

- 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: The paper does not involve human subjects and therefore does not require IRB approval.

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 rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper does not incorporate large language models (LLMs) as part of its core methodology. LLMs were not used in the development, analysis, or implementation of any scientific methods or experiments.

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.

References

- [DSFNMO] DSFNMO. Godae sfcobs surface temperature observations (1998–present).
- [2] Aouni, A. E., Gaudel, Q., Regnier, C., Van Gennip, S., Drevillon, M., Drillet, Y., and Lelouche, J.-M. (2024). Glonet: Mercator’s end-to-end neural forecasting system. *arXiv preprint arXiv:2412.05454*.
- [3] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2022). Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*.
- [4] Bloom, S., Takacs, L., Da Silva, A., and Ledvina, D. (1996). Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124(6):1256–1271.
- [5] Brasseur, P. and Verron, J. (2006). The seek filter method for data assimilation in oceanography: a synthesis. *Ocean Dynamics*, 56:650–661.
- [6] 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(1):2268.

- [7] Dai, A., Qian, T., Trenberth, K. E., and Milliman, J. D. (2009). Changes in continental freshwater discharge from 1948 to 2004. *Journal of climate*, 22(10):2773–2792.
- [8] Dee, D. P., Uppala, S., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al. (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597.
- [9] Divakaran, P., Brassington, G., Ryan, A., Regnier, C., Spindler, T., Mehra, A., Hernandez, F., Smith, G., Liu, Y., and Davidson, F. (2015). Godae oceanview inter-comparison for the australian region. *Journal of Operational Oceanography*, 8(sup1):s112–s126.
- [10] Dobricic, S. and Pinardi, N. (2008). An oceanographic three-dimensional variational data assimilation scheme. *Ocean modelling*, 22(3-4):89–105.
- [11] DSCMS. Global ocean - near real-time (nrt) in situ quality controlled observations.
- [12] DSCMS. Global ocean along-track sea surface height anomalies (sla) from altimeter satellites.
- [DSMOI] DSMOI. Insitu_glo_phy_uvassim_discrete_nrt_013_054: Near-real-time drifter velocity product (filtered and assimilated, irregular time).
- [14] Hernandez, F., Bertino, L., Brassington, G., Chassignet, E., Cummings, J., Davidson, F., Drévilon, M., Garric, G., Kamachi, M., Lellouche, J.-M., et al. (2009). Validation and intercomparison studies within godae. *Oceanography*, 22(3):128–143.
- [15] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049.
- [16] Jiang, N., Zhu, C., Hu, Z.-Z., McPhaden, M. J., Chen, D., Liu, B., Ma, S., Yan, Y., Zhou, T., Qian, W., et al. (2024). Enhanced risk of record-breaking regional temperatures during the 2023–24 el niño. *Scientific Reports*, 14(1):2521.
- [17] Johnson, J. E., Febvre, Q., Gorbunova, A., Metref, S., Ballarotta, M., Le Sommer, J., and fablet, r. (2023). Oceanbench: The sea surface height edition. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78275–78295. Curran Associates, Inc.
- [18] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2022). Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.
- [19] Lellouche, J.-M., Greiner, E., Ruggiero, G. and Bourdallé-Badie, R., Testut, C.-E., Le Galloudec, O., and Benkiran, M. G. G. (2023). Evolution of the copernicus marine service global ocean analysis and forecasting high-resolution system: potential benefit for a wide range of users.
- [20] Madec, G. et al. (2015). Nemo ocean engine.
- [21] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.
- [22] Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. (2020). Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203.
- [23] Rasp, S., Hoyer, S., Meroze, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al. (2024). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019.
- [24] Rozier, D., Birol, F., Cosme, E., Brasseur, P., Brankart, J.-M., and Verron, J. (2007). A reduced-order kalman filter for data assimilation in physical oceanography. *SIAM review*, 49(3):449–465.
- [25] Ryan, A., Regnier, C., Divakaran, P., Spindler, T., Mehra, A., Smith, G., Davidson, F., Hernandez, F., Maksymczuk, J., and Liu, Y. (2015). Godae oceanview class 4 forecast verification framework: global ocean inter-comparison. *Journal of Operational Oceanography*, 8(sup1):s98–s111.
- [26] Tranchant, B., Testut, C.-E., Ferry, N., and Brasseur, P. (2006). Sam2: The second generation of mercator assimilation system. *European Operational Oceanography: Present and Future*, 650:650–655.

- [27] 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*.

A Appendix: ML Models and Observational Datasets

This appendix provides a detailed overview of the machine learning (ML) models benchmarked within the OceanBench framework, including their architectural designs, training protocols, and forecasting strategies. These models span a range of neural approaches tailored for global ocean prediction and are evaluated across the three benchmarking tracks defined in this study: models-to-analysis, models-to-reanalysis, and models-to-observation. In addition to describing the models, this appendix introduces the datasets used for the models-to-observations track, which assesses model forecast skill directly against independent, near-real-time ocean observations. This observation-based validation, grounded in the IV-TT Class-4 framework (14; 25; 9), provides an operationally relevant benchmark by leveraging high-quality drifters measurements.

A.1 GLONET, A Neural Global Ocean Forecasting System

GLONET (2) is a data-driven, global ocean forecasting model targeting short-range (10-day) predictions of key ocean state variables, including 3D temperature and salinity, sea surface height (SSH), and surface currents. It operates at a horizontal resolution of $1/4^\circ$ with 21 vertical levels, and is trained on GLORYS12 reanalyses interpolated to the target grid. The architecture follows a hybrid neural operator design that fuses multiple modeling paradigms. Large-scale patterns (e.g., gyres, equatorial currents) are captured using Fourier Neural Operators (FNOs), while CNNs enhance representation of finer-scale dynamics. A hierarchical transformer backbone models long-range spatial dependencies, particularly important for resolving land–sea boundaries and coastal complexities. The model follows an encoder-decoder structure that integrates multi-scale spatial and temporal features into a coherent latent representation. GLONET employs an autoregressive forecasting strategy over 10 days, where predictions are recursively used as inputs for subsequent steps. It does not perform online data assimilation; instead, it leverages the observational constraints already embedded in the GLORYS12 reanalyses used during training. For initialization, GLONET uses daily near-real-time analyses from GLO12. The system produces daily 10-day forecasts as daily mean fields, and outputs are experimentally distributed via the European Digital Twin Ocean (EDITO) platform. In OceanBench, GLONET represents the flagship AI-based model benchmarked against the physics-based GLO12 system, providing insight into the performance of deep learning approaches in operational forecasting contexts.

A.2 XiHe: A Global Ocean Eddy-Resolving Forecasting System

XiHe (27) is a data-driven global ocean forecasting model designed to capture mesoscale and large-scale dynamics with high spatial resolution. It operates at $1/12^\circ$ horizontal resolution with 23 vertical levels and is trained on 25 years of daily GLORYS12 reanalyses, enriched with near-surface wind fields from ERA5 and high-resolution SST from OSTIA. At its core, XiHe employs a hierarchical transformer architecture tailored for ocean forecasting. Custom ocean-specific self-attention blocks capture both local and global spatial dependencies, enabling the model to represent regional variability and inter-basin teleconnections. A land–ocean mask is applied to restrict learning to oceanic regions, improving spatial focus and reducing boundary artifacts. XiHe adopts a modular, temporally stratified design: 20 independent transformer-based models are trained separately for each forecast day (1 to 10) and vertical region (upper or lower ocean). This setup avoids the error accumulation common in autoregressive strategies and allows each model to specialize in lead-time– and depth-specific dynamics.

A.3 WenHai: Forecasting the Eddying Ocean with a Deep Neural Network

WenHai (6) is a global eddy-resolving ocean forecasting model based on deep learning, designed to predict upper ocean dynamics at $1/12^\circ$ resolution across 23 vertical levels. Trained on 25 years of daily GLORYS12 reanalyses and ERA5 atmospheric forcings, WenHai focuses on mesoscale features such as eddies and sharp thermohaline gradients. Rather than predicting ocean state variables directly, WenHai forecasts their daily tendencies changes in temperature, salinity, sea surface height (SSH), and surface currents, which are applied recursively to update the ocean state over a 10-day forecast horizon. This tendency-based, autoregressive formulation emphasizes learning temporal dynamics. The model architecture is built on the Swin Transformer, leveraging localized self-attention to capture long-range spatial dependencies. Physical priors are embedded via bulk formulae for surface fluxes of momentum, heat, and freshwater. A volume-weighted loss prioritizes upper-ocean accuracy, aligning model training with regions of strong mesoscale variability and better observational coverage.

A.4 IV-TT Class-4 Observation Dataset

The Class-4 framework, developed by the Intercomparison and Validation Task Team (IV-TT), defines a standardized and operationally relevant protocol for assessing ocean forecasting systems within the observation space (14; 25; 9). By directly comparing model outputs to near-real-time, independent observations at coincident spatial and temporal locations, this approach enables an unbiased evaluation of forecast skill across multiple variables and lead times, ranging from day 0 (best analysis) to 10-day forecasts.

Adopted in OceanBench, this framework complements analysis- and reanalysis-based evaluations by anchoring performance assessment in real observations, thereby supporting both scientific benchmarking and operational utility. The observational period considered spans the year 2024, with all datasets produced in near-real-time mode, thus aligning with the Class-4 philosophy of independence, timeliness, and applicability to operational oceanography.

The following observation datasets are employed for Class-4 validation:

- **Surface currents:** Validated against Lagrangian drifter velocities from INSITU_GLO_PHY_UVASSIM_DISCRETE_NRT_013_054 (DSMOI), which provides quality-controlled, near-real-time measurements from the global drifter array.
- **Temperature and salinity vertical profiles:** Sourced from the Argo program via INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030 (11), which offers multi-depth, multi-parameter observations from autonomous profiling floats.
- **Sea level anomalies (SLA):** Evaluated using gridded satellite altimetry from SEALEVEL_GLO_PHY_L3_NRT_008_044 (12), a Level 3 near-real-time product merging multiple satellite tracks.
- **Sea surface temperature (SST):** Assessed using in-situ measurements from the FNMOC GODAE SFCOBS dataset (DSFNMOC), distributed via the GODAE Monterey Server and compiled from ships, moored and drifting buoys, and Coastal-Marine Automated Network (CMAN) stations.

B Appendix B: Derived Physical Diagnostics

This appendix outlines the methodology used to compute key derived quantities for process-oriented evaluation of ocean forecasts. These diagnostics, Mixed Layer Depth (MLD), geostrophic currents, and Lagrangian trajectories serve as physically meaningful benchmarks for assessing the internal consistency and dynamical realism of model outputs. While not directly optimized during training, these variables are inferred from predicted state fields (e.g., temperature, salinity, sea surface height, velocity) and thus provide a stringent test of whether neural forecasting systems capture the underlying physical processes of the ocean. The following subsections detail the mathematical formulations and computational procedures used to derive each diagnostic.

B.0.1 Mixed Layer Depth (MLD)

MLD is a key indicator of ocean vertical mixing and stratification. Accurately predicting MLD is essential for simulating air-sea interactions, heat exchange, and biological productivity. MLD is derived from forecasted temperature and salinity profiles and is commonly defined based on a density threshold criterion, such that the mixed layer is the depth at which the density difference from the surface equals a specified threshold. The MLD can be approximated as:

$$\text{MLD} = \min \{z \mid \rho_z - \rho_0 \geq \Delta\rho\} \quad (3)$$

where ρ_z represents the density at depth z , ρ_0 is the density at the surface, and $\Delta\rho$ is a threshold value typically set to a small increment (e.g., 0.03 kg/m^3) to capture the mixed layer's depth relative to surface conditions.

B.0.2 Geostrophic Currents

Derived from sea surface height, geostrophic currents provide a diagnostic of large-scale ocean circulation. Accurate prediction of these currents is critical for understanding ocean transport and dynamics. Geostrophic currents are derived from forecasted SSH under the geostrophic approximation:

$$\mathbf{v}(\phi, \theta, t) = gf^{-1}\nabla^\perp\eta(\phi, \lambda, t) \quad (4)$$

where g is the acceleration of gravity, f presents the Coriolis coefficient, and $\eta(\phi, \lambda, t)$ is the sea surface height (SSH), which serves as a noncanonical Hamiltonian for surface velocity. \perp stands for a 90° anticlockwise rotation of the gradient vector, producing a perpendicular flow direction as dictated by geostrophic balance.

B.0.3 Lagrangian Trajectory

Lagrangian drift analysis offers insight into a model’s ability to capture the advection of ocean particles over time, which is critical for applications involving transport processes such as pollutant dispersion, larval connectivity, and passive tracer dynamics. By simulating the motion of synthetic particles advected by model-predicted velocity fields, we assess whether the flow structures are coherent and physically realistic. Let’s consider the ocean currents field:

$$\mathbf{v}(\mathbf{x}, t), \quad \mathbf{x} \in \mathbb{R}^2, \quad t \in [t_0, t_f] \quad (5)$$

and its associated ordinary differential equation:

$$\dot{\mathbf{x}} = \mathbf{v}(\mathbf{x}, t), \quad \mathbf{x} \in \mathbb{R}^2, \quad t \in [t_0, t_f] \quad (6)$$

where \mathbf{v} the U and V components of ocean currents, defined on a possibly time-dependent spatial domain $\mathcal{U}(t) \in \mathbb{R}^2 \times [t_0, t_f]$.

Lagrangian trajectories are defined as:

$$\mathbf{x}(t_f, t_0, \mathbf{x}_0) = \mathbf{x}_0 + \int_{t_0}^{t_f} \mathbf{v}(\mathbf{x}(\tau), \tau) d\tau \quad (7)$$

To quantitatively evaluate the fidelity of Lagrangian trajectories, we compute the Euclidean distance between model-predicted and reference (GLORYS12) particle positions at each time step. It is expressed in kilometers and averaged over all particles:

$$\text{Lagrangian drift deviation}(t) = \frac{1}{N} \sum_n^N \left| \mathbf{x}_i^f(t) - \mathbf{x}_i^r(t) \right| \quad (8)$$

This metric provides a time-resolved diagnostic of trajectory divergence, helping identify whether modeled flow fields maintain coherent transport pathways. Small Lagrangian errors suggest a physically plausible flow structure, which is particularly important for data-driven models not constrained by conservation laws.

C Appendix C: Benchmark Track Results and Model Intercomparison

This appendix presents a consolidated analysis of model performance across the two core benchmarking tracks defined in OceanBench: models-to-analysis and observations-to-analysis. It brings together a comprehensive intercomparison of forecasting approaches, examining their behavior across spatial and temporal scales through a range of qualitative and quantitative diagnostics. The goal is to provide deeper insight into the strengths and limitations of each model in capturing ocean dynamics, fostering a more nuanced understanding of their generalization ability under realistic forecasting scenarios.

C.1 Models-to-Observations Track

The models-to-observations track provides a direct evaluation of forecast skill against independent in situ and satellite observations. Among the assessed variables, surface ocean currents, evaluated at a reference depth of 15 meters, exhibit notable skill differentials between modeling approaches. ML-based models demonstrate superior performance in this regime, with GLONET in particular achieving consistently lower errors relative to both traditional physics-based and other ML-based models. This improved performance likely stems from the capacity of ML models to capture advective structures and mesoscale variability present in historical training data.

The skill observed in surface velocity fields is mirrored to some extent in sea level anomaly (SLA) forecasts, which are used to derive geostrophic surface currents. Certain ML models also achieve

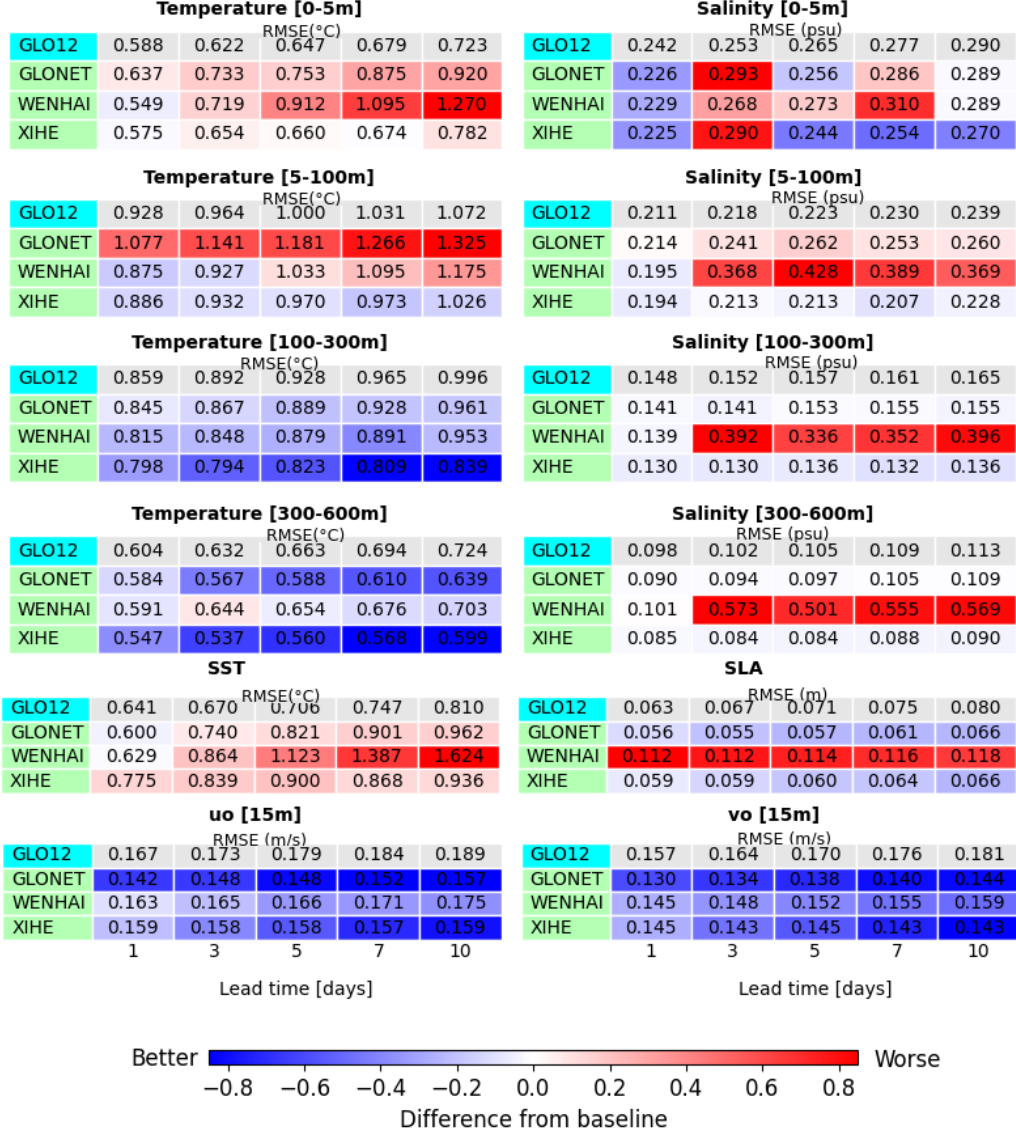


Figure 3: Models to observations track.

competitive performance in SLA prediction, indicating an emerging ability to learn coherent surface dynamics from data alone, though the degree of success varies across architectures.

Forecast skill in temperature and salinity fields reveals distinct depth-dependent behavior. For temperature, ML performance tends to improve with depth, particularly at intermediate levels (e.g., 100–300 m), where thermocline structure is both stable and predictable. This suggests that data-driven models can internalize persistent stratification patterns when supported by sufficient historical context. In contrast, salinity forecasts tend to degrade with increasing depth, likely reflecting the more heterogeneous and patchy nature of salinity fields, which pose greater challenges for interpolation and learning. While regional breakdowns of performance are not available in the present analysis, it is reasonable to hypothesize that ML-based models gains are more pronounced in regions characterized by high mesoscale activity and dense observation coverage. Further spatial disaggregation would be required to confirm such patterns.

C.2 Models-to-Analysis Track

The models-to-analysis track evaluates model forecasts against the GLO12 analysis. Unsurprisingly, all models underperform relative to the GLO12 baseline in this track, as the reference analysis is itself

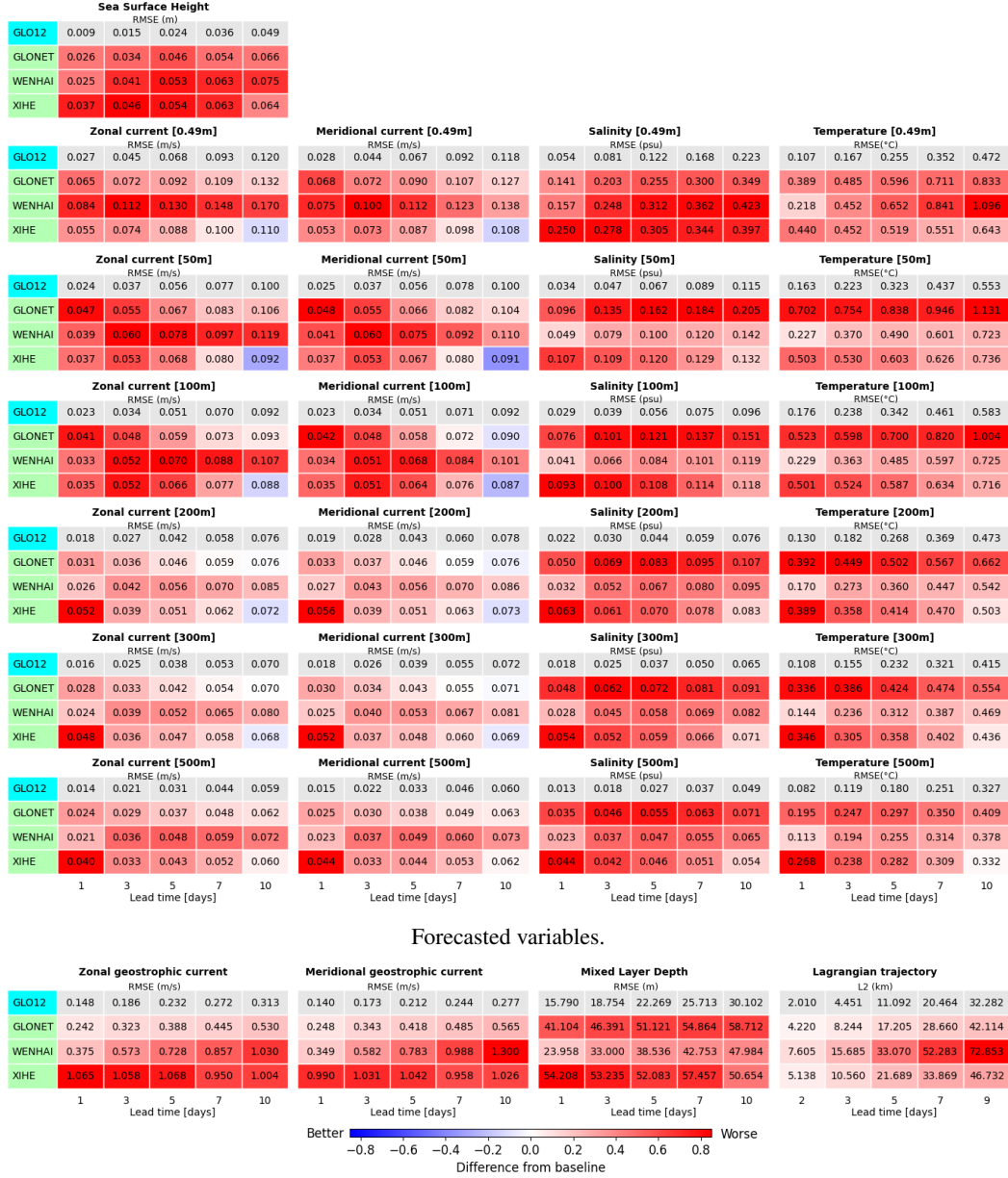


Figure 4: Models to analysis track.

generated from the GLO12 forecast model. Specifically, the GLO12 analysis is produced through a weekly data assimilation cycle applied to GLO12 forecasts, meaning it inherits both the dynamical structure and biases of the underlying numerical model. This setup creates a structural advantage for GLO12-consistent models, and correspondingly poses a higher bar for systems that diverge in design. This structural bias is clearly reflected in the results: Wenhai, which incorporates physically inspired components such as bulk formulae for surface forcing, exhibits error evolution patterns that closely mirror those of GLO12, particularly for surface currents.

Such similarity suggests that shared physical assumptions lead to convergent dynamical behavior under this evaluation framework. In contrast, more flexible data-driven models tend to display different error trajectories, with some demonstrating improved accuracy at longer lead times, potentially due to a reduced coupling with the reference model's assimilation dynamics. For scalar variables such as temperature and salinity, however, no clear systematic trends emerge across models. This may reflect the more complex vertical structure and reduced observational constraint at depth, which

weaken the influence of both physical priors and learned data patterns in shaping model skill under this benchmark.

In summary, the models-to-analysis track is best interpreted as a measure of structural consistency with the GLO12 system, rather than as an unbiased indicator of real-world forecast skill. It complements the observation-based track by revealing how different model classes align or diverge from an established operational baseline, and underscores the importance of using multiple benchmarks to robustly assess forecast performance across frameworks.

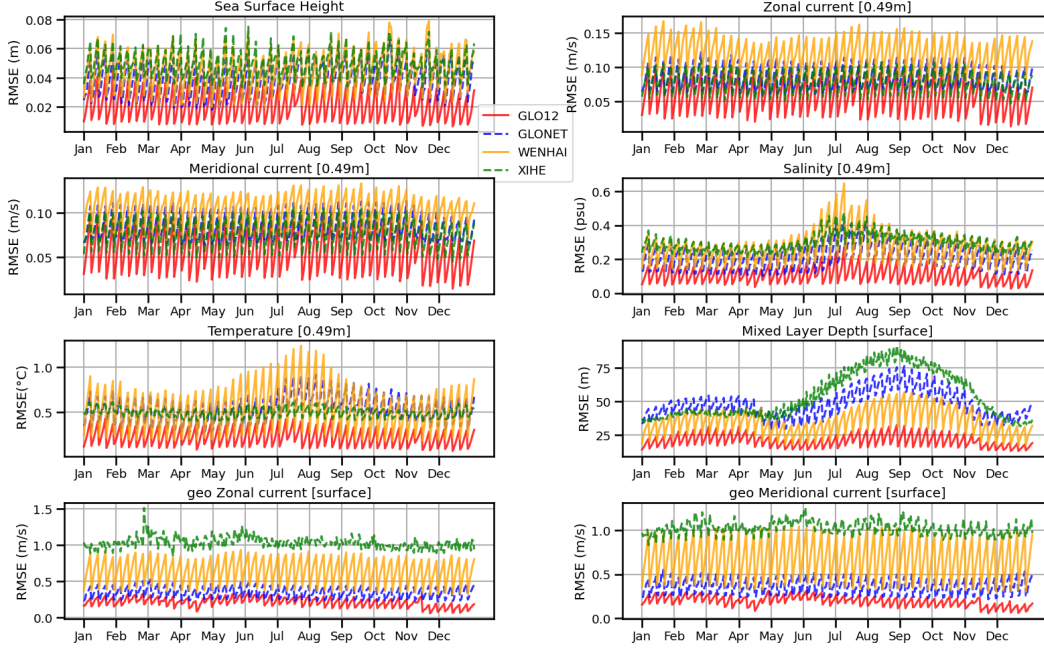


Figure 5: Time series of RMSE evolution throughout 2024 for all models in the Models-to-Analysis track.

C.2.1 Temporal Structure of Forecast Errors

To complement the overall skill metrics reported in the Models-to-Analysis Track, we analyze the temporal evolution of model performance over the full calendar year of 2024 as shown in Figure 5. This evaluation provides a time-resolved perspective on how forecast accuracy evolves in relation to both seasonal and sub-seasonal variability, using daily 7-day RMSE scores stratified by variable type. Compared to the Reanalysis Track, models appear more tightly clustered in performance, both at large scales (seasonal modulations) and at higher frequencies, where alignment with the weekly forecast cycle of the GLO12 analysis becomes apparent.

Geostrophic Currents and Dynamical Fields. For geostrophic surface currents, the spread in error among models is markedly reduced compared to the Reanalysis Track. This convergence may reflect the structural imprint of the GLO12 forecast system on the analysis product, which effectively narrows the range of permissible dynamical behaviors. While short-term error fluctuations are still observed, likely tied to the 7-day assimilation cycle, the relative ranking of models remains consistent with the Reanalysis evaluation.

Temperature and Salinity. Seasonal modulation in tracer forecast errors is also diminished relative to the Reanalysis Track. Whereas clear annual cycles were previously observed, particularly strong in machine learning models, temperature and salinity RMSEs now exhibit weaker amplitude and reduced variability across models. Notably, the GLO12 baseline displays little to no seasonal pattern, likely due to its assimilation-driven correction toward climatological states. This damping effect appears to propagate into the analysis, thereby reducing the sensitivity of evaluation metrics to seasonal forcing signals. As a result, the Models-to-Analysis Track offers a more constrained and homogenized assessment of model fidelity, shaped in part by the characteristics of the reference itself.

D Spatial Structure and Scale-Resolved Evaluation

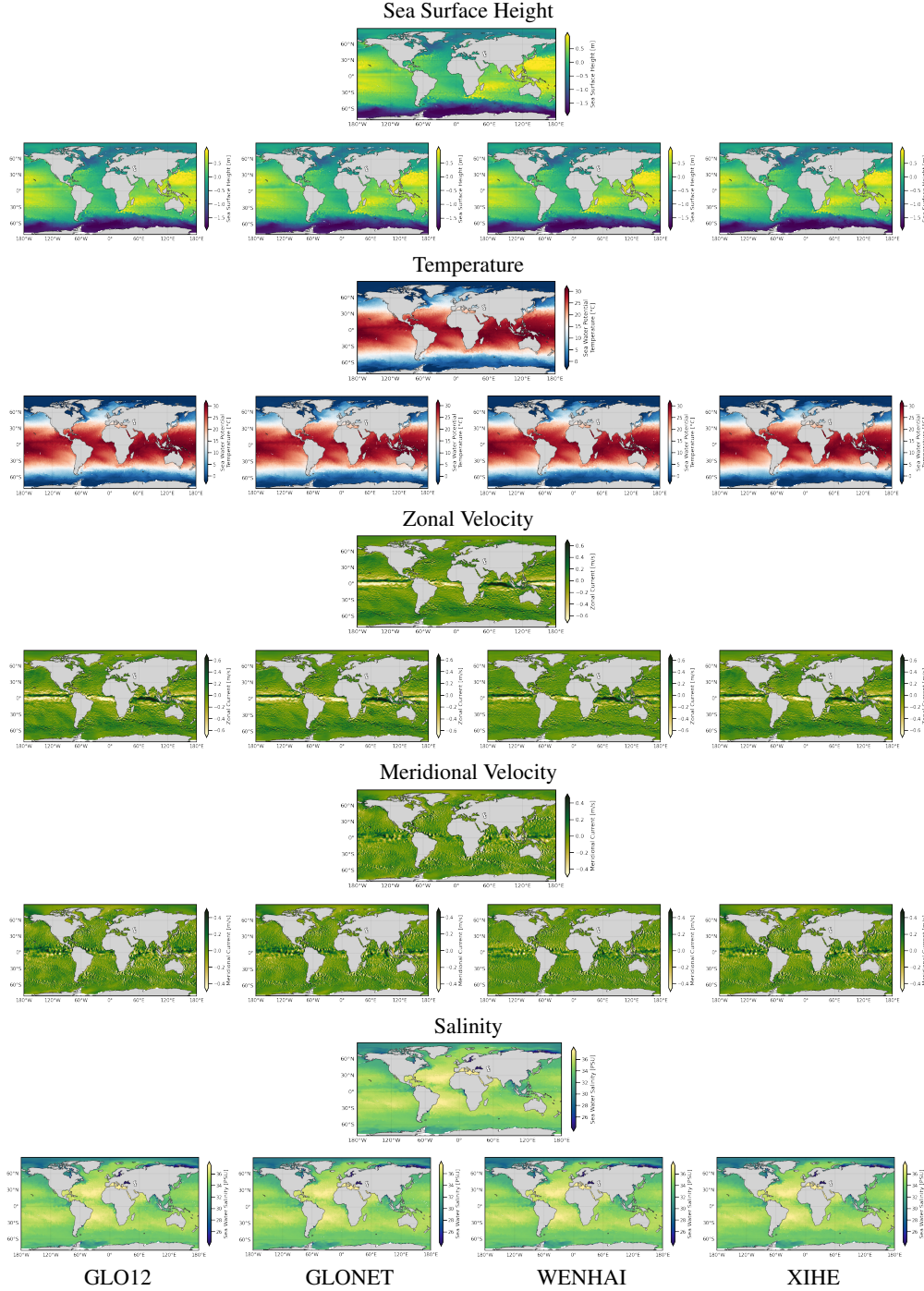


Figure 6: A set of results for world map for the forecasted variables. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity All of these are for lead time 1 for the date 2024-01-03.

Beyond aggregate scores and temporal trends, spatial diagnostics provide essential insight into the qualitative behavior of ocean forecasting models. This section presents a series of spatially explicit analyses that complement the benchmark metrics by offering a visual and scale-aware assessment of model fidelity. These diagnostics not only reveal how errors manifest across different oceanographic regimes but also highlight the structural differences in model output, particularly in terms of resolved spatial scales and noise characteristics.

D.1 Visual comparison of model outputs.

Qualitative analysis of the model outputs reveals a high degree of similarity in the spatial distribution and dynamical structure across all evaluated systems (see Figure 6). Core ocean state variables including sea level anomaly, surface temperature, salinity, and surface currents, exhibit coherent mesoscale features and basin-scale gradients that are well captured by all models, despite differences in resolution or architectural design. This convergence suggests a shared ability to reconstruct the dominant patterns of ocean variability present in the reanalysis datasets used during training or evaluation. Notably, planetary-scale wave structures are clearly visible in the surface velocity fields of some of the models, closely resembling those observed in the reference systems. These large-scale features, which are often indicative of baroclinic and barotropic wave dynamics, are generally more difficult to capture in data-driven models but appear to be well preserved across the ensemble. Their presence points to a broader capacity among models to internalize low-frequency, dynamically consistent patterns, even in the absence of explicit physical constraints or assimilation cycles. Such visual coherence provides a qualitative complement to quantitative metrics and reinforces the notion that evaluated models reproduce not only the mean state but also the spatial structure of ocean circulation.

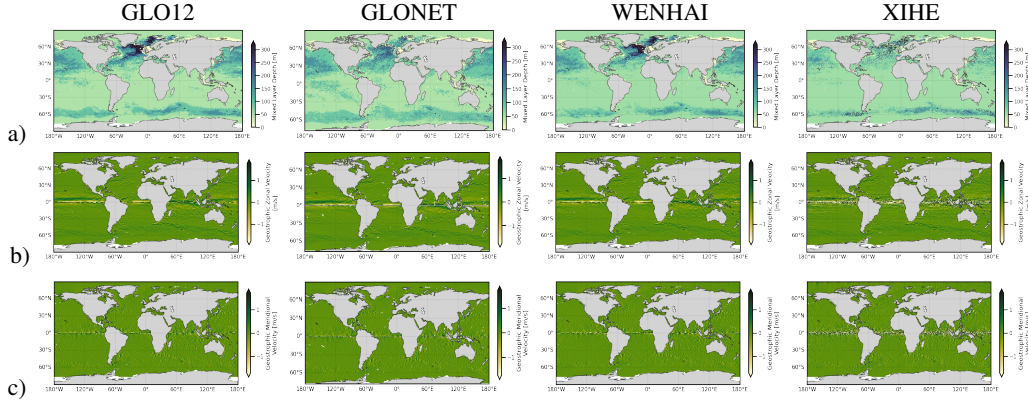


Figure 7: A set of results for world map for diagnostic variables. The following physical variables are as follows: a) Geostrophic Zonal Velocity, b) Geostrophic Meridional Velocity, and c) Mixed Layer Depth. All of these are for lead time 1 on the date 2025-01-02.

When examining derived diagnostic variables such as geostrophic surface currents and mixed layer depth (MLD), the model outputs reveal varying degrees of structural coherence and persistence of artifacts (see Figure 7). In general, most models demonstrate a consistent alignment between forecasted dynamical fields and their diagnostics, suggesting a reasonable preservation of physical dependencies across variables. However, notable differences emerge based on model architecture and forecasting strategy. Xihe, employing non-autoregressive forecasting strategy, tend to exhibit reduced spatial coherence and increased noise, particularly evident in fragmented geostrophic current patterns and highly irregular MLD fields. Wenhai also shows some edge-related noise in meridional geostrophic velocities near the northern and southern boundaries, but produces MLD fields that are well-structured and closely aligned with the reference GLO12. These patterns underscore the varying sensitivity of diagnostic outputs to architectural design, and highlight the value of visual diagnostics in assessing the internal physical consistency of model forecasts.

Spatial distribution of errors. To better understand the regional distribution of forecast skill and the origin of discrepancies, we present spatial maps of root mean square error (RMSE) relative to the GLORYS12 reference (see Figures 8–12). These maps reveal a striking degree of consistency in the geographical structure of forecast errors across models, particularly for sea surface height and salinity. Elevated errors are systematically found in western boundary current systems, equatorial regions, and zones of intense mesoscale activity, areas that are difficult to forecast due to their high dynamical variability and sensitivity to initial and boundary conditions.

While these broad spatial patterns are largely shared, noticeable inter-model differences are observed in the velocity components and temperature fields. For zonal and meridional currents, the RMSE distributions vary in both magnitude and slightly in localization, reflecting differences in how models represent and propagate dynamical features. Similarly, temperature fields exhibit some variability in error structure, likely tied to each model’s handling of thermal gradients and stratification processes.

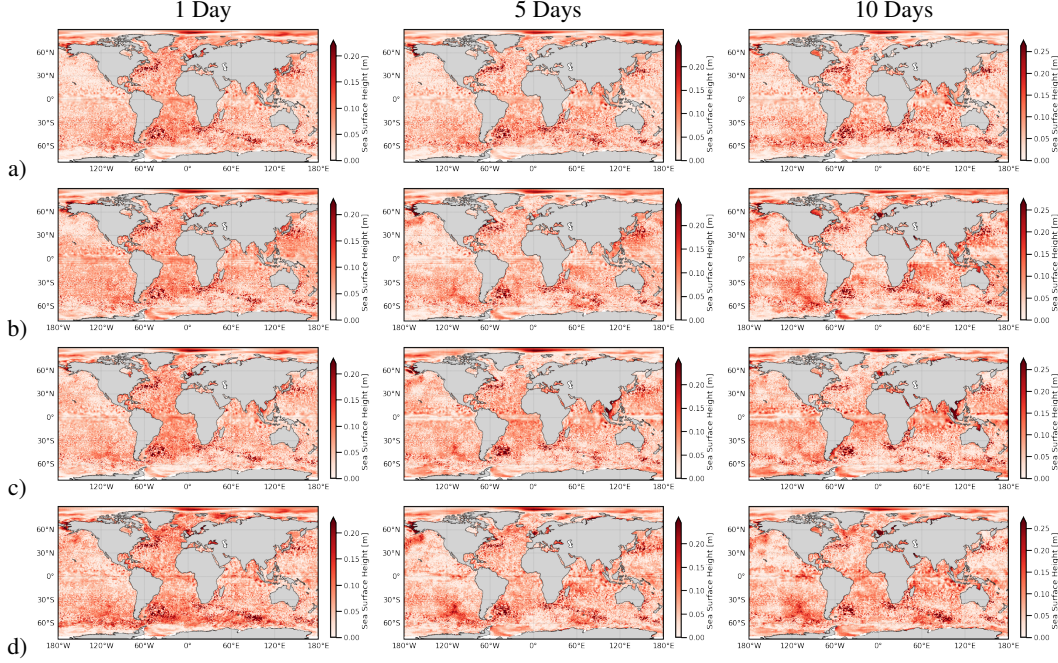


Figure 8: Error Maps of Root Mean Squared Error as a function of lead time for Sea Surface Height. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

These variable-dependent differences suggest that, although models contend with common physical and observational constraints, their ability to represent ocean dynamics and thermodynamics diverges in meaningful ways. Overall, the spatial diagnostics reinforce the robustness of the benchmarking framework while highlighting the importance of evaluating model skill across individual physical variables.

D.2 Power Spectral Density (PSD) Analysis

To quantitatively assess the models' ability to reproduce ocean variability across spatial scales, we analyze the power spectral density (PSD) of predicted oceanographic fields. PSD offers a robust scale-resolved diagnostic that complements RMSE and visual inspection by identifying noise artifacts, structural inconsistencies, and dynamical fidelity in model outputs.

Given a two-dimensional spatial field $f(x, y)$ defined on a regular grid, we compute its isotropic PSD as follows. First, we remove the spatial mean to eliminate the zero-frequency component: $\tilde{f}(x, y) = f(x, y) - \bar{f}$. We then apply a two-dimensional discrete Fourier transform (2D-DFT) to obtain the spectral coefficients:

$$\hat{f}(k_x, k_y) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \tilde{f}(x, y) \exp \left(-2\pi i \left(\frac{k_x x}{N_x} + \frac{k_y y}{N_y} \right) \right), \quad (9)$$

where k_x and k_y are the zonal and meridional wavenumbers, respectively. The two-dimensional PSD is given by the squared magnitude of the Fourier coefficients:

$$\text{PSD}(k_x, k_y) = |\hat{f}(k_x, k_y)|^2. \quad (10)$$

To obtain a one-dimensional isotropic spectrum, we perform radial averaging in spectral space by binning values according to the radial wavenumber $k = \sqrt{k_x^2 + k_y^2}$. This results in $\text{PSD}(k)$, a spectrum that describes the distribution of variance across spatial scales, enabling direct comparison of model performance in resolving fine to coarse features.

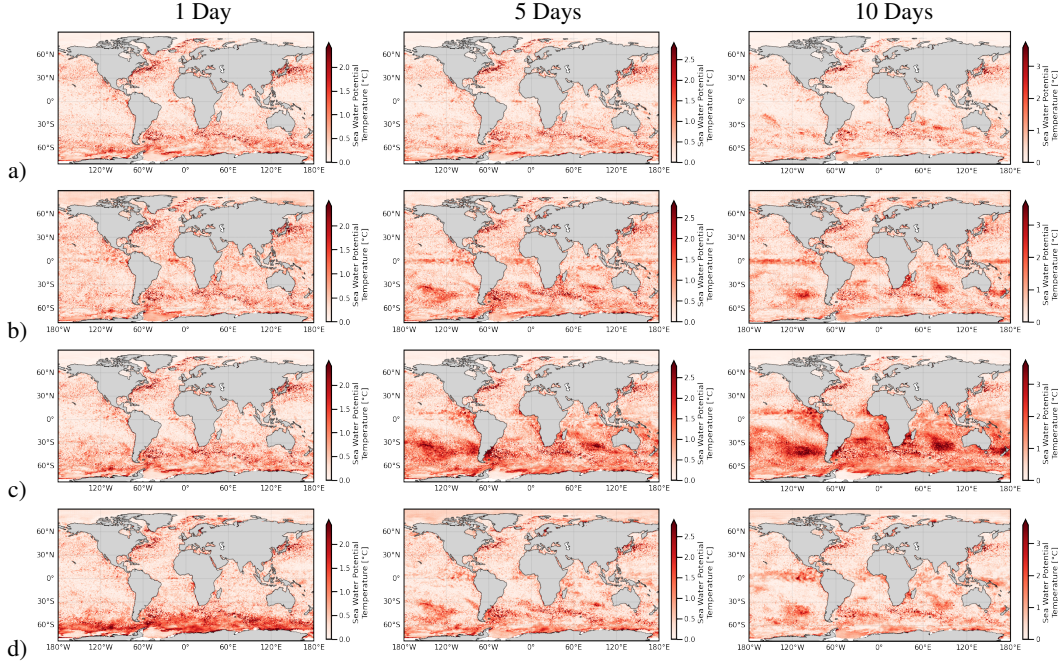


Figure 9: Error Maps of Root Mean Squared Error as a function of lead time for Temperature. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

Global-scale analysis. At the global scale and short lead time (day 1), the PSDs reveal consistent spectral patterns that distinguish the forecasting systems (see Figure 13). Wenhai stands out across most variables by exhibiting elevated spectral plateaus at high wavenumbers, indicative of pervasive high-frequency noise and a lack of effective small-scale filtering. This characteristic suggests that Wenhai’s outputs are generally noisier and less dynamically coherent at fine scales, even at early lead times. In contrast, Xihe shows more variable behavior: while its spectral decay in scalar fields such as temperature and salinity is more moderate, its forecasts of vector quantities, specifically zonal and meridional velocities exhibit pronounced short-wavelength artifacts. These directional inconsistencies point to a model architecture that struggles to resolve or stabilize fine-scale dynamical structures, particularly in the representation of currents. Together, these global PSD diagnostics underscore the importance of physical constraints in mitigating high-wavenumber noise, even at the outset of the forecast horizon.

Regional-scale analysis. The computed PSDs over the gulf stream region (Figure 14) reveal important differences across systems and lead times. For sea surface height (SSH), GLO12 and models architecturally aligned with it exhibit the expected monotonic spectral decay, consistent with geophysical fluid dynamics. In contrast, both Xihe and Wenhai display oscillatory behavior at short wavelengths, suggestive of unresolved dynamics or spurious high-frequency noise. These discrepancies are further amplified at longer lead times (e.g., days 5 and 10), where the reduction in fine-scale energy becomes more pronounced. While such decay is a natural consequence of forecast uncertainty, the extent of energy loss and spectral distortion is particularly notable in certain ML-based systems.

Similar spectral anomalies are observed in the temperature field, where Xihe shows both a reduced overall spectral power and pronounced short-scale oscillations, indicating a degradation in multi-scale thermal fidelity. The decline in spectral energy with lead time is consistent across systems but disproportionately affects models with weaker physical priors.

These trends extend to the zonal and meridional velocity components, where Xihe continues to exhibit the lowest spectral energy and elevated high-frequency artifacts. The salinity spectra follow a similar pattern, reinforcing the finding that some architectures are more prone to high-wavenumber noise and less capable of preserving physical structure over time.

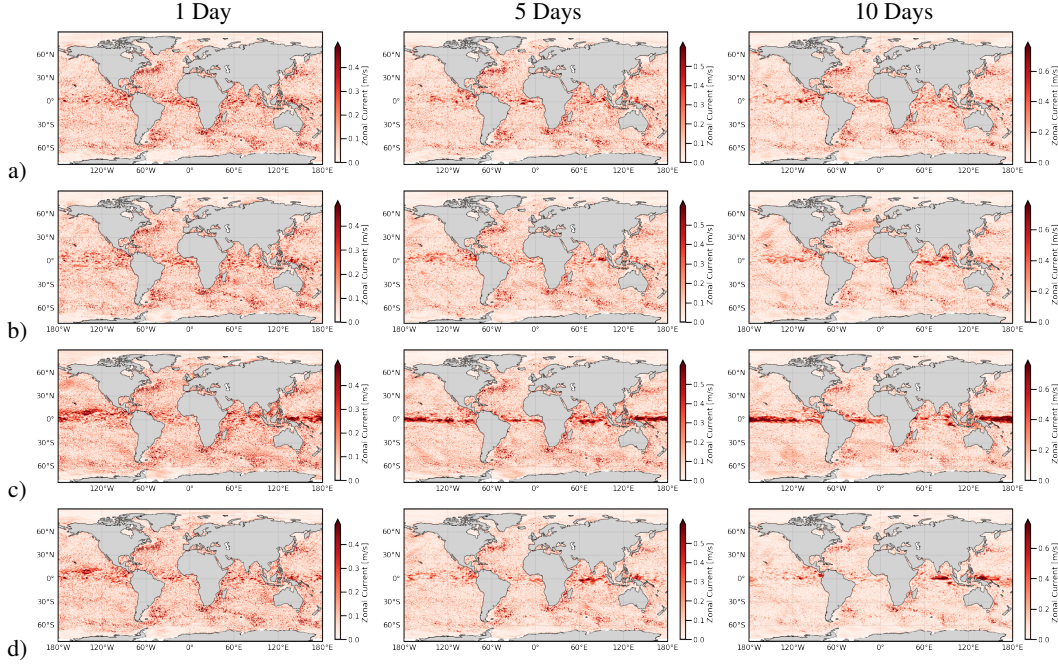


Figure 10: Error Maps of Root Mean Squared Error as a function of lead time for the Zonal Velocity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

Overall, the PSD analysis provides a rigorous and interpretable framework for evaluating the scale-resolving skill of forecasting systems. It highlights the robustness of physically grounded models in maintaining spectral coherence and exposes the challenges that certain ML-based alternatives face, particularly at extended lead times in preserving energy across scales.

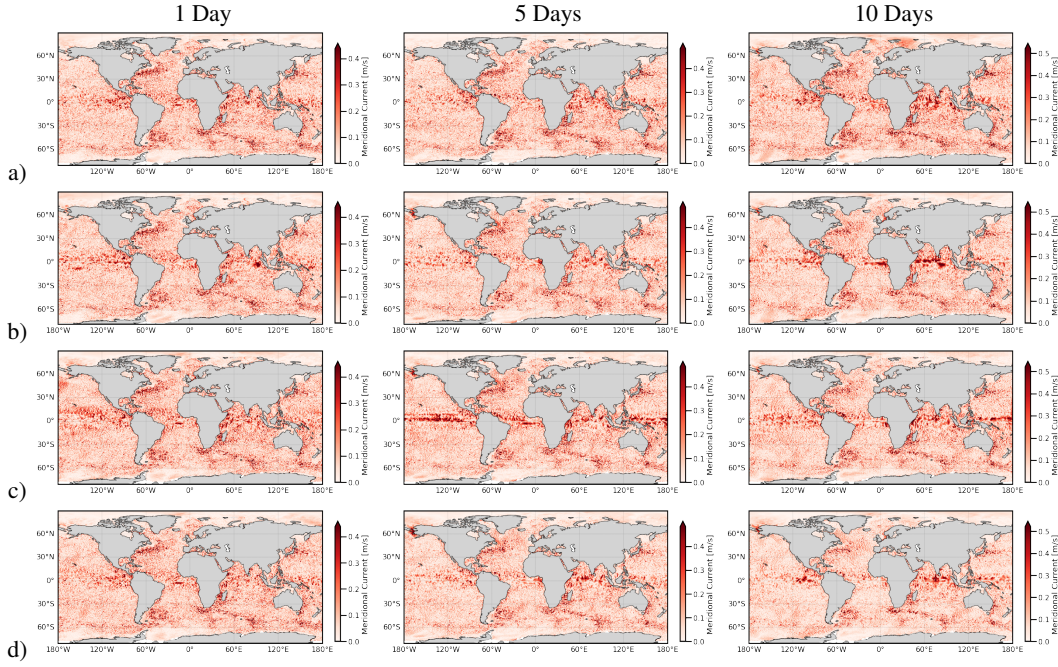


Figure 11: Error Maps of Root Mean Squared Error as a function of lead time for the Meridional Velocity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

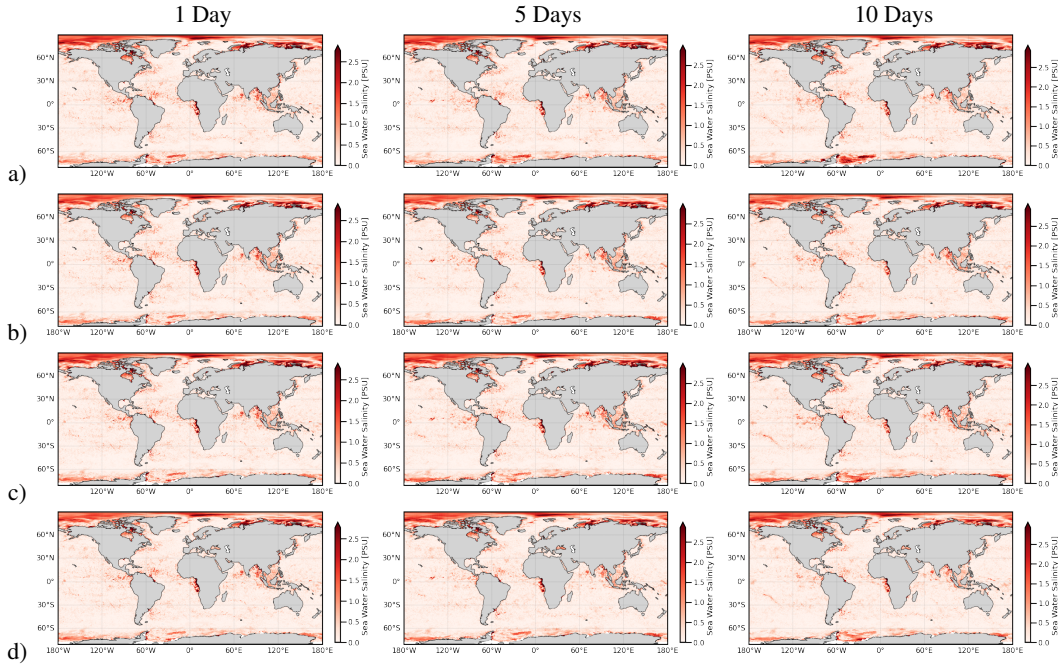


Figure 12: Error Maps of Root Mean Squared Error as a function of lead time for Sea Surface Salinity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

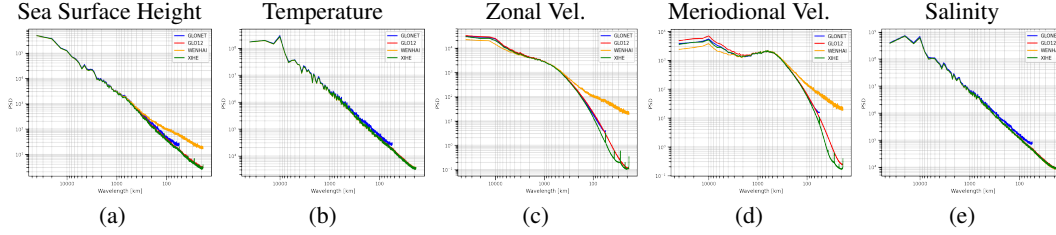


Figure 13: A set of results for the power spectrum for the zonal direction averaged over the latitude and time(2024) for the whole globe. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity. This figure only shows a lead time of 1.

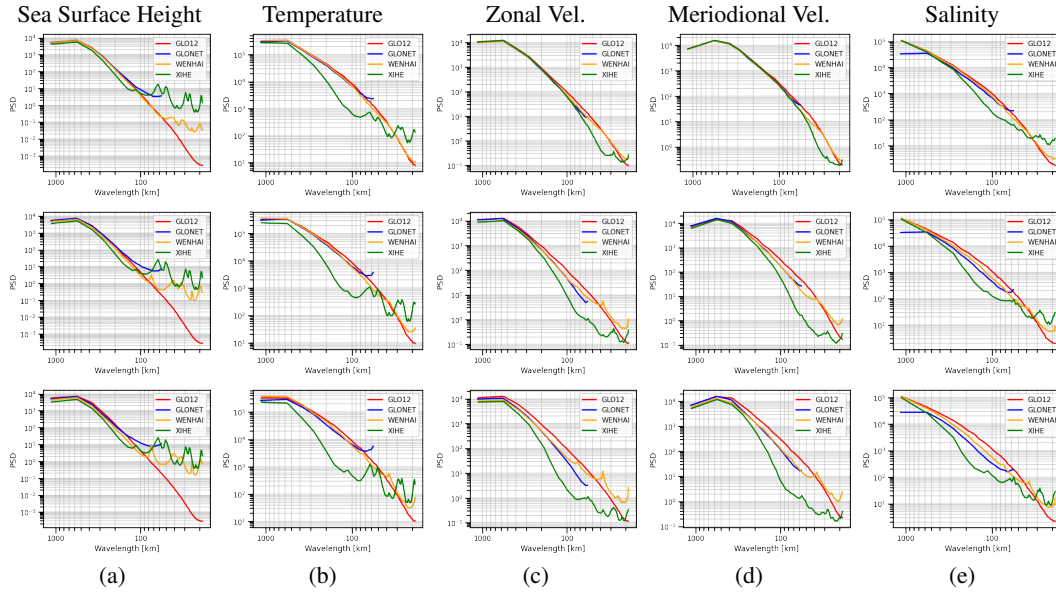


Figure 14: A set of results for the power spectrum for the zonal direction averaged over the latitude and time(2024) over the Gulf Stream. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity. Rows 1, 2, and 3 are the lead times of 1 day, 5 days, and 10 days respectively.