Error patterns intensify primarily in western boundary current regions (Gulf Stream, Kuroshio Current) and the Antarctic Circumpolar Current, where intense mesoscale eddy activity creates inherent predictability challenges [31, 12]. Notably, due to the high uncertainty of changes in these regions, even observational data from satellites and other sources can be significantly affected, leading to instability in model results [32, 12]. Salinity (S) forecasts display a similar spatial pattern but with larger relative errors in regions of significant freshwater influence, such as major river outflows and high-precipitation zones [25]. The equatorial Pacific shows particularly strong performance across all lead times, maintaining low error values even at 10-day forecasts. At shorter forecast times, the error distribution locations of the current velocity components U and V are highly similar. As the forecast lead time increases, significant differences in U and V near the equator emerge, which aligns with the physical phenomenon of the inconsistent rotation directions of warm currents in the Northern and Southern Hemispheres. Western boundary currents show the highest error magnitudes, consistent with their intrinsic variability. Nevertheless, FuXi-Ocean maintains reasonable accuracy in these challenging regions, with errors remaining within operational tolerance thresholds even at longer lead times. Crucially, our 6-hour predictions show substantially lower errors than daily forecasts across all variables, validating the model's ability to capture sub-daily ocean dynamics.

Vertical Performance Analysis. The vertical profile of the ocean environment is a crucial aspect for evaluating the effectiveness of ocean prediction systems. In Fig. 3, we present the variation of RMSE for salinity and temperature with depth. First, FuXi-Ocean maintains consistent prediction skill throughout the water column, with RMSE degrading gradually with depth. Second, the thermocline region (approximately 100-300m) shows relatively higher errors compared to both surface and deep layers, reflecting the inherent difficulty in predicting this dynamically complex interface. Despite this challenge, FuXi-Ocean's RMSE still maintains effective values. Additionally, the error growth with increasing forecast lead time shows an interesting depth-dependent pattern. Thanks to our Mixture-of-Time (MoT) approach, which adaptively weights temporal dependencies for each variable, even though surface layers (0-300m) are sensitive to rapidly changing conditions while deeper layers (beyond 500m) are not, the error growth across layers remains relatively stable without any layer performing significantly worse.

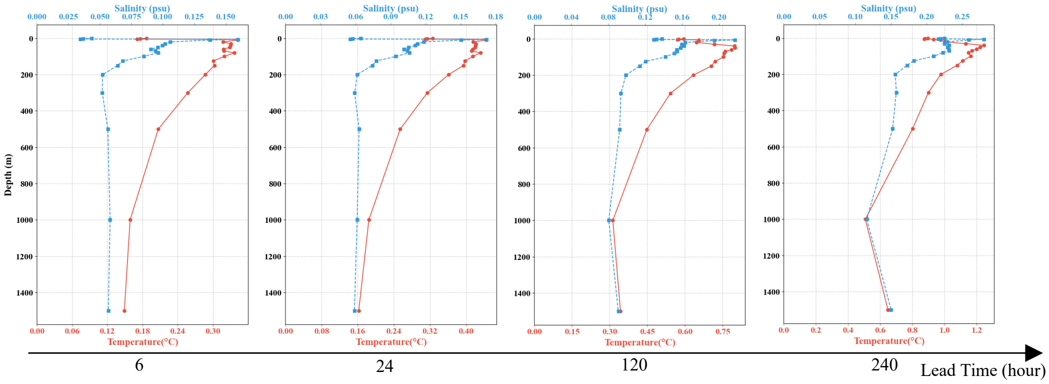

Figure 3: **Depth-dependent RMSE Distributions of salinity and temperature varying with lead time.** Each subplot represents the RMSE (lower is better) varying with depth at the current lead time. Blue represents salinity results, while red represents temperature results.

5.2 Comparison

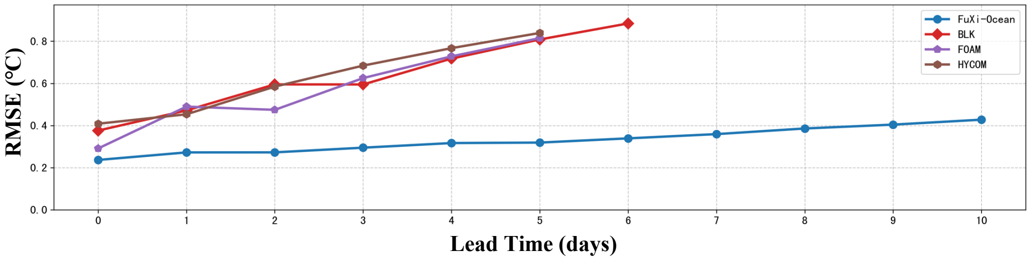

Figure 4: **The SST comparison of different methods based on the IV-TT evaluation framework.** The x-axis represents the forecast time, and the y-axis represents the RMSE (lower is better).

To evaluate the actual operational performance, we used sea surface temperature observation data from 2022 within the IV-TT framework to compare FuXi-Ocean with the HYCOM, BLK, and FOAM methods. For this comparison, we averaged our 6-hourly outputs to produce daily means that can be directly compared with the daily forecasts of other methods. To ensure a fair comparison, we also evaluated the RMSE at the initial time (0 day). As shown in Fig. 4, the RMSE of FuXi-Ocean is lower than that of other methods, and the cumulative error growth is significantly smaller. Note that the most recent training data for FuXi-Ocean does not exceed 2014, and it uses only oceanic variables as input information, highlighting that FuXi-Ocean effectively captures the intrinsic patterns of ocean system changes.

5.3 Ablation

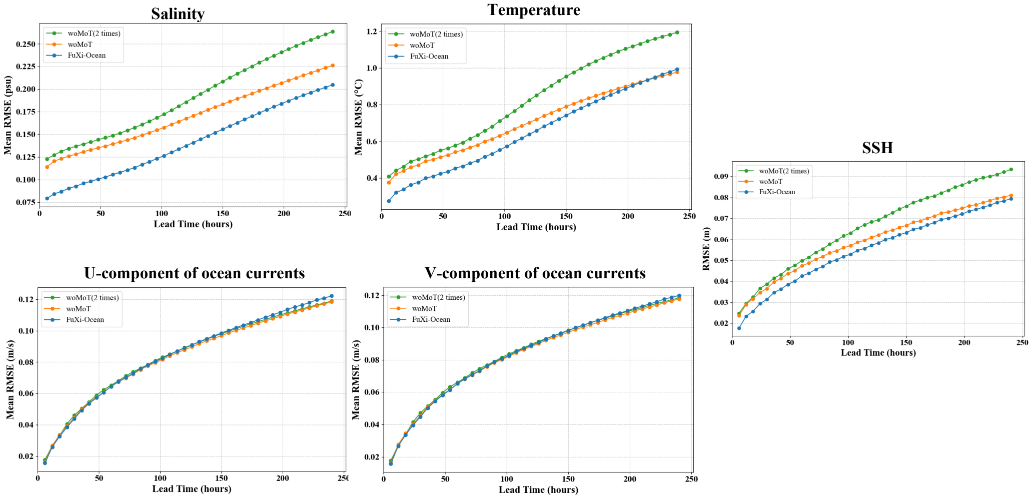

Figure 5: **The ablation study of different methods on the HYCOM-RD.** The five subplots respectively show salinity, temperature, ocean current UV components, and sea surface height. The x-axis represents the forecast lead time, and the y-axis represents the RMSE (lower is better). Note that for each variable, we average the RMSE over depth.

To evaluate the contribution of our key innovations, we compare FuXi-Ocean with two ablated variants: (1) "woMoT," which removes the MoT module, and (2) "woMoT(2times)," which removes MoT and reduces input time steps from four to two.

Figure 5 presents the results across all five forecasted variables. For salinity, temperature, and SSH, removing the MoT module leads to substantial performance degradation, with the increasing of RMSE peaked at nearly 40% across forecast lead times. This degradation is particularly pronounced for short-term forecasts, where adaptive temporal context selection proves critical for capturing rapidly evolving processes. These findings confirm that variable-specific temporal modeling significantly improves prediction accuracy for thermohaline variables that exhibit complex spatiotemporal dynamics. Furthermore, reducing the historical context from four to two time steps (woMoT(2times)) causes additional performance deterioration, especially for temperature and salinity predictions beyond 48 hours. This contrasts with experiences in atmospheric forecasting models, where shorter historical windows often suffice. Interestingly, ocean current components (U and V) show less sensitivity to both the removal of MoT and the reduction of temporal context. This observation aligns with the physical reality that surface currents are largely driven by recent wind forcing and geostrophic balance rather than their own history. The minimal performance difference between ablated models for current forecasting suggests that specialized architectures may be warranted for different ocean variablesa direction we leave for future work.

6 Conclusion

We present FuXi-Ocean, the first data-driven global ocean forecasting system achieving 6-hour temporal resolution at a 1/12° spatial scale with coverage from surface to 1500 m depth. Our Mixture-of-Time module adaptively captures different temporal dependencies across ocean variables, addressing a key limitation of previous models. FuXi-Ocean outperforms operational numerical systems in predicting sea surface temperature while requiring significantly fewer computational resources and relying only on ocean variables as input. Its 6-hour forecasts demonstrate superior accuracy than daily-averaged predictions from state-of-the-art numerical models, confirming the effectiveness of our approach for high-frequency ocean prediction.

The successful application of deep learning to sub-daily ocean forecasting opens several promising avenues for future research directions. First, while our current model covers depths up to 1500 m, extending its range to abyssal depths would provide a more complete representation of the ocean system. Second, incorporating physical conservation laws as additional constraints could further improve prediction stability and ensure physical consistency. Third, increasing the forecast frequency beyond 6 hours toward hourly predictions could enable applications requiring even finer temporal resolution, such as tidal forecasting and coastal hazard warning systems.

Acknowledgments and Disclosure of Funding

This work was supported by AI for Science Program (ID: 2025-GZL-RGZN-BTBX-02017), Shanghai Municipal Commission of Economy and Information

References

[1] Adcroft, A., Hallberg, R., Harrison, M., 2008. A finite volume discretization of the pressure gradient force using analytic integration. Ocean Modelling 22, 106–113. doi:https://doi.org/10.1016/j.ocemod.2008.02.001.

[2] Aouni, A.E., Gaudel, Q., Regnier, C., Gennip, S.V., Drevillon, M., Drillet, Y., Lellouche, J.M., 2024. Glonet: Mercator's end-to-end neural forecasting system. Preprint at https://arxiv.org/abs/2412.05454.

[3] Barton, N., Metzger, E.J., Reynolds, C.A., Ruston, B., Rowley, C., Smedstad, O.M., Ridout, J.A., Wallcraft, A., Frolov, S., Hogan, P., et al., 2021. The navy's earth system prediction capability: A new global coupled atmosphere-ocean-sea ice prediction system designed for daily to subseasonal forecasting. Earth and Space science 8, e2020EA001199.

[4] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3d neural networks. Nature 619, 533–538.

[5] Blockley, E., Martin, M., McLaren, A., Ryan, A., Waters, J., Lea, D., Mirouze, I., Peterson, K., Sellar, A., Storkey, D., 2014. Recent development of the met office operational ocean forecasting system: an overview and assessment of the new global foam forecasts. Geoscientific Model Development 7, 2613–2638.

[6] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al., 2024. Aurora: A foundation model of the atmosphere. arXiv preprint arXiv:2405.13063 .

[7] Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M., 1994. Two deterministic half-quadratic regularization algorithms for computed imaging, in: Proceedings of 1st International Conference on Image Processing, pp. 168–172 vol.2. doi:10.1109/ICIP.1994.413553.

[8] Chassignet, E.P., Hurlburt, H.E., Smedstad, O.M., Halliwell, G.R., Hogan, P.J., Wallcraft, A.J., Baraille, R., Bleck, R., 2007. The hycom (hybrid coordinate ocean model) data assimilative system. Journal of Marine Systems 65, 60–83. URL: https://www.sciencedirect.com/science/article/pii/S0924796306002855, doi:https://doi.org/10.1016/j.jmarsys.2005.09.016. marine Environmental Monitoring and Prediction.

[9] Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.J., Chen, X., Ma, L., Zhang, T., Su, R., et al., 2023a. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv preprint arXiv:2304.02948 .

[10] Chen, L., et al., 2023b. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. npj Climate and Atmospheric Science , 1–11.

[11] Cui, Y., Wu, R., Zhang, X., Zhu, Z., Liu, B., Shi, J., Chen, J., Liu, H., Zhou, S., Su, L., et al., 2025. Forecasting the eddying ocean with a deep neural network. Nature Communications 16, 2268.

[12] Dong, C., You, Z., Dong, J., Ji, J., Sun, W., Xu, G., Lu, X., Xie, H., Teng, F., Liu, Y., et al., 2025. Oceanic mesoscale eddies. Ocean-Land-Atmosphere Research 4, 0081.

[13] Fox-Kemper, B., et al., 2019. Challenges and prospects in ocean circulation models. Frontiers in Marine Science Volume 6 - 2019. doi:10.3389/fmars.2019.00065.

[14] Garraffo, Z.D., Cummings, J.A., Paturi, S., Hao, Y., Iredell, D., Spindler, T., et al., 2020. Rtofs-da: real time ocean-sea ice coupled three dimensional variational global data assimilative ocean forecast system. Research activities in Earth system modelling .

[15] Griffies, S., 2018. Fundamentals of ocean climate models. Princeton university press.

[16] Griffies, S., Adcroft, A., Banks, H., Böning, C., Chassignet, E., Danabasoglu, G., Danilov, S., Deleersnijder, E., Drange, H., England, M., et al., 2009. Problems and prospects in large-scale ocean circulation models. Proceedings of OceanObs 9, 410–431.

[17] Griffies, S.M., Harrison, M.J., Pacanowski, R.C., Rosati, A., et al., 2004. A technical guide to mom4. GFDL Ocean Group Tech. Rep 5, 371.

[18] Guo, Y., Wang, C., Yu, S.X., McKenna, F., Law, K.H., 2022. Adaln: a vision transformer for multidomain learning and predisaster building information extraction from images. Journal of Computing in Civil Engineering 36, 04022024.

[19] Gurvan, M., Bourdallé-Badie, R., Bricaud, C., Bruciaferri, D., Calvert, D., Chanut, J., Clementi, E., Coward, A., Delrosso, D., Ethé, C., et al., 2017. Nemo ocean engine .

[20] Ham, Y.G., Kim, J.H., Luo, J.J., 2019. Deep learning for multi-year enso forecasts. Nature 573, 568–572.

[21] Jayne, S.R., 2009. The impact of abyssal mixing parameterizations in an ocean general circulation model. Journal of Physical Oceanography 39, 1756–1775.

[22] Jean-Michel, L., Eric, G., Romain, B.B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L.G., Charly, R., et al., 2021. The copernicus global 1/12 oceanic and sea ice glorys12 reanalysis. Frontiers in Earth Science 9, 698876.

[23] Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980. preprint at https://arxiv.org/abs/1412.6980.

[24] Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., Anandkumar, A., 2023. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators, in: Proceedings of the platform for advanced scientific computing conference, pp. 1–11.

[25] Lagerloef, G., Schmitt, R., Schanze, J., Kao, H.Y., 2010. The ocean and the global water cycle. Oceanography 23, 82–93.

[26] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al., 2023. Learning skillful medium-range global weather forecasting. Science 382, 1416–1421.

[27] Lellouche, J.M., Greiner, E., Le Galloudec, O., Regnier, C., Benkiran, M., Testut, C.E., Bourdalle-Badie, R., Drevillon, M., Garric, G., Drillet, Y., 2018. The mercator ocean global high-resolution monitoring and forecasting system. New Frontiers in Operational Oceanography 1, 563–592.

[28] Liu, Z., et al., 2022. Swin transformer v2: Scaling up capacity and resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11999–12009.

[29] Loshchilov, I., Hutter, F., 2017a. Decoupled weight decay regularization, in: International Conference on Learning Representations.

[30] Loshchilov, I., Hutter, F., 2017b. Sgdr: Stochastic gradient descent with warm restarts. Preprint at https://arxiv.org/abs/1608.03983.

[31] Morrow, R., Fu, L.L., Farrar, J.T., Seo, H., Le Traon, P.Y., 2017. Ocean eddies and mesoscale variability, in: Satellite altimetry over oceans and land surfaces. CRC press, pp. 315–342.

[32] Morrow, R., Le Traon, P.Y., 2012. Recent advances in observing mesoscale ocean dynamics with satellite altimetry. Advances in Space Research 50, 1062–1076.

[33] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A., 2023. Climax: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343 .

[34] Palmer, T.N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G.J., Steinheimer, M., Weisheimer, A., 2009. Stochastic parametrization and model uncertainty .

[35] Paszke, A., et al., 2017. Automatic differentiation in pytorch, in: NIPS 2017 Workshop on Autodiff.

[36] Piomelli, U., 2014. Large eddy simulations in 2030 and beyond. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 372, 20130320.

[37] Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al., 2025. Probabilistic weather forecasting with machine learning. Nature 637, 84–90.

[38] Quach, B., Glaser, Y., Stopa, J.E., Mouche, A.A., Sadowski, P., 2020. Deep learning for predicting significant wave height from synthetic aperture radar. IEEE Transactions on Geoscience and Remote Sensing 59, 1859–1867.

[39] Ryan, A., Régnier, C., Divakaran, P., Spindler, T., Mehra, A., Smith, G.C., Davidson, F., Hernandez, F., Maksymczuk, J., Liu, Y., 2015. Godae oceanview class 4 forecast verification framework: global ocean inter-comparison. Journal of Operational Oceanography 8, s111 – s98. URL: https://api.semanticscholar.org/CorpusID:118680054.

[40] Schiller, A., Brassington, G.B., Oke, P., Cahill, M., Divakaran, P., Entel, M., Freeman, J., Griffin, D., Herzfeld, M., Hoeke, R., et al., 2020. Bluelink ocean forecasting australia: 15 years of operational ocean service delivery with societal, economic and environmental benefits. Journal of Operational Oceanography 13, 1–18.

[41] Schiller, A., Davidson, F.J.M., DiGiacomo, P.M., Wilmer-Becker, K., 2016. Better informed marine operations and management: Multidisciplinary efforts in ocean forecasting research for socioeconomic benefit. Bulletin of the American Meteorological Society 97, 1553–1559. URL: https://api.semanticscholar.org/CorpusID:56107855.

[42] Smith, G.C., Roy, F., Reszka, M., Surcel Colan, D., He, Z., Deacu, D., Belanger, J.M., Skachko, S., Liu, Y., Dupont, F., et al., 2016. Sea ice forecast verification in the canadian global ice ocean prediction system. Quarterly Journal of the Royal Meteorological Society 142, 659–671.

[43] Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., Wang, H., Wang, S., Zhu, J., Xu, J., et al., 2024. Xihe: A data-driven model for global ocean eddy-resolving forecasting. arXiv preprint arXiv:2402.02995 .

[44] Wunsch, C., Ferrari, R., 2004. Vertical mixing, energy, and the general circulation of the oceans. Annu. Rev. Fluid Mech. 36, 281–314.

[45] Xiong, W., Xiang, Y., Wu, H., Zhou, S., Sun, Y., Ma, M., Huang, X., 2023. Ai-goms: Large ai-driven global ocean modeling system. arxiv. arXiv preprint arXiv:2308.0315 .

[46] Yang, N., Wang, C., Zhao, M., Zhao, Z., Zheng, H., Zhang, B., Wang, J., Li, X., 2024. Langya: Revolutionizing cross-spatiotemporal ocean forecasting. arXiv preprint arXiv:2412.18097 .

[47] Zheng, G., Li, X., Zhang, R.H., Liu, B., 2020. Purely satellite data–driven deep learning forecast of complicated tropical instability waves. Science advances 6, eaba1482.

[48] Zhou, L., Zhang, R.H., 2022. A hybrid neural network model for enso prediction in combination with principal oscillation pattern analyses. Advances in Atmospheric Sciences 39, 889–902.

[49] Zhou, L., Zhang, R.H., 2023. A self-attention–based neural network for three-dimensional multivariate modeling and its skillful enso predictions. Science Advances 9, eadf2827.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: section abstract and introduction in 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we discuss this in Section A

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: this paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: as shown in Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, data, and checkpoints will be released publicly upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: as shown in Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, reporting error bars is too computationally expensive and is not standard in data-driven ocean forecasting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify this part in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: : The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of the paper in Section B

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used and cited HYCOM-RD, an open-source dataset, and the validation framework IV-TT.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Limitations

Despite these promising results, our approach has several limitations that warrant further investigation. The reliance on reanalysis data for training, while necessary given data availability constraints, introduces potential biases from the underlying numerical models. This challenge is compounded by the scarcity of publicly available high-quality sub-daily ocean reanalysis data. In future work, we plan to address this limitation by integrating HYCOM analysis fields with Argo observational data and other in-situ measurements to create a more robust dataset. For subsurface validation, this will involve a rigorous data matching process, where observational profiles are spatio-temporally collocated with the model grid points, and quality control filters are applied to remove outliers and inconsistent measurements before comparison. The filters use HYCOM analysis field data as an anchor to remove outlier data with high variability. Additionally, our evaluation thus far focus primarily on standard RMSE metrics, which, while useful for comparing forecast systems, may not fully capture the model's ability to represent specific ocean phenomena like mesoscale eddies, western boundary currents, or seasonal thermocline transitions. Future analysis will incorporate phenomenon-specific evaluation metrics and physical consistency measures to provide deeper insights for model improvement. Finally, while our system's forecast accuracy remains robust through 10 days, performance for longer-term predictions (seasonal to interannual) remains unexplored, representing another important direction for future research.

B Broader Impacts

FuXi-Ocean represents a significant advancement with far-reaching implications across multiple domains. By enabling high-resolution, sub-daily ocean forecasting at global scales, our work creates opportunities for numerous scientific and societal applications. In the realm of maritime safety and operations, the 6-hour temporal resolution provides critical advantages for shipping navigation, offshore energy production, and search and rescue missions. Particularly for emergency response scenarios such as oil spill tracking and marine accident response, the ability to forecast ocean currents at 6-hour intervals significantly improves source identification and trajectory prediction, potentially saving lives and reducing environmental damage. For marine resource management, FuXi-Ocean can enhance fisheries operations through improved forecasting of ocean conditions that influence fish migration and aggregation patterns. The comprehensive vertical coverage (0-1500 m) is especially valuable for this application, as it captures the habitats of commercially important species that undergo diel vertical migration. For coastal communities and small island nations, FuXi-Ocean provides enhanced capability for predicting coastal hazards like storm surge, coastal flooding, and harmful algal blooms, all of which can develop rapidly and require high-frequency forecasting systems for adequate warning. The eddy-resolving capabilities of our model are particularly important for these applications, as smaller-scale coastal processes often drive the most damaging impacts. From a computational efficiency perspective, FuXi-Ocean demonstrates that high-quality forecasts can be generated with significantly reduced computational resources compared to traditional numerical models. This efficiency makes sophisticated ocean forecasting more accessible to researchers and institutions with limited computational infrastructure, potentially democratizing access to advanced oceanographic tools.

However, our work also presents potential risks that warrant careful consideration. Relying on deep learning models for critical ocean forecasting applications requires robust verification systems, particularly when extending predictions to new or unprecedented scenarios not represented in the training data. The reliance on reanalysis data means that biases in the underlying numerical models could be perpetuated or even amplified by our approach. Additionally, as with any forecasting system, there is risk in over-reliance on model outputs for critical decision-making without appropriate uncertainty quantification.

C Additional results

C.1 Performance Test Supplement

We present our test results for ocean current UV in Fig.6 and 7. Similar to Fig.3 in the main text, the RMSE here is tested on the HYCOM reanalysis data for 2015. From the results, it can be seen that

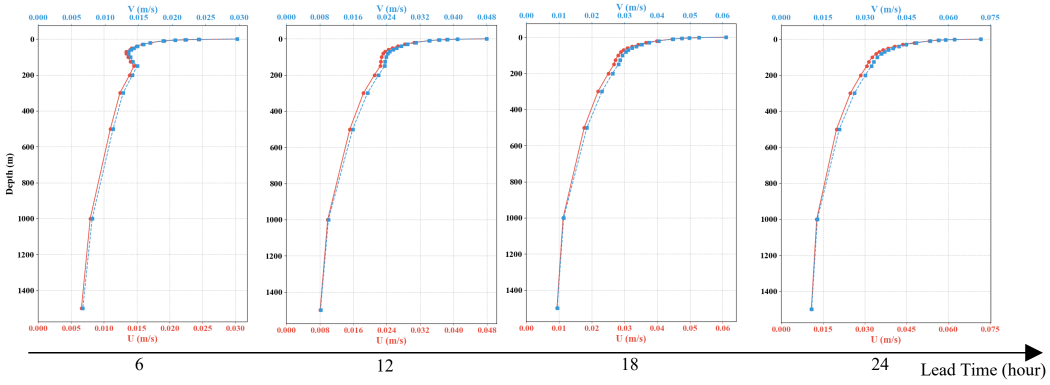

Figure 6: **Depth-dependent RMSE Distributions of ocean current UV components varying with lead time (6 - 24 hours).** Each subplot represents the RMSE (lower is better) varying with depth at the current lead time.

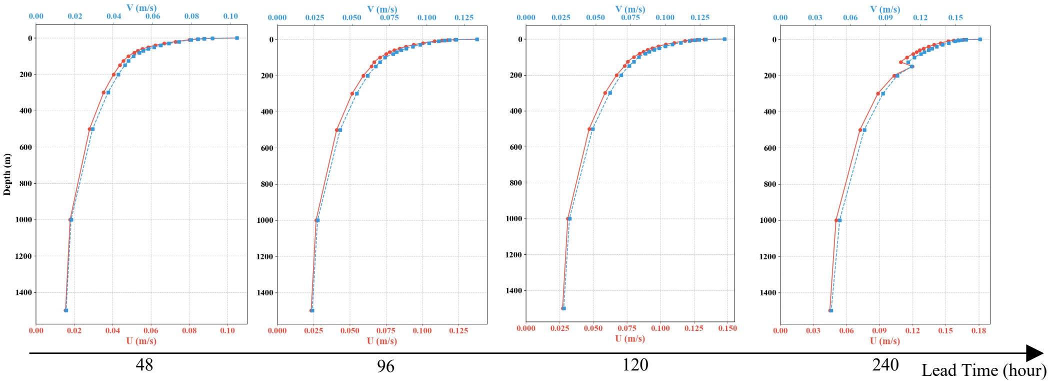

Figure 7: **Depth-dependent RMSE Distributions of ocean current UV components varying with lead time (48 - 240 hours).** Each subplot represents the RMSE (lower is better) varying with depth at the current lead time.

the RMSE values for ocean current UV maintain high accuracy and exhibit high stability in error accumulation. As lead time increases, the error growth is gradual and uniform, without significant differences between intra-day and daily intervals. Notably, in the 6-hour forecast, the stratification of change characteristics near the ocean surface is evident, indicating that the changes in ocean currents are more easily influenced by recent changes, such as sea surface winds.

C.2 Performance Test Supplement

To comprehensively evaluate the model's performance, we conducted tests on ocean current UV components under the observation data (Buoy data on CMEMS). The WenHai* and Glory* shown in the fig.8 and 9 are direct comparisons based on descriptions from paper[11]. Although this comparison may not be entirely fair (mainly due to inconsistent testing years), FuXi-Ocean's performance can still serve as a simple reference.

D Data

As shown in the Tab.1, compared with the long-term training data used by mainstream ocean models such as Xihe, Wenhai, Langya, etc., we used shorter time range data for training and achieved better prediction results. For data-driven deep learning methods, we can accurately capture ocean phenomena and features by relying solely on pure ocean variables such as T, S, U, V and SSH. **Observation Data.** For the IV-TT data, we list the available public links in Fig.2, which include salinity, temperature, and SST variables. Through experimentation, it was found that due to practical

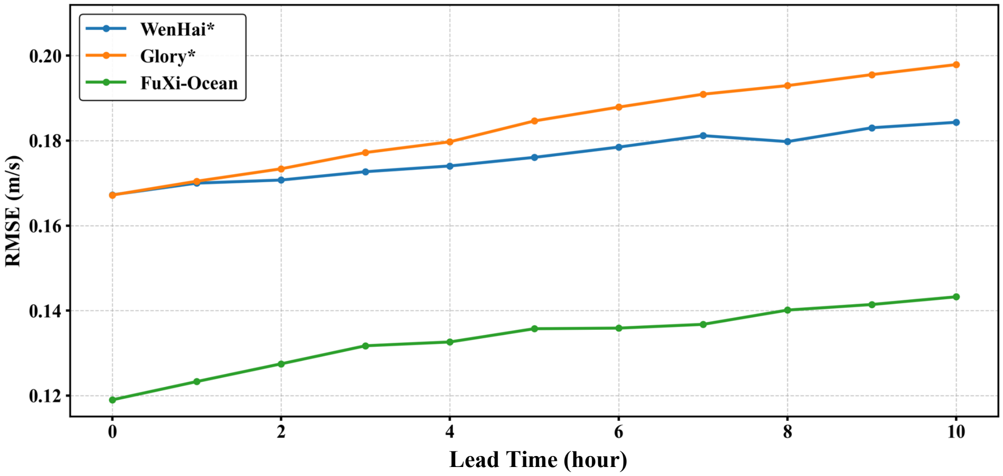

Figure 8: **The ocean current U comparison of different methods based on the observation data.** The x-axis represents the forecast time, and the y-axis represents the RMSE (lower is better). Note that the results of WenHai and Glory are derived from their papers, hence marked with an asterisk (*).

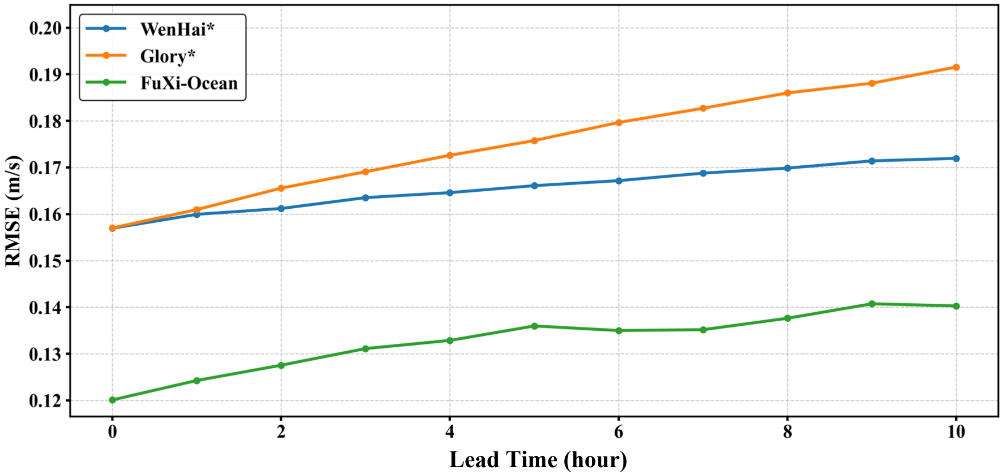

Figure 9: **The ocean current V comparison of different methods based on the observation data.** The x-axis represents the forecast time, and the y-axis represents the RMSE (lower is better). Note that the results of WenHai and Glory are derived from their papers, hence marked with an asterisk (*).

issues, only partial data is available. Firstly, due to the presence of outliers in the temperature and salinity vertical profile data from IV-TT, these anomalies significantly impact the evaluation results, causing the errors to be abnormally large. Secondly, because the ocean contains thermoclines and haloclines, where temperature and salinity change rapidly and non-linearly, using IV-TT to perform linear interpolation in the vertical direction to align with the model grid introduces excessive errors, thereby reducing the reliability of the evaluation results. Therefore, we ultimately use SST for

Table 1: Comparison of Training Data for Existing Ocean Models

Name of the Model	Training Data Time Range	Input Variables
Xihe	1993-2020	SST, U10, V10, T, S, U, V, SSH
Langya	1993-2021	SST, U10, V10, T, S, U, V, SSH
WenHai	1993-2020	SST, U10, V10, T, S, U, V, SSH
FuXi-Ocean	2006-2015	T, S, U, V, SSH

Table 2: The source of observation data

Variable	Data type	Source of data
T	In situ Argo profiles	Argo GDAC (from http://www.usgodae.org/argo/argo.html or http://www.coriolis.eu.org/)
S	In situ Argo profiles	Argo GDAC (as above)
U/V	In situ Argo profiles	Argo, Ocean Sites, GOSUD, EGO https://data.marine.copernicus.eu/product/INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030
SLA	Satellite altimeter from Jason-1, Jason-2 and Envisat	CLS Aviso Level 3 data
SST	In situ surface drifter data	From USGODAE server: http://www.usgodae.org/cgi-bin/datalist.pl?dset=fnmoc_obs_sfcobs&summary=Go

evaluation and comparison.

The observational data from ARGO is not gridded, and its degree of discreteness is shown in Fig.10. Globally, the distribution of stations at key locations is relatively uniform and comprehensive.

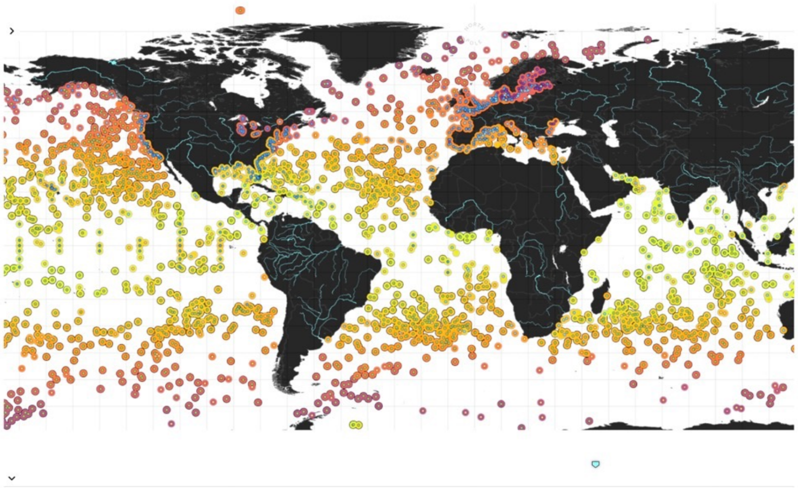

Figure 10: **Global distribution map of ARGO and other buoys.**

E Supplementary details

In practice, we set $K = 1$ to achieve the best performance of the MoT module. In the module design, we assume that there will always be an optimal moment of information that can optimize the results. Of course, if the number of historical moments involved, N, increases significantly, $K = 1$ is unlikely to be a good choice.