OceanBench: A Benchmark for Data-Driven Global Ocean Forecasting systems

Anass El Aouni[1]* **Quentin Gaudel**[1] **Juan Emmanuel Johnson**[2]

Charly Regnier[1] **Julien Le Sommer**[3] **Simon Van Gennip**[1] **Ronan Fablet**[4]

Marie Drevillon[1] **Yann Drillet**[1] **Pierre-Yves Le Traon**[1]

[1]Mercator Ocean International, Toulouse, France
[2]International Methane Emissions Observatory, UNEP, Paris, France
[3]Université Grenoble Alpes, CNRS, Grenoble, France
[4] IMT Atlantique, Brest, France

Abstract

Data-driven approaches, particularly those based on deep learning, are rapidly advancing Earth system modeling. However, their application to ocean forecasting remains limited despite the ocean's pivotal role in climate regulation and marine ecosystems. To address this gap, we present OceanBench, a benchmark designed to evaluate and accelerate global short-range (1-10 days) data-driven ocean forecasting. OceanBench is constructed from a curated dataset comprising first-guess trajectories, nowcasts, and atmospheric forcings from operational physical ocean models, typically unavailable in public datasets due to assimilation cycles. Matched observational data are also included, enabling realistic evaluation in an operational-like forecasting framework. The benchmark defines three complementary evaluation tracks: (i) Model-to-Reanalysis, where models are compared against the reanalysis dataset commonly used for training; (ii) Model-to-Analysis, assessing generalization to a higher-resolution physical analysis; and (iii) Model-to-Observations, Intercomparison and Validation (IV-TT) CLASS-4 evaluation against independent observational data. The first two tracks are further supported by process-oriented diagnostics to assess the dynamical consistency and physical plausibility of forecasts. OceanBench includes key ocean variables: sea surface height, temperature, salinity, and currents, along with standardized metrics grounded in physical oceanography. Baseline comparisons with operational systems and state-of-the-art deep learning models are provided. All data, code, and evaluation protocols are openly available at https://github.com/mercator-ocean/oceanbench, establishing OceanBench as a foundation for reproducible and rigorous research in data-driven ocean forecasting.

1 Introduction

Global ocean forecasting is a cornerstone of Earth system prediction. Accurate forecasts of ocean dynamics underpin a broad range of societal and scientific needs, from climate monitoring and carbon budgeting to maritime safety, fisheries management, and disaster response. Like its atmospheric counterpart, ocean forecasting has historically been dominated by numerical models based on physics that solve the governing equations of fluid motion, thermodynamics, and biogeochemistry.

*Correspondence to: aelaouni@mercator-ocean.eu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

These models have matured significantly over recent decades and now offer multiscale forecasts at increasingly high resolution. However, they remain computationally intensive, dependent on complex data assimilation workflows, and slow to iterate with typical development cycles that stretch over years. Moreover, despite steady gains, key physical processes such as submesoscale turbulence, deep convection, or eddy-mean flow interactions remain underresolved or poorly parameterized.

By contrast, data-driven ocean prediction remains in its infancy. This is due in part to the inherent complexity of ocean dynamics, the challenges of sparse and heterogeneous observations, and a historical lack of investment relative to atmospheric science. These factors have contributed to the slower adoption of machine learning (ML) methods in operational ocean forecasting, despite their recent success in other Earth system domains. Contributing to this gap is also the absence of a standardized benchmark which hinders progress by making it difficult to assess model skill, diagnose failure modes, or identify promising approaches.

To address this gap, we introduce OceanBench: a community benchmark for global short-range ocean forecasting (1-10 days). OceanBench is built on curated datasets from operational forecasting systems, including model first-guesses [2], nowcasts, and associated atmospheric forcings data which is typically omitted from public reanalyses. It also integrates matched observational datasets from both satellite and in-situ sources, enabling evaluation in a realistic real-time forecasting framework. The benchmark defines two complementary evaluation tracks: a model-to-observation track using standard skill scores and a model intercomparison track that supports traditional and process-oriented diagnostics. Finally, OceanBench includes baseline results from both operational systems and recent ML-based models, along with open-source tools for data access, training, and evaluation. By establishing a standardized and extensible framework, OceanBench aims to foster reproducibility, encourage collaboration, and accelerate progress in AI-enabled ocean forecasting.

2 Related work

In recent years, the success of deep learning in atmospheric sciences has spurred interest in data-driven Earth system prediction. Enabled by advances in model architecture and the availability of large-scale reanalysis datasets, several ML models have achieved competitive or superior performance compared to traditional numerical weather prediction (NWP) systems. Notable examples include GraphCast, Pangu-Weather, and FourCastNet, which demonstrate strong forecasting skill across various deterministic and probabilistic metrics, particularly for medium-range forecasts (1-10 days) [22, 3, 25]. Much of this progress has been driven by standardized benchmarks such as WeatherBench and its successor WeatherBench2, which provide high-quality datasets, reproducible pipelines, and unified evaluation protocols [26, 27]. These benchmarks have been instrumental in enabling rapid model iteration, fair comparison, and transparent reporting of progress.

In the ocean domain, machine learning applications have gained traction more slowly. The inherent complexity of ocean dynamics, including nonlinear interactions across a wide range of spatial and temporal scales, poses unique challenges for data-driven modeling. These challenges are compounded by limited observational coverage, especially below the surface, and by the diversity of forcing mechanisms such as wind stress, tides, and boundary currents that influence ocean behavior. Nevertheless, recent work has begun to demonstrate the viability of ML for global-scale ocean forecasting. Emerging global models like GLONET, XiHe, and WenHai [2, 32, 6], along with regional models [19, 14], have shown that it is possible to learn from historical analyses and reanalyses, or even directly from observations to generate skillful forecasts of key ocean state variables [16]. These models operate efficiently at inference time and are often focused on short-range forecasting, where satellite observations are more abundant and predictability is higher.

Despite recent momentum, progress in ocean ML forecasting has been limited by the absence of shared datasets and evaluation protocols. Framework "OceanBench: The SSH Edition" had a first iteration which attempted to standardize benchmarking for ocean observations and simulations but they only focused on a very small subset of variables, i.e., sea surface height and sea surface temperature, over a very small region, i.e., Gulf-stream [21]. OceanBench builds on the principles established by domain-specific efforts (such as "OceanBench: The SSH Edition") with the completeness and

[2] In operational oceanography and data assimilation, the first-guess (also known as the background) is the short-term forecast from a previous model run, typically used as the initial estimate of the ocean state before assimilating new observational data. It represents the model's best estimate of the current conditions based solely on past information and dynamical evolution. During assimilation, this first-guess is combined with newly available observations to produce an updated analysis, or nowcast, that better reflects the true state of the ocean.

robustness of broader benchmarks (such as WeatherBench2), adapting them to the ocean domain through curated datasets, standardized metrics, and baseline models. By doing so, it is a foundational resource for the growing ocean forecasting community, supporting reproducible research, consistent and fair comparison across both physical and machine learning models, and diagnostic evaluation tailored to the unique challenges of ocean dynamics. As more ML and hybrid models are developed, they can be integrated into the benchmark over time.

3 Core OceanBench Forcing and Evaluation Systems

This section introduces the forecasting systems and data sources that underpin the OceanBench benchmark: 1) the initialization, 2), the baselines, and 3) the reference datasets. OceanBench evaluates both physics-based models, which constitute the foundation of operational oceanography (introduced in Appendix A), and data-driven models that learn to predict ocean dynamics directly from historical data using machine learning models. These systems vary in spatial resolution, temporal coverage, and forecasting methodology and encompass both autoregressive and direct prediction approaches. All models are initialized from a common ocean state and are subjected to the same atmospheric forcing fields, ensuring fair and synchronized forecast launches. Model performance is evaluated using a consistent set of reference data, which includes both reanalysis products, providing the best available estimates of the full 3D ocean state, and operational analysis fields, which offer high-frequency, observation-constrained estimates used in real-time systems. In addition, observational data sets support resolution-agnostic comparisons through standardized evaluation protocols inspired by operational oceanography. This unified framework combines physics-based models, machine learning forecasts, and reanalysis and analysis references, which support rigorous, reproducible, and comparative evaluation of global ocean forecasting systems. A complete description of the candidate models for machine learning, as well as the observation data used for evaluation, is provided in Appendix B.

3.1 GLORYS12 Global Ocean Physics Reanalysis: Training and Reference

GLORYS12 is a global eddy-resolving ocean and sea ice reanalysis produced by the Copernicus Marine Environment Monitoring Service (CMEMS), spanning from 1993 to the present. It is based on the NEMO ocean model [24], configured at a horizontal resolution of $1/12^\circ$ (approximately 8km) with 50 vertical levels. The vertical grid is refined near the surface, with 22 levels in the upper 100m to better resolve upper-ocean variability. Atmospheric forcing is provided by the ECMWF (European Centre for Medium-Range Weather Forecasts) reanalyses, using ERA-Interim for earlier years and transitioning to ERA5 for recent periods [8, 18]. The data assimilation methodology combines a reduced-order Kalman filter [28], which ingests altimeter sea level anomalies, satellite sea surface temperatures, sea ice concentrations, and in situ temperature and salinity profiles, with a three-dimensional variational (3D-VAR) scheme [23] to correct large-scale temperature and salinity biases. GLORYS12 outputs daily and monthly mean fields of temperature, salinity, currents, sea level, mixed layer depth, and sea ice parameters on a regular global grid at $1/12^\circ$ resolution. Within OceanBench, it is used both to train AI-based forecasting models and to evaluate their performance. Its high resolution, physical consistency, and integration of observations make it a strong reference for benchmarking both data-driven and physics-based systems.

3.2 GLO12 Global Analysis and Forecast System: Nowcast, Baseline, and Reference

GLO12 is an operational global ocean forecasting system built on the same NEMO-LIM3 configuration as GLORYS12, employing a $1/12^\circ$ (8 km) horizontal grid, 50 vertical levels, and enhanced surface resolution to resolve upper-ocean and mixed-layer dynamics. Bathymetry is constructed using a blend of ETOPO1 and GEBCO8 datasets for depths between 200-300 m. Unlike GLORYS12, which is a delayed-mode reanalysis, GLO12 operates in near-real time. It is driven by high-resolution ($1/10^\circ$) ECMWF IFS forecasts that provide momentum, heat, and freshwater fluxes at sub-daily frequency. Tidal forcing is incorporated using prescribed constituents, and river discharge accounts for input from over 100 major rivers and polar ice sheet meltwater [7]. The GLO12 system employs the SAM2 data assimilation framework [31], combining a 4D SEEK filter [5], Incremental Analysis Updates (IAU) [4], and 3D-Var bias correction [10]. Observations assimilated include ODYSSEA SST, OSI SAF sea ice, AVISO sea level anomalies, and CORIOLIS in situ profiles. Deep-ocean constraints (below 2000 m) rely on WOA2013 climatology with non-Gaussian error formulations, and mass conservation is enforced via a global sea surface height constraint.

GLO12 produces both near-real-time analyses and 10-day forecasts on a rolling basis. Each week, a nowcast (i.e., ocean analysis field approximately 1-8 days behind real time) is generated and used to initialize daily-updated forecasts. The most recent 11-15 days are reprocessed to ensure temporal consistency across the analysis and forecast products. The outputs include 3D fields of temperature, salinity, and currents, as well as 2D fields such as sea surface height, sea ice parameters, and mixed layer depth, with daily, 6-hourly, and hourly resolutions. Within the OceanBench framework, GLO12 plays a triple role:

Nowcast (Initialization): Weekly nowcast snapshots from GLO12 are used to initialize all candidate AI forecasting models. These represent the best short-term estimate of the ocean state, incorporating both model dynamics and assimilated observations. Specifically, one nowcast per week is selected (every Tuesday, when the most accurate and up-to-date nowcast fields become available following data assimilation and quality control) to ensure consistency in model initializations across the 2024 evaluation period, while capturing seasonal variability. These nowcasts are distributed through OceanBench, as they are not available via the standard Copernicus catalog.

Forecast (Baseline): The operational 10-day forecasts issued from each nowcast serve as the baseline prediction, representing the performance of a physics-based state-of-the-art model.

Analysis (Reference): The corresponding GLO12 analysis fields are used as the reference product for evaluating forecast skill, providing an observation-constrained, quality-assured benchmark.

This unified use of GLO12 ensures that candidate models are initialized from a common ocean state, compared against the same forecast baseline, and validated against a consistent reference, thereby supporting robust, reproducible, and fair benchmarking across different modeling approaches.

3.3 IFS Atmospheric Forcing Fields: Operational Forcings

To ensure a consistent and realistic forcing environment across all candidate models, OceanBench provides operational atmospheric forecast fields from the ECMWF Integrated Forecasting System (IFS). These fields represent the actual atmospheric forecasts used in the real-time execution of the GLO12 system, thereby replicating the conditions under which operational ocean models are forced in practice. The IFS products used here correspond to the high-resolution (HRES) configuration of the ECMWF deterministic forecast, produced daily at a spatial resolution of approximately $1/10^{\circ}$. The forcings span the full set of ocean-relevant surface variables, including: 10-meter wind components (U10, V10), surface air temperature and humidity, surface pressure, downward shortwave and longwave radiation, precipitation and evaporation, zonal and meridional wind stress, surface heat flux and freshwater flux components.

In OceanBench, these 10-days IFS forecast fields are made available alongside the GLO12 nowcasts, corresponding to the same initialization time, every Tuesday the the whole period of 2024. This unified setup ensures that all data-driven models are not only initialized from the same oceanic state but are also forced with the exact same atmospheric conditions as those used in GLO12 forecasts. By standardizing both the initial state and the atmospheric forcing, this framework provides a fair and reproducible foundation for evaluating the performance of candidate models. Moreover, it offers a realistic setup aligned with current operational forecasting protocols, allowing future deployment scenarios to be mirrored during benchmark evaluation.

4 Evaluation set-up and metrics

To comprehensively assess the forecasting skill of candidate ocean models, we adopt a unified evaluation framework that is both temporally and spatially consistent. The evaluation protocol ensures that all models are assessed under the same initialization frequency and forecast horizon while respecting their native resolution to avoid unfair penalization of coarser models. Through the materialization of a harmonized challenger dataset, this setup establishes a fair and meaningful basis for comparison across a wide range of modeling approaches. Table 1 summarizes the datasets and variables used in the evaluation framework. This includes datasets used as references (i) for ML-based model trainings, (ii) as forecast inputs to generate OceanBench challenger datasets, and (iii) for the evaluations of the challenger datasets.

Building upon this foundation, we implement a multifaceted evaluation strategy (illustrated in Figure 1) that captures different aspects of model performance. This includes (i) observation-based intercomparison, (ii) reference-model benchmarking, and (iii) process-oriented diagnostics derived

Short name / Variable	Description / Source	(Spatial) dim. / type	Units	Dataset / Product ID
(A) Ocean				
Ocean spatial dimensions				
`lat`	Latitude	1D	°	GLORYS / GLO12 / Challengers
`lon`	Longitude	1D	°	GLORYS / GLO12 / Challengers
`depth`	Depth	1D	m	GLORYS / GLO12 / Challengers
Oceanbench challenger temporal dimensions				
`lead_day_index`	Lead Day Index	1D	int(1-10)	Challengers
`first_day_datetime`	Forecast First Day Datetime	1D	datetime	Challengers
Ocean state variables				
`zos`	Sea Surface Height	2D	m	GLORYS / GLO12 / Challengers
`thetao`	Temperature	3D	°C	GLORYS / GLO12 / Challengers
`so`	Salinity	3D	PSU	GLORYS / GLO12 / Challengers
`uo`	Zonal Current	3D	$\mathrm{m\,s^{-1}}$	GLORYS / GLO12 / Challengers
`vo`	Meridional Current	3D	$\mathrm{m\,s^{-1}}$	GLORYS / GLO12 / Challengers
Ocean derived variables				
	Zonal Geostrophic Current	2D	$\mathrm{m\,s^{-1}}$	
	Meridional Geostrophic Current	2D	$\mathrm{m\,s^{-1}}$	
	Mixed Layer Depth	2D	m	
	Lagrangian trajectory	2D	°	
Observations				
Surface Currents	Lagrangian drifter velocities (DSMOI)	In-situ	$\mathrm{m\,s^{-1}}$	`INSITU_GLO_PHY_UVASSIM_DISCRETE_NRT_013_054`
Temperature & Salinity Profiles	Argo profiling floats	In-situ, multi-depth	°C / PSU	`INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030`
Sea Level Anomaly (SLA)	Merged satellite altimetry	Gridded (L3)	m	`SEALEVEL_GLO_PHY_L3_NRT_008_044`
Sea Surface Temperature (SST)	FNMOC GODAE SFCOBS dataset	In-situ surface	°C	`DSFNMOC`
(B) Atmosphere				
Atmospheric forcing fields				
`sotemair`	2 m Air Temperature	2D	K	IFS
`sowinu10`	10 m Zonal Wind Component	2D	$\mathrm{m\,s^{-1}}$	IFS
`sowinv10`	10 m Meridional Wind Component	2D	$\mathrm{m\,s^{-1}}$	IFS
`sosudosw`	Downward Shortwave Radiation	2D	$\mathrm{W\,m^{-2}}$	IFS
`sosudolw`	Downward Longwave Radiation	2D	$\mathrm{W\,m^{-2}}$	IFS
`sowaprec`	Precipitation Rate	2D	$\mathrm{kg\,m^{-2}\,s^{-1}}$	IFS
`sod2m`	2 m Dew Point Temperature	2D	K	IFS
`somslpre`	Mean Sea Level Pressure	2D	Pa	IFS

Table 1: Summary of all datasets used in OceanBench, including model input variables (ocean and atmospheric fields) and observation datasets for CLASS-4 validation. All observation products correspond to near-real-time (NRT) data from 2024. Variables marked as 3D have multiple vertical levels; 2D variables are surface fields. An OceanBench challenger dataset is a multidimensional spatio-temporal datacube (five variables over 3 spatial dimensions and 2 temporal dimensions). Derived variables are computed as part of OceanBench process-oriented diagnostics. Lead day refers to the time elapsed between the model initialization and the target forecast time. Forecast first day datetimes are the 52 Wednesdays of year 2024.

from physically meaningful variables. Together, these components provide a holistic view of each model's ability to reproduce observed ocean dynamics, maintain internal physical consistency, and generalize beyond the training regime.

4.1 Evaluation window and forecast initializations

In this benchmark, we evaluate all forecasting models for the full year of 2024. This period was selected for several reasons. First, 2024 lies well beyond the training periods of most AI-based models, which were typically trained on data up to 2019. This temporal gap provides a rigorous test of the generalization capacity and robustness of the models to out-of-distribution conditions. Second, the year 2024 offers a sufficiently long evaluation window to compute reliable statistics for both pointwise and process-oriented metrics. While some phenomena, such as marine heatwaves or extreme transport events, may benefit from longer multi-year analyses, a full annual cycle remains a strong baseline for most oceanographic diagnostics. To ensure a fair comparison across models, we initialize each forecast using the same nowcast on every Tuesday in 2024. From each initialization date, all models perform a 10-day forecast, enabling a consistent temporal sampling framework. This weekly cadence strikes a balance between computational efficiency and temporal resolution for skill evaluation. This evaluation strategy allows us to assess forecast skill under realistic deployment scenarios, where models must extrapolate far beyond their training horizon. It also avoids contamination from data seen during training, ensuring that metrics reflect true predictive skill rather than overfitting to historical conditions.

4.2 Evaluation grid and domain

To ensure a fair and resolution-consistent evaluation across models of varying spatial granularity, each model is assessed at its native resolution. Rather than regridding all model outputs to a common

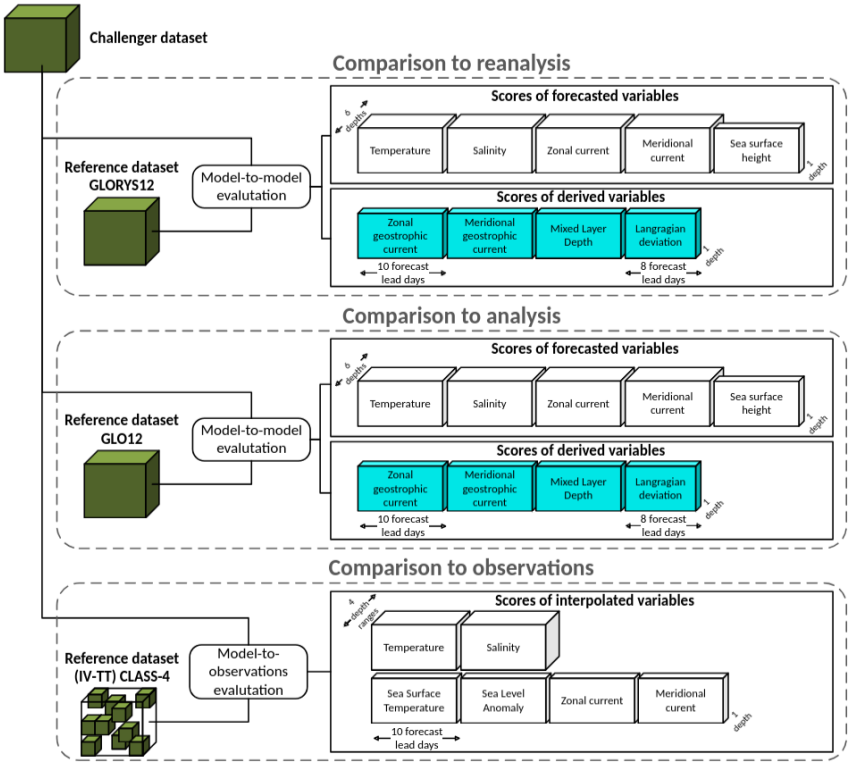


Figure 1: Overview of OceanBench's evaluation process.

target, we regrid the reference dataset, GLORYS12, to match the resolution of each challenger model. This approach avoids artificially inflating errors that could arise when comparing high-resolution truth with coarse predictions. By respecting each model's native resolution during evaluation, OceanBench avoids introducing biases and enables a clearer attribution of skill to the model architecture and training regime rather than its grid configuration. In contrast, CLASS-4 evaluation metrics [17, 29, 9] are inherently resolution-agnostic. They rely on comparisons against sparse observational data, such as along-track satellite or in-situ profiles, allowing for a direct and uniform comparison across models regardless of grid spacing. This dual evaluation strategy resolution-aware reanalysis comparison and resolution-invariant observation-based assessment provides a more balanced understanding of model performance across spatial scales, as it combines two complementary perspectives: one capturing the internal dynamical consistency of forecasts relative to a coherent reanalysis field, and the other assessing external realism against independent observations. Together, these perspectives help disentangle grid-dependent skill from true physical fidelity. All results reported in this manuscript are computed over the global ocean.

Symbol	Range	Description
f		Forecast value
o		Observation value
r		Reference (GLORYS12 or GLO12)
t	$1, \ldots, T$	Evaluation time index
l	$1, \ldots, L$	Lead time index
i	$1, \ldots, I$	Latitude index
j	$1, \ldots, J$	Longitude index
d	$1, \ldots, D$	Depth index
n	$1, \ldots, N$	Lagrangian trajectory index

Table 2: Evaluation metrics notations.

4.3 Models to observations evaluation track

The IV-TT CLASS-4 framework provides a benchmark for model validation by operating within the observation space, enabling a direct comparison between observed and modeled values across

both spatial and temporal dimensions. For each observation, the corresponding model counterpart is extracted at the same spatial and temporal location across various evaluation times and their corresponding forecast lead times, ranging from the best analysis (day 0) to ten-day forecasts. The CLASS-4 dataset includes observations of temperature and salinity from Argo profiles, sea surface temperature (SST) from surface drifting buoys, sea level anomaly (SLA) from along-track satellite measurements and surface current observations at 15m depth from GDP drifters buoys. This framework serves as a robust tool for intercomparison, facilitating a comprehensive assessment of the forecasting models' performance.

4.3.1 Root Mean Squared Error (RMSE)

RMSE is a standard metric to quantify the average magnitude of error between predicted and observed variables. It penalizes large errors more severely, making it particularly effective in identifying models that may occasionally produce large deviations from observations.

$$\mathrm{RMSE}_{(l,d)} = \sqrt{\frac{1}{TIJ}\sum_{t}^{T}\sum_{i}^{I}\sum_{j}^{J}\left(f_{t,l,i,j,d} - x_{t,i,j,d}\right)^2}, \quad x \in \{o, r\} \tag{1}$$

4.4 Models to Analysis and Reanalysis tracks

This track performs a dense, global, and vertically resolved pointwise evaluation using the GLORYS12 and GLO12 ocean reanalysis and analysis as references. Unlike the models-to-observation track, which relies on sparse observational data, it leverages the high-resolution 3D structure of these products to assess full-field accuracy.

4.5 Process-Oriented Evaluation

Beyond conventional pointwise error metrics, process-oriented evaluation provides a deeper examination of the physical consistency and generalization ability of data-driven ocean forecasting models. Unlike traditional physics-based models that encode dynamical constraints through governing equations, neural network-based systems are primarily optimized to minimize loss functions, such as RMSE, often without explicit enforcement of physical laws. While this training approach can yield high pointwise accuracy, it does not guarantee that the resulting forecasts maintain coherent and physically plausible relationships between ocean variables.

To bridge this gap, we assess whether forecasted fields can be used to derive key oceanographic quantities that reflect underlying dynamics and processes. This includes diagnostics such as mixed layer depth (MLD) and geostrophic currents; none of which were directly used during training. These derived quantities provide a stringent test of physical realism, as they depend on accurate interplay between state variables like temperature, salinity, and sea surface height. Moreover, we evaluate trajectory coherence using Lagrangian diagnostics, which test whether velocity fields support realistic particle advection patterns over time.

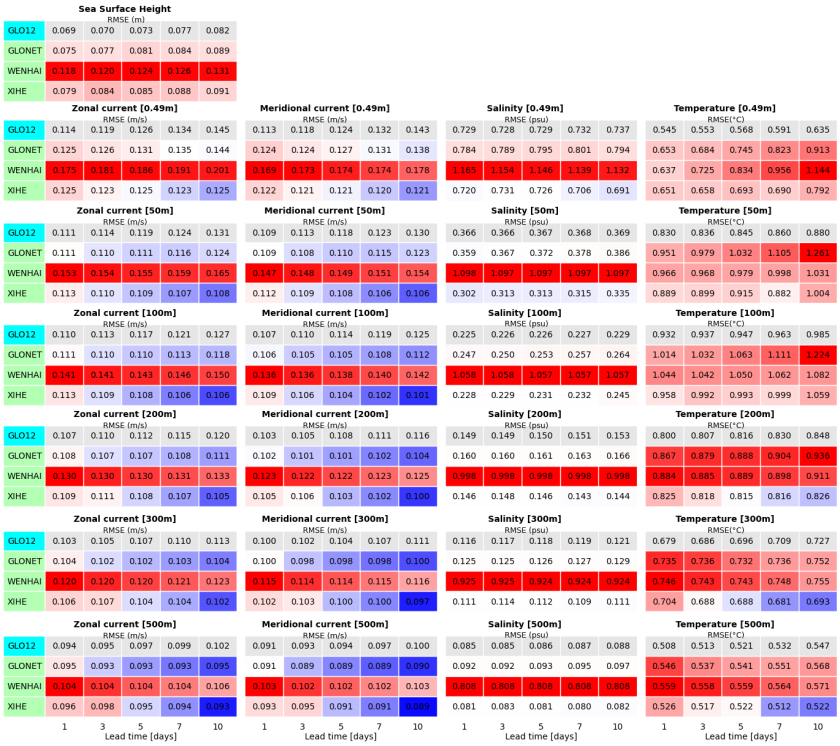
This process-oriented approach thus serves as a critical complement to pointwise metrics, enabling a holistic assessment of whether a model not only forecasts individual variables accurately but also captures the physical integrity of the ocean system it aims to simulate. Details on the computation of these derived diagnostics including MLD, geostrophic currents, and Lagrangian trajectories are provided in Appendix C.

5 Benchmark Results

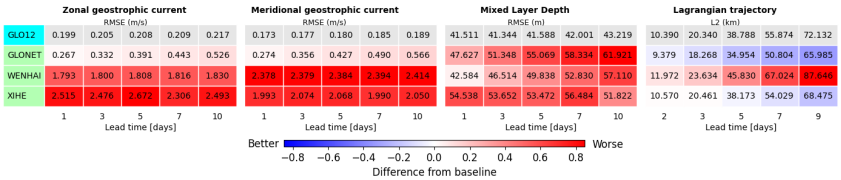
This section presents the headline results for the **Reanalysis Track**, which evaluates the short-range forecasting skill of each model against a high-quality reference reanalysis product. The summary scores, shown in Figure 2, offer a compact yet comprehensive view of model performance across key ocean state variables and diagnostic metrics. Results for the remaining benchmark tracks are provided in the Appendix D.

We assess five key physical state variables: potential temperature (thetao), salinity (so), and the zonal and meridional components of ocean velocity (uo, vo), all evaluated as three-dimensional fields at five standard depths (0.49 m, 50 m, 100 m, 200 m, 300 m, and 500 m). Sea surface height (zos) is evaluated alongside the surface-layer variables and included at the 0.49 m level to ensure

consistency in surface diagnostics. In addition to these pointwise state variables, process-oriented diagnostics are included to evaluate the physical realism and dynamical coherence of model outputs. These diagnostics comprise the mixed layer depth (MLD), geostrophic surface currents (u_{geo}, v_{geo}) derived from sea surface height gradients, and Lagrangian drifts deviation based on particle advection in surface velocity fields. The latter offers insight into the model's ability to preserve coherent flow structures and tracer transport over time. Collectively, the headline scores capture both state estimation accuracy and process-level fidelity, supporting a comprehensive evaluation of model performance across space, depth, and time.



Forecasted variables.



Diagnostic variables.

Figure 2: This table shows the absolute RMSE scores for the reanalysis track. These are the deterministic scores for the forecasted variables and physically-consistent diagnostic variables. Values show absolute RMSE. The colors denote % difference to the GLO12 baseline.

ML versus Physics-based Models. When comparing machine learning (ML) models to traditional physics-based systems, we observe that ML approaches tend to perform better on dynamical variables, particularly surface velocities (uo, vo). These variables reflect the advective structure of the flow field, which ML models appear to capture effectively from historical data. In contrast, ML models generally exhibit lower skill on scalar tracers such as potential temperature (thetao). However, their performance on salinity (so) is more variable and, in some cases, competitive or even slightly superior in specific tracks. This discrepancy may arisefrom the relatively low temporal variability and strong vertical stratification of tracers, which are more directly governed by conservation laws and vertical mixing processes; elements explicitly represented in physics-based models but only implicitly learned by ML models. Additionally, scalar tracers are more sensitive to atmospheric forcing in forecasting scenarios. Since such forcings are often more accurately represented in physical models

through prescribed boundary conditions, they may further contribute to the superior tracer predictions observed in physics-based systems.

Performance Differences Among ML Models. Among the ML-based approaches, we find significant performance variability, particularly in relation to the extent to which models incorporate explicit physical knowledge. Notably, WenHai, which includes bulk formulae and parameterizations similar to those used in operational systems, consistently ranks lowest among ML models across most metrics. This suggests that while physically inspired forcing mechanisms may add realism, they may also constrain the model's capacity to generalize or adapt to data-driven patterns when not fully integrated into the training pipeline. In contrast, more flexible data-driven architectures with less rigid physics priors achieve higher fidelity in reproducing reanalysis targets.

Forecasting Strategy: Recursive versus Direct. A key differentiator among ML models is their forecast design paradigm. XiHe, which employs a direct prediction strategy, forecasting future ocean states in a single forward pass, exhibits stable error evolution across lead times. In contrast, recursive models that predict iteratively (auto-regressively) tend to accumulate errors in a manner similar to traditional numerical forecast models. This compounding of small-scale inaccuracies over time limits their skill at extended lead times. These findings support prior observations in weather modeling that direct forecasting can mitigate temporal instability and better preserve long-range predictability.

Impact of Physical Constraints. Introducing physical constraints into machine learning (ML) architectures does not universally improve the accuracy of all forecasted variables. Among the ML-based systems, only the model that explicitly embeds physical constraints within its architectural design consistently outperforms others, most notably in the prediction of geostrophic surface currents $(u_{\text{geo}}, v_{\text{geo}})$. This suggests that architectural inductive biases aligned with physical principles, rather than external forcings or training objectives alone, are key to improving dynamical fidelity. For Lagrangian drift diagnostics, models tend to follow similar performance trends as those observed for surface velocity components (u, v), indicating that advection-dominated features are learned in tandem. However, the improved coherence in derived diagnostics such as $u_{\text{geo}}, v_{\text{geo}}$ further highlights the value of incorporating physically meaningful structure into ML model design.

5.1 Seasonal Dynamics of Forecast Skill

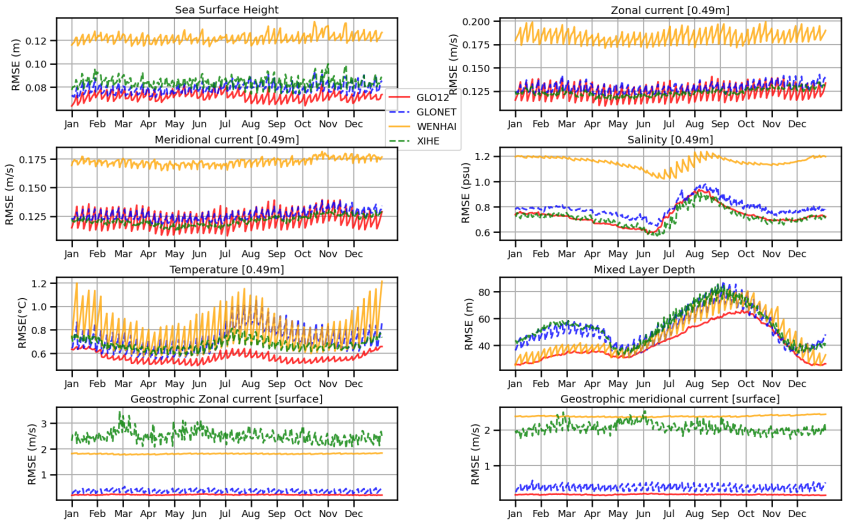


Figure 3: Time series of RMSE evolution throughout 2024 for all models in the Models-to-Reanalysis track.

To complement the headline skill scores, we further examine how model performance evolves throughout 2024. Specifically, we initialize a 7-day forecast every Tuesday and compute the corresponding RMSE for each forecast day. This produces a continuous time series of weekly forecast errors across the full year, providing insight into the temporal (and seasonal) modulation of model skill. We report, in Figure 3, the 7-day forecast RMSE computed daily throughout the year, stratified by variable type, providing a fine-grained view of each model's robustness over time.

Dynamical Variables. For core dynamical variables, zonal and meridional velocities (u, v) and sea surface height (zos), we observe high-frequency fluctuations in RMSE that are consistent across all models and forecast cycles. These fluctuations likely reflect short-term variability and the accumulation of forecast error over the 7-day horizon. However, the absence of any systematic trend or recurring seasonal pattern suggests that model performance for these variables remains temporally stable and does not exhibit sensitivity to seasonal regimes. This holds for both physics-based and data-driven systems, indicating robustness of dynamical state prediction across seasonal transitions.

Tracer Fields. In contrast, strong seasonal patterns emerge in the forecast errors of tracer variables salinity (so) and temperature (thetao). All models, whether machine learning or physics-based, exhibit a clear seasonal modulation in RMSE. Salinity forecast errors tend to reach a minimum around mid-June and peak in September, while temperature forecast errors are lowest in mid-May and peak in August for most models. This modulation likely reflects seasonal changes in surface forcing, stratification, and vertical mixing. However, part of the observed variability, especially in salinity may also contain an interannual component, potentially linked to large-scale climate modes such as El Nino-Southern Oscillation (ENSO), particularly given the strong Nino conditions observed in 2024 [20]. Notably, the seasonal amplitude is less pronounced in the physics-based GLO12 system for temperature, possibly due to its regular data assimilation cycles and physical constraints. In contrast, ML-based systems show stronger seasonal variability, suggesting greater sensitivity to seasonal and possibly interannual variations in upper-ocean structure.

6 Conclusion

We introduce OceanBench, a unified benchmark for evaluating machine learning-based ocean forecasting models with a focus on scientific relevance and reproducibility. By leveraging consistent datasets, rigorous evaluation protocols, and metrics grounded in oceanographic practice, OceanBench aims to standardize model comparison across the community. Our initial release includes a suite of tasks that emphasize not only point-wise accuracy but also process-aware diagnostics, such as Lagrangian drift and mixed layer depth evaluation. OceanBench serves as a bridge between operational oceanography and modern data-driven methods, fostering cross-disciplinary collaboration.

Limitation: While OceanBench provides a robust and standardized framework for evaluating data-driven ocean forecasting models, several limitations remain in its current version. First, the benchmark focuses exclusively on deterministic forecasting tasks, with no support yet for probabilistic models or associated uncertainty-aware evaluation metrics, limiting its applicability for ensemble or stochastic forecasting approaches. Second, although OceanBench is designed for global-scale evaluation, it does not currently provide region-specific breakdowns of performance, which are critical for understanding model behavior in dynamically distinct areas such as western boundary currents or high-latitude regions. Third, the benchmark exclusively targets physical ocean variables such as sea surface height, temperature, and velocity without incorporating biogeochemical fields, which are increasingly important for monitoring marine ecosystems and carbon cycling. These limitations highlight opportunities for future extensions of OceanBench to support uncertainty quantification, regional diagnostics, and interdisciplinary ocean prediction tasks.

Future Work: We plan to extend the evaluation period to assess model performance over longer timescales and diverse oceanic conditions. New evaluation metrics will be incorporated to better capture spatially and dynamically varying forecast skill, giving appropriate emphasis to regions and depths of higher physical relevance. Additional measures will also address uncertainty quantification, extreme events, and physical consistency. Current OceanBench release is designed to accommodate different computational capacities, from GPU-limited research setups to large operational systems, and hence includes a coarser 1° resolution configuration enabling preparation for future updates that will further expand evaluation tracks to facilitate large-scale intercomparison and benchmarking. OceanBench is designed to remain a living benchmark, continually evolving to reflect the latest developments in ocean forecasting and machine learning. To support this vision, we maintain an open-source codebase, an up-to-date website, and an active GitHub repository. We encourage contributions and feedback from the broader community to help shape future versions of the benchmark and ensure its continued impact.

NeurIPS Paper Checklist

For all authors.

1. **Claims**

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

   Answer: [Yes]

   Justification: The paper introduces OceanBench as a standardized benchmark for AI-based ocean forecasting and delivers on this by providing curated datasets, evaluation tracks, baseline models, and open-source tools. These contributions are clearly stated upfront and consistently supported throughout the paper.

   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: The paper includes a dedicated Limitations section that clearly discusses several constraints of the current version of OceanBench. Specifically, the authors acknowledge that the benchmark currently focuses only on deterministic forecasting tasks, without incorporating probabilistic models or uncertainty-aware evaluation metrics. They also note that while OceanBench is designed for global-scale assessments, it does not provide region-specific performance breakdowns, which limits interpretability in areas with distinct ocean dynamics. Furthermore, the benchmark is restricted to physical ocean variables and does not yet include biogeochemical fields, which are important for ecological and climate-related applications. These reflections demonstrate an awareness of the benchmark's scope and potential areas for future improvement.

   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: [No]

   Justification: The paper does not present any formal theoretical results, assumptions, or proofs. It focuses on the design and implementation of a benchmarking framework for data-driven ocean forecasting rather than on developing new theoretical foundations.

   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: The paper provides detailed descriptions of the dataset construction, evaluation protocols, and baseline models. It also includes open-source tools and code, enabling others to reproduce the benchmark setup and replicate the reported results with transparency.

   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: The paper provides open access to all benchmark datasets, baseline model code, and evaluation tools through public repositories, along with clear instructions for setup and replication. This ensures that the main experimental results are fully reproducible.

   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: [NA]

   Justification: The OceanBench paper introduces a benchmark and evaluation protocol but does not propose a new model or conduct original training experiments. Instead, it aggregates existing datasets and provides standardized evaluation procedures for model comparison. As such, detailed training configurations are out of scope.

   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: [Yes]

Justification: The benchmark includes standardized evaluation metrics such as RMSE and L2 distance, and baseline results are reported with appropriate error quantification over spatial and temporal domains. The methods for computing and interpreting these metrics are clearly documented in the supplemental material and code repository.

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: [No]

   Justification: While the OceanBench paper introduces baseline results and benchmark datasets, it does not systematically report the compute resources used to run these baselines. There is no mention of GPU/CPU types, memory, execution time, or total compute estimates.

   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 paper adheres to the NeurIPS Code of Ethics. It releases open-source benchmarking tools and builds on publicly available datasets with appropriate licenses and acknowledgments. There is no indication of ethical violations or concerns in data use, model application, or reproducibility.

   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: The paper discusses the positive impacts of improving ocean forecasting, which has implications for climate science, environmental monitoring, and societal safety (e.g., marine hazards). It also acknowledges the need for transparency and community-driven development, though potential negative impacts are more lightly addressed.

    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: The benchmark datasets and code released in the paper do not pose a high risk of misuse. They are based on physical and environmental oceanographic data and do not include sensitive personal information or general-purpose generative models.

    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: The paper provides appropriate credit for all reused datasets and acknowledges their licensing terms. The code and benchmarks released also include license information and citations where applicable.

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: [Yes]

Justification: The OceanBench benchmark introduces new curated datasets and baseline implementations, all of which are well-documented in the GitHub repository. The documentation includes usage examples, data structure, metric definitions, and links to tutorials and metadata descriptions.

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: The paper does not involve crowdsourcing or human subject research.

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: 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 rigorousness, 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

[DSFNMOC] DSFNMOC. Godae sfcobs surface temperature observations (1998–present).

[2] Aouni, A. E., Gaudel, Q., Regnier, C., Van Gennip, S., Drevillon, M., Drillet, Y., and Lellouche, 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] Epicoco, I., Donno, D., Accarino, G., Norberti, S., Grandi, A., Giurato, M., McAdam, R., Elia, D., Clementi, E., Nassisi, P., Scoccimarro, E., Coppini, G., Gualdi, S., Aloisio, G., Masina, S., Boccaletti, G., and Navarra, A. (2025). Medformer: a data-driven model for forecasting the mediterranean sea.

[15] Flemming, N. C. (2002). Strategic planning for operational oceanography. In *Ocean Forecasting: Conceptual Basis and Applications*, pages 1–17. Springer.

[16] Garcia, P., Larroche, I., Pesnec, A., Bull, H., Archambault, T., Moschos, E., Stegner, A., Charantonis, A., and Béréziat, D. (2025). Orcast: Operational high-resolution current forecasts. *Artificial Intelligence for the Earth Systems*.

[17] Hernandez, F., Bertino, L., Brassington, G., Chassignet, E., Cummings, J., Davidson, F., Drévillon, M., Garric, G., Kamachi, M., Lellouche, J.-M., et al. (2009). Validation and intercomparison studies within godae. *Oceanography*, 22(3):128–143.

[18] 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.

[19] Holmberg, D., Clementi, E., Epicoco, I., and Roos, T. (2025). Accurate mediterranean sea forecasting via graph-based deep learning. *arXiv preprint arXiv:2506.23900*.

[20] 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.

[21] 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.

[22] 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*.

[23] 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.

[24] Madec, G. et al. (2015). Nemo ocean engine.

[25] 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*.

[26] 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.

[27] Rasp, S., Hoyer, S., Merose, 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.

[28] 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.

[29] 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.

[30] Tonani, M., Balmaseda, M., Bertino, L., Blockley, E., Brassington, G., Davidson, F., Drillet, Y., Hogan, P., Kuragano, T., Lee, T., Mehra, A., Paranathara, F., Tanajura, C. A., and Wang, H. (2015). Status and future of global and regional ocean prediction systems. *Journal of Operational Oceanography*, 8(sup2):s201–s220.

[31] 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.

[32] 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: Preliminaries

This appendix provides key background information on ocean forecasting concepts, datasets, and terminology used throughout the paper. Operational oceanography is the provision of scientifically based information and forecasts about the state of the sea (including its chemical and biogeochemical components) on a routine basis and with sufficient speed, so that users can act on the information and make decisions before the relevant conditions have changed significantly or become unpredictable [15].

A.1 Operation oceanographic systems

Operational oceanographic systems are historically based on numerical modelling of the ocean dynamic and data-assimilation schemes for the blending of the observations into the model to provide the most accurate description of the past and the future [30].

Figure 4 shows an overview of the lifecycle of an ocean state estimate at a given time t in the context of operational oceanography. In addition to observation assimilation, operational oceanographic systems are usually forced at the air-sea interface by atmospheric fields produced by operational atmospheric centers. Updated atmospheric fields and newly acquired observations are systematically released and thus used in the rerunning of the systems to produce more accurate ocean state estimates. As long as time t is in the future, the state estimate is called *forecast*. When time t is in the recent past and observations are assimilated in dribs and drabs, the state estimate is called *analysis*. After some point and depending on their operational constraints and data dissemination strategy, atmospheric and observation centers usually release best-quality datasets, allowing more updates of the state estimate, now called *reanalysis*.

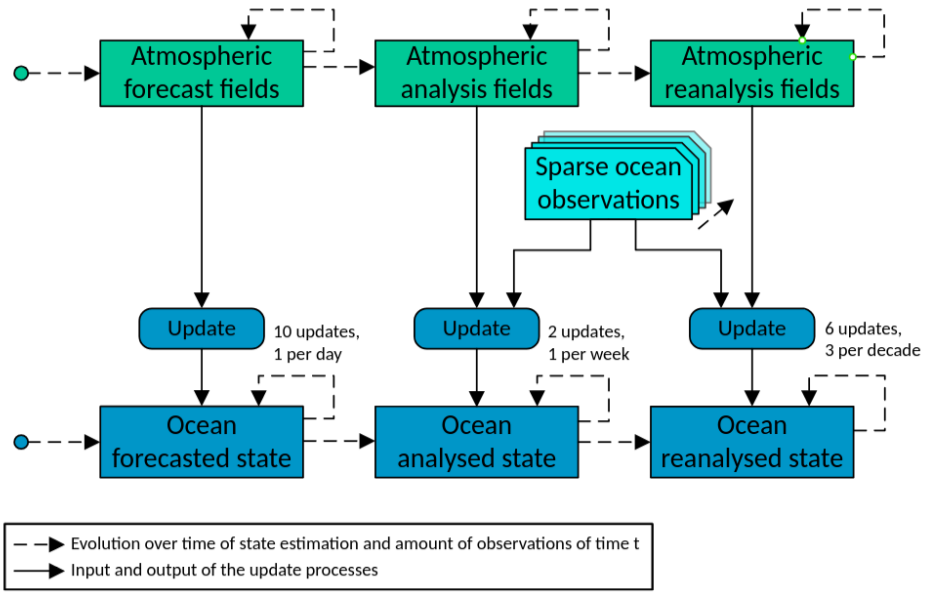


Figure 4: Lifecycle of an ocean state estimate at time t in operational oceanography: Operational oceanographic systems usually estimate ocean states by forcing atmospheric fields at air-sea interface and by assimilating observations acquired over time, enabling the production of more accurate state estimates. For future t (e.g., 10 days), the state estimate is a *forecast* and is updated regularly (e.g., daily) with new atmospheric forecasts. For recent past t (e.g., 2 weeks), the state estimate is an *analysis* and is updated regularly (e.g., weekly) with new atmospheric analysis and observations. Updates with high-quality atmospheric and observational datasets produce *reanalysis* (e.g., 6 updates over 2 decades). A system may include one or more of these components and the update frequencies may vary.

A.2 Headline Metrics and Application Relevance

The ocean headline metrics exposed in Table 1 capture essential oceanic processes and are directly linked to a wide range of marine applications. The horizontal velocity components (u_o, v_o) and sea

surface height (z_{os}) together describe the total surface circulation, encompassing both geostrophic and ageostrophic components that govern surface transport. These fields are critical for navigation, shipping route optimization, and the drift prediction of floating objects. Sea surface salinity (s_{os}) and temperature (θ_o) represent thermohaline variability and air-sea interactions, underpinning applications in fisheries management, marine ecosystem monitoring, and climate diagnostics. The geostrophic current, derived from gradients in z_{os}, isolates the balanced circulation and mesoscale eddy activity, providing insight into nutrient transport and biologically productive regions relevant to fishing and aquaculture. The mixed layer depth (MLD) further characterizes vertical mixing and stratification, controlling heat and nutrient exchanges. Finally, Lagrangian trajectory assessments provide an integrated measure of transport skill, relevant for pollutant dispersion, search-and-rescue operations.

B 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 [17, 29, 9], provides an operationally relevant benchmark by leveraging high-quality drifters measurements.

B.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.

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

XiHe [32] 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.

B.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

Model	Type	Autoregressive	Initialization	Horizontal Res.	Vertical Levels / Depth
GLO12	Physical (Forecast & Analysis)	Yes	Self-initialized (own forecast) + IFS	$1/12^\circ$	50 levels (0m-seafloor)
GLONET	AI-based	Yes	From GLO12	$1/4^\circ$	21 levels (0m-seafloor)
XiHe	AI-based	No (Direct)	From GLO12 + IFS (u10, v10)	$1/12^\circ$	23 levels (0-600 m)
WenHai	AI-based	Yes	From GLO12 + IFS (t2m, d2m, u10, v10, mtpr, ssr, strd, msl)	$1/12^\circ$	23 levels (0-600 m)

Table 3: Summary of the forecasting models used in OceanBench, including their type, initialization, and resolution. Autoregressive models produce forecasts iteratively from previous outputs, while direct models predict specific lead times independently.

Table 3 summarizes the key characteristics of the four forecasting systems considered in this study. Their computational requirements differ notably: the physics-based GLO12 model is the most demanding, as it integrates the full ocean dynamics at high spatial and temporal resolution using large-scale HPC resources. In contrast, the ML-based models require substantial resources during training but are considerably more efficient during inference, producing multi-day forecasts in a fraction of the time. This distinction highlights the potential advantages of data-driven approaches for scalable and real-time global ocean prediction.

B.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 [17, 29, 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.

C Appendix: 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.

C.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 profilesm 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:

$$\text{MLD} = \min\{z \mid \rho_z - \rho_0 \geq \Delta\rho\} \tag{2}$$

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.

C.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:

$$\mathbf{v}(\phi, \theta, t) = gf^{-1}\nabla^{\perp}\eta(\phi, \lambda, t) \tag{3}$$

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.

C.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:

$$\mathbf{v}(\mathbf{x}, t), \quad \mathbf{x} \in \mathbb{R}^2, \quad t \in [t_0, t_f] \tag{4}$$

and its associated ordinary differential equation:

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

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:

$$\mathbf{x}(t_f, t_0, \mathbf{x}_0) = \mathbf{x}_0 + \int_{t_0}^{t_f} \mathbf{v}(\mathbf{x}(\tau), \tau)\, d\tau \tag{6}$$

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:

$$\text{Lagrangian drift deviation}(t) = \frac{1}{N}\sum_{n}^{N}\left|\mathbf{x}_i^f(t) - \mathbf{x}_i^r(t)\right| \tag{7}$$

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.

D Appendix: 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.

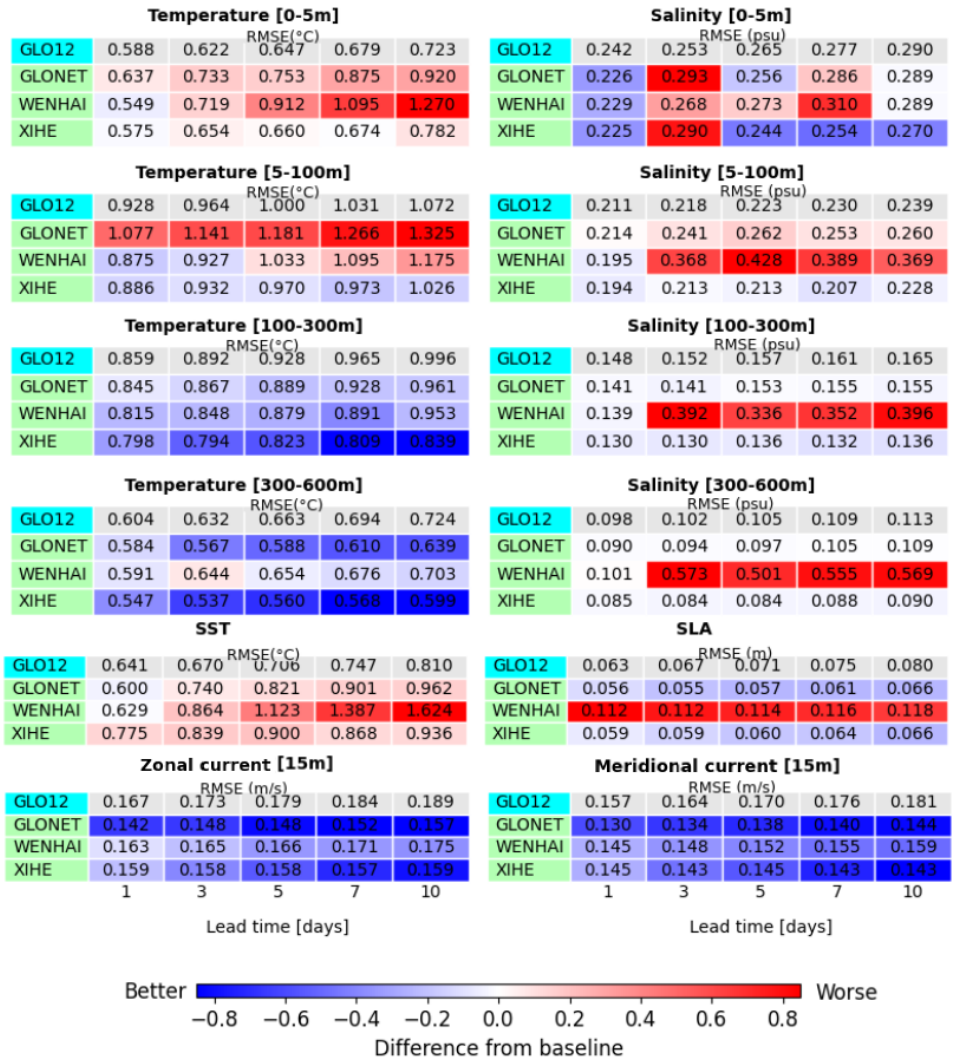


Figure 5: Models to observations track.

D.1 Models-to-Observations Track

The models-to-observations track provides a direct evaluation of forecast skill against independent in situ and satellite observations (see Figure 5). 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 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.

D.2 Models-to-Analysis Track

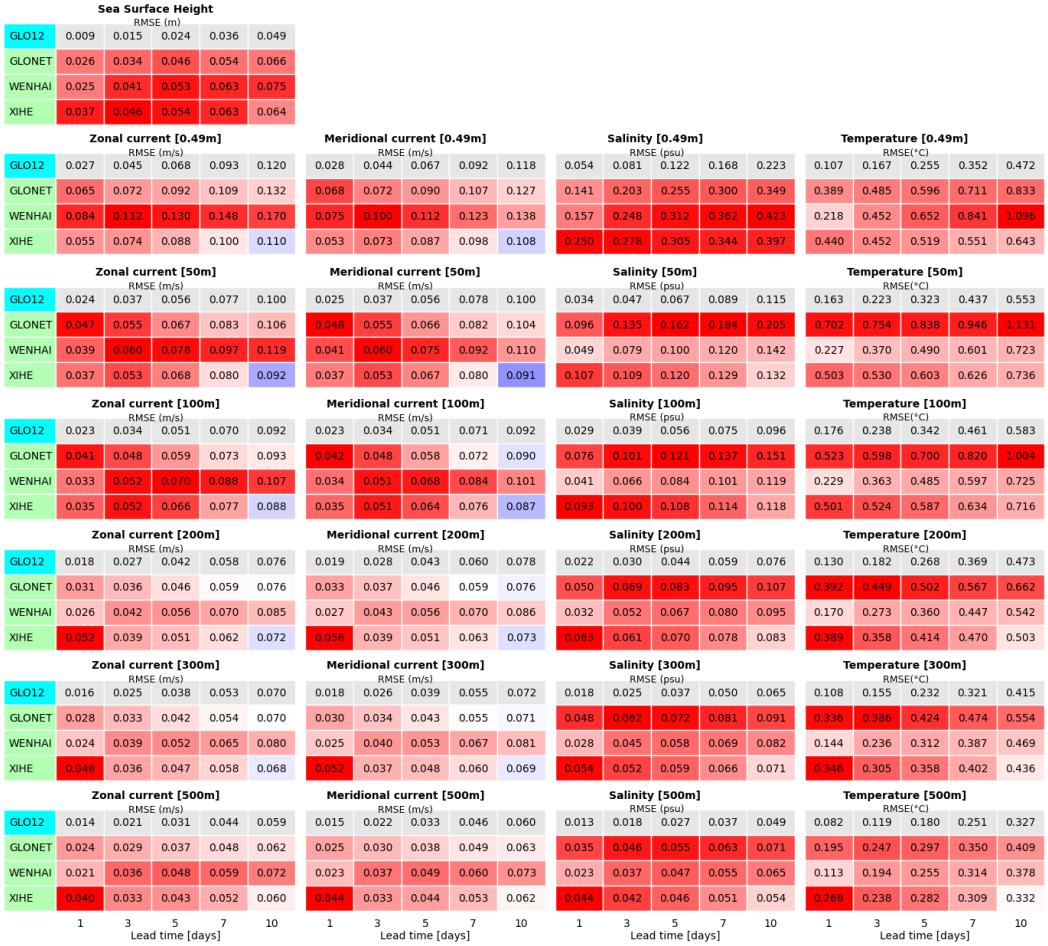
The models-to-analysis track evaluates model forecasts against the GLO12 analysis (see Figure 6). Unsurprisingly, all models underperform relative to the GLO12 baseline in this track, as the reference analysis is itself generated from the GLO12 forecast model. Specifically, the GLO12 analysis is produced 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.

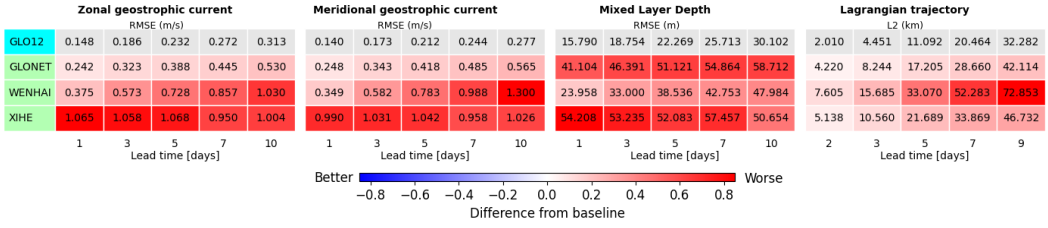
In summary, the models-to-analysis track is best interpreted as a measure of structural consistency with the GLO12 system, rather than as an unbiased indicator of real-world forecast skill. It complements the observation-based track by revealing how different model classes align or diverge from an established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

D.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 7. 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.



Forecasted variables.



Diagnostic variables.

Figure 6: Models to analysis track.

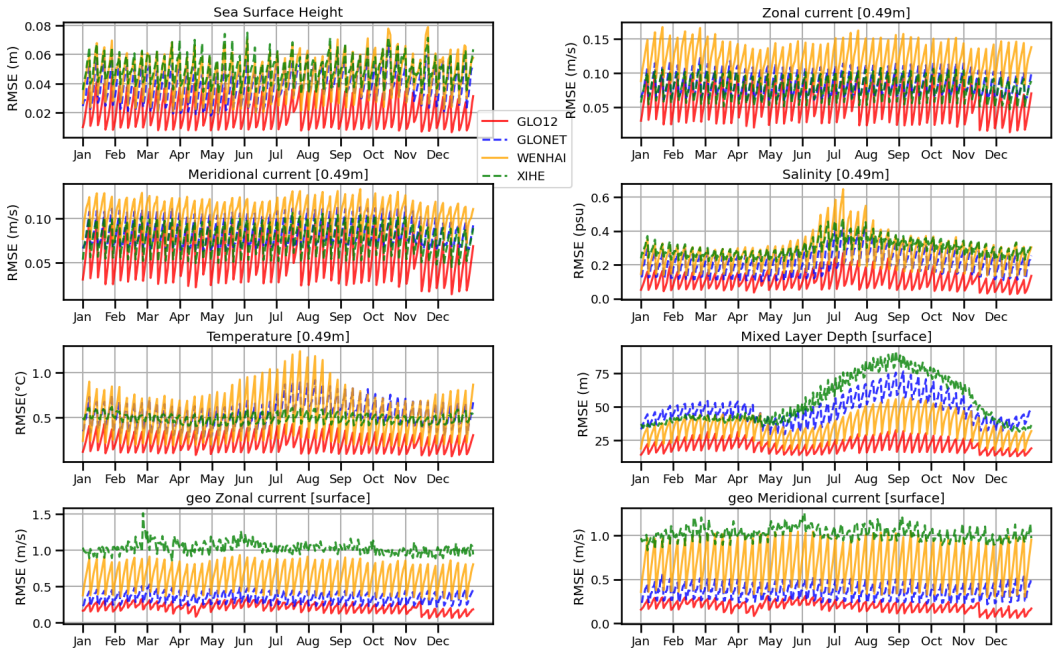


Figure 7: Time series of RMSE evolution throughout 2024 for all models in the Models-to-Analysis track.

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.

E Spatial Structure and Scale-Resolved Evaluation

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.

E.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 8). 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

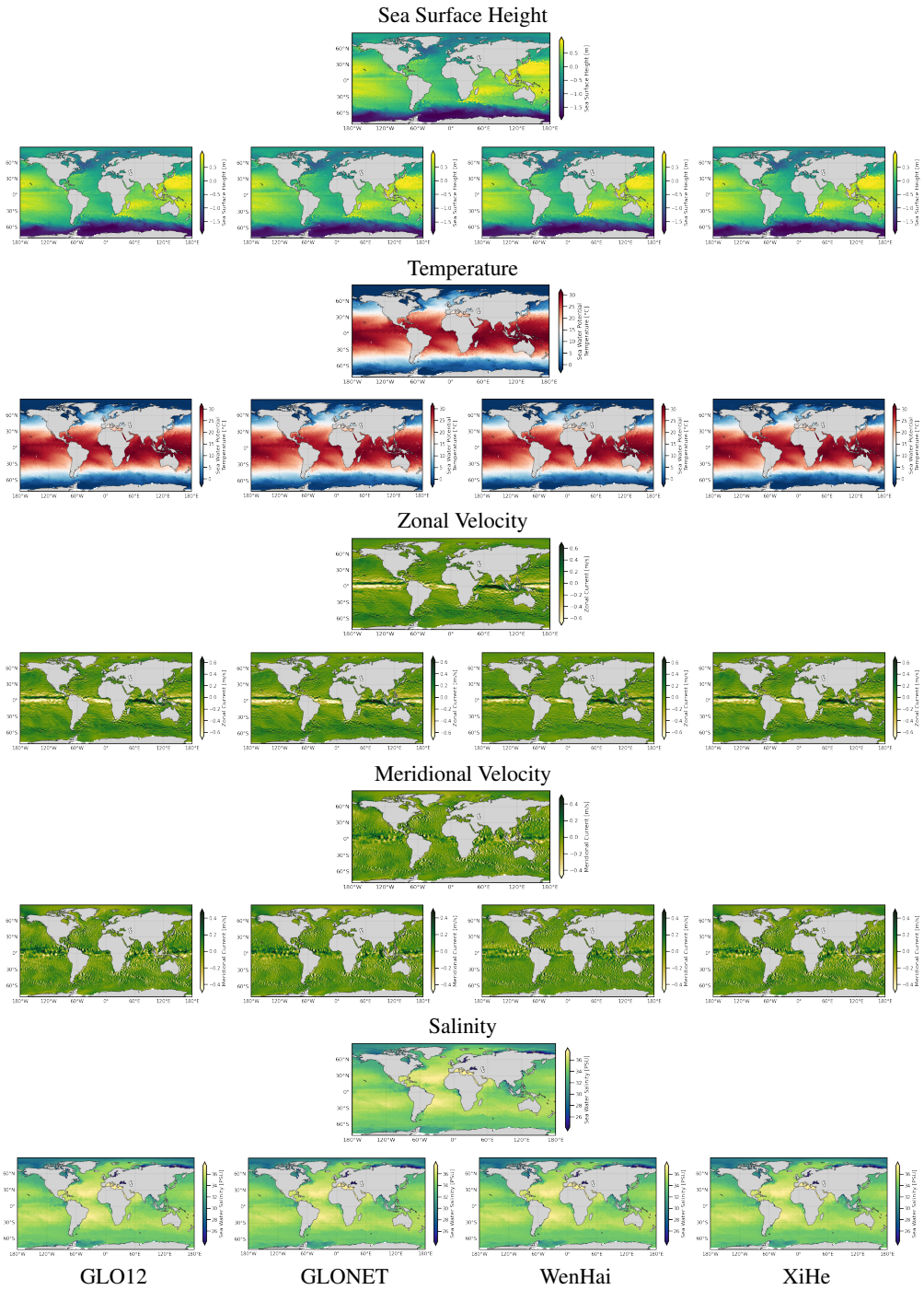


Figure 8: A set of results for world map for the forecasted variables. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity All of these are for lead time 1 for the date 2024-01-03.

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.

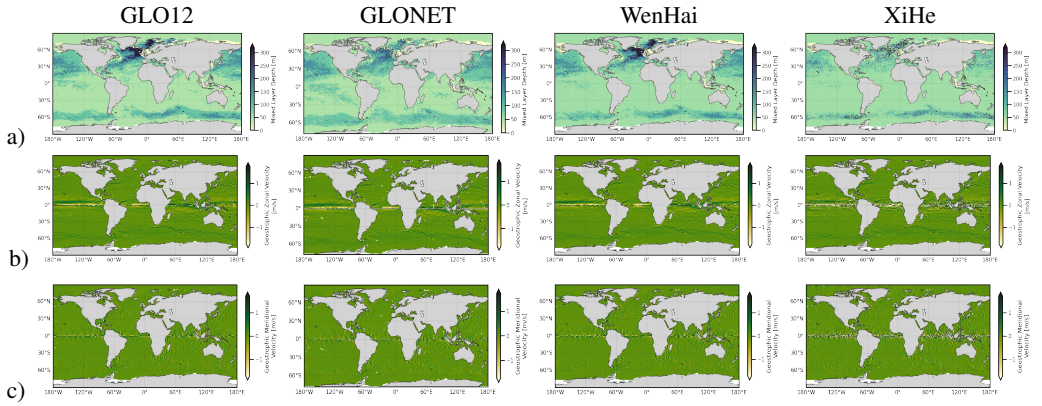


Figure 9: A set of results for world map for diagnostic variables. The following physical variables are as follows: a) Geostrophic Zonal Velocity, b) Geostrophic Meridional Velocity, and c) Mixed Layer Depth All of these are for lead time 1 on the date 2025-01-02.

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 9). 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 10-14). 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.

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.

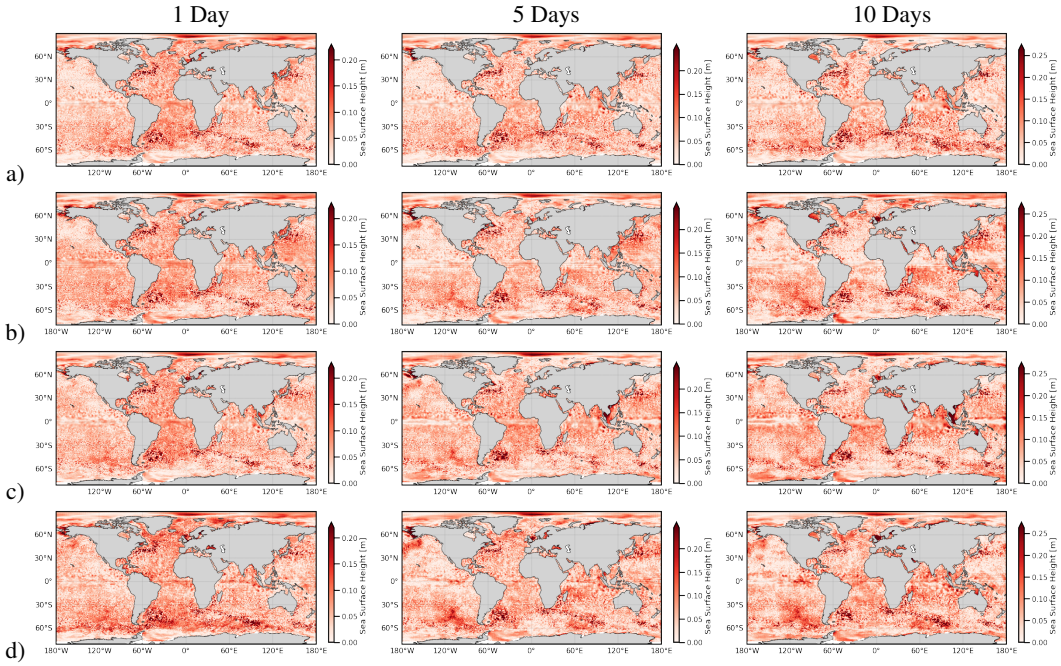


Figure 10: Error Maps of Root Mean Squared Error as a function of lead time for Sea Surface Height. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03,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

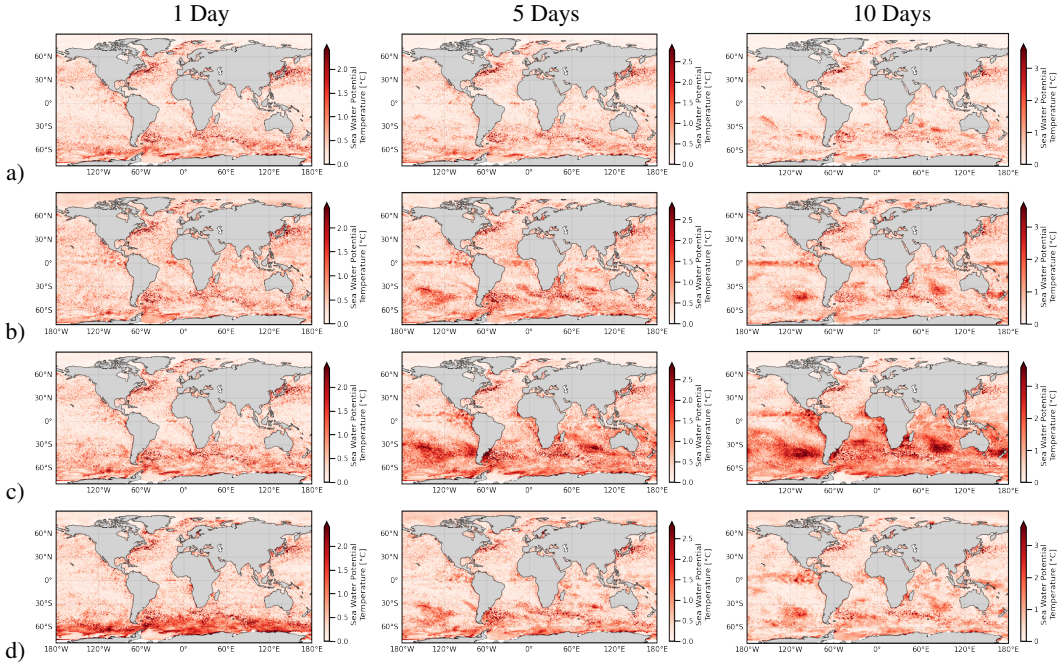


Figure 11: Error Maps of Root Mean Squared Error as a function of lead time for Temperature. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

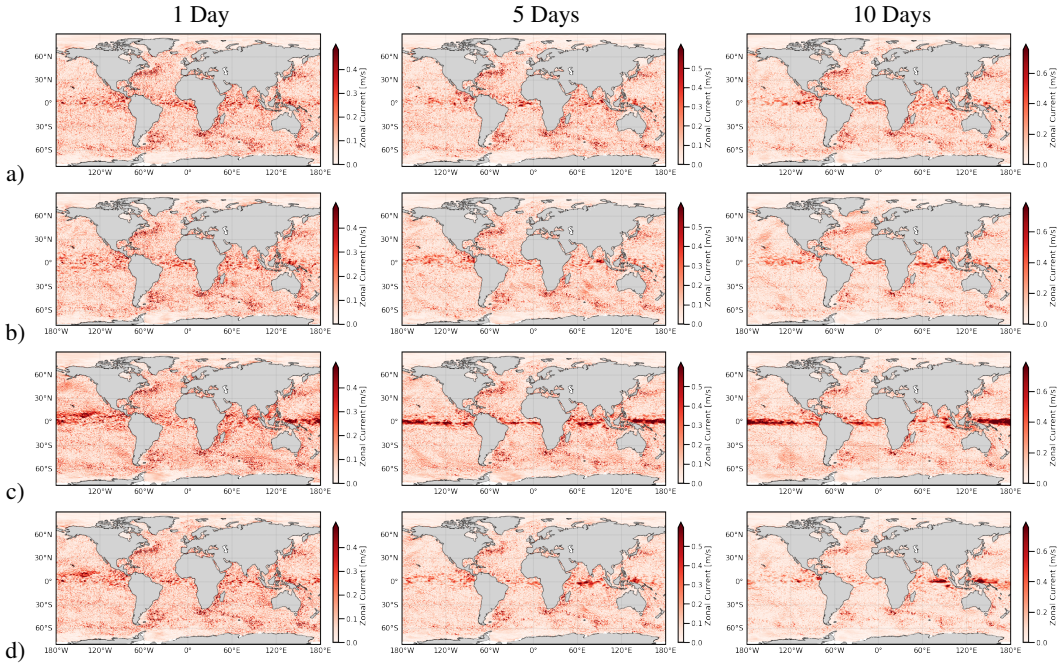


Figure 12: Error Maps of Root Mean Squared Error as a function of lead time for the Zonal Velocity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03,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

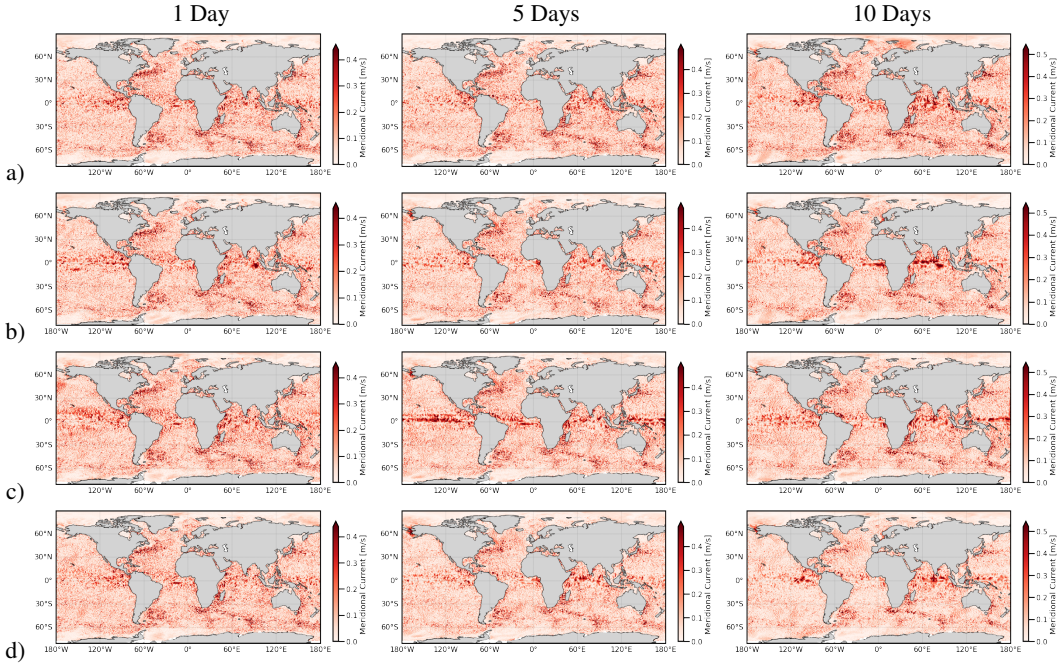


Figure 13: Error Maps of Root Mean Squared Error as a function of lead time for the Meridional Velocity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

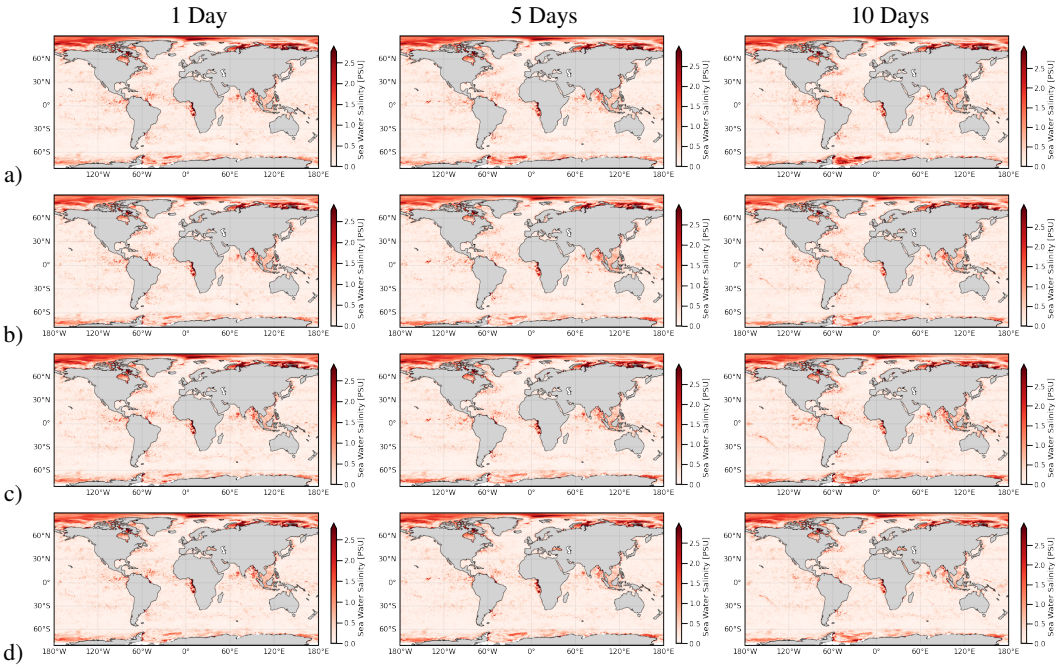


Figure 14: Error Maps of Root Mean Squared Error as a function of lead time for Sea Surface Salinity. Each model is compared with the GLORYS12 Reanalysis on 2024-01-03 with a lead time of 1, 5, and 10 respectively. We showcase the following variables: a) GLO12, b) GLONET, c) WenHai, d) XiHe. 2024-01-03, 2024-01-05,

E.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:

$$\hat{f}(k_x, k_y) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \tilde{f}(x, y) \exp\left(-2\pi i \left(\frac{k_x x}{N_x} + \frac{k_y y}{N_y}\right)\right), \tag{8}$$

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:

$$\text{PSD}(k_x, k_y) = |\hat{f}(k_x, k_y)|^2. \tag{9}$$

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.

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 15). 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.

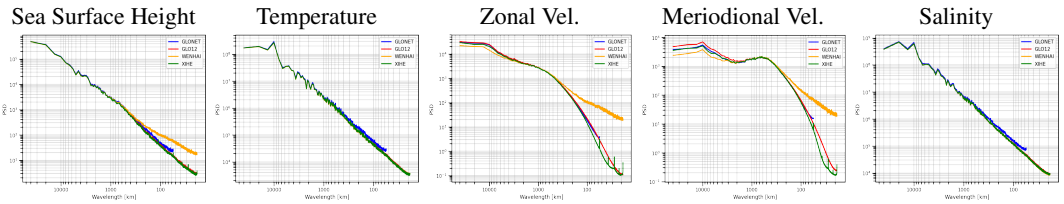


Figure 15: A set of results for the power spectrum for the zonal direction averaged over the latitude and time(2024) for the whole globe. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity. This figure only shows a lead time of 1.

Regional-scale analysis. The computed PSDs over the gulf stream region (Figure 16) 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.

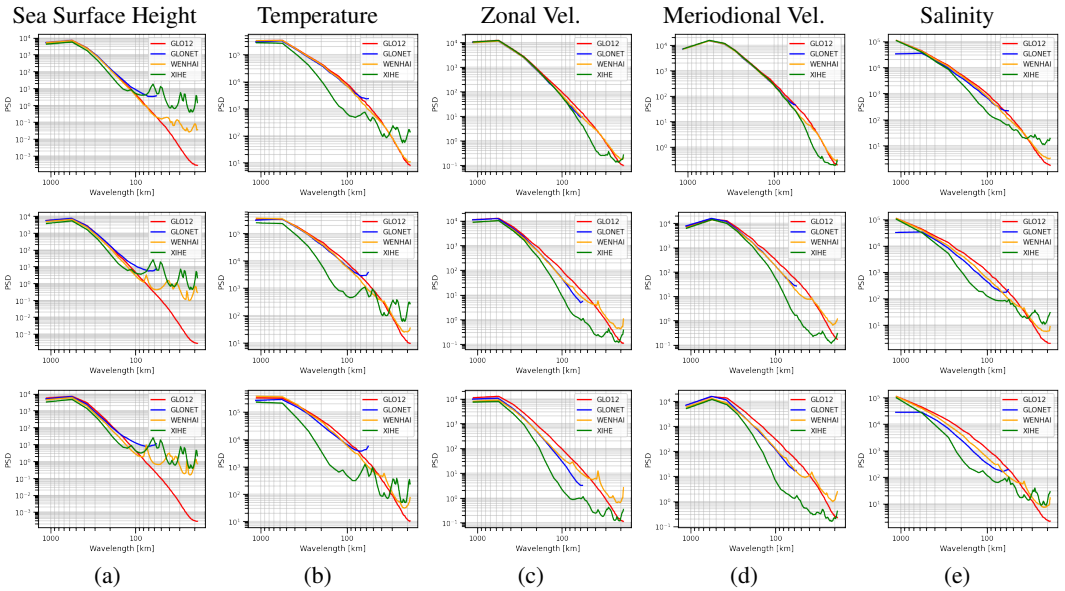


Figure 16: A set of results for the power spectrum for the zonal direction averaged over the latitude and time(2024) over the Gulf Stream. The following physical variables are as follows: a) Sea Surface Height, b) Seawater Potential Temperature, c) Zonal Current, d) Meridional Current, e) Salinity. Rows 1, 2, and 3 are the lead times of 1 day, 5 days, and 10 days respectively.

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.

Overall, the PSD analysis provides a rigorous and interpretable framework for evaluating the scale-resolving skill of forecasting systems. Although it does not directly disentangle specific physical processes, it effectively characterizes how each model distributes and evolves energy across spatial and temporal scales, in line with their common forecasting objective. This perspective 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.