PHYSICS-ALIGNED NEURAL MODELING OF OCEAN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-driven models have achieved strong performance in geophysical forecasting, particularly in the atmosphere, but transferring these approaches to the ocean remains challenging due to fundamental differences in geometry and dynamics. The ocean is defined on a discontinuous domain with complex coastlines, strong stratification, and highly anisotropic, surface-driven transport, which are poorly aligned with standard learning architectures. As a result, existing models often exhibit physically inconsistent behavior, including land-sea mixing, unrealistic vertical transport, and incorrect propagation of surface forcing. We propose a physics-aligned neural architecture for ocean forecasting that embeds these structural constraints directly into the model design. The framework operates exclusively on the wet domain, respects boundary-driven forcing, decouples horizontal and vertical transport mechanisms, and incorporates geometry-aware interactions adapted to the structure of the ocean system. The resulting model produces stable long-horizon autoregressive simulations with coherent spatiotemporal structure across physical, sea ice, and biogeochemical variables, while remaining robust under perturbed forcing conditions. Additional analyses further demonstrate stable spectral behavior, physically consistent cross-variable coupling, and improved long-term stability relative to architectures that rely on generic feature-space interactions. These results highlight the importance of incorporating domain-specific geometry and inductive biases directly at the architectural level, rather than relying solely on data scale or loss-based physical constraints.

1 INTRODUCTION

Recent progress in machine learning has significantly reshaped geophysical forecasting, with particularly strong advances in atmospheric modeling. Data-driven approaches now achieve competitive, and in some regimes superior, performance compared to traditional numerical weather prediction systems, while offering substantial gains in computational efficiency (Rasp et al., 2024). State-of-the-art data-driven models are able to learn the spatiotemporal evolution of large-scale atmospheric dynamics directly from data with remarkable fidelity (Lang et al., 2024; Price et al., 2023; Bonev et al., 2025; Lang et al., 2026; Kurth et al., 2023; Lam et al., 2023). These successes are partly enabled by the relatively regular structure of the atmosphere, which can be represented on quasi-continuous spherical domains with limited geometric heterogeneity and relatively simple internal boundary conditions.

These advances have naturally motivated extensions to ocean forecasting. A growing body of work explores learning ocean dynamics directly from reanalysis and observational data for short-term prediction, spanning surface-focused models (Botvynko et al., 2025; Garcia et al., 2025), fully three-dimensional formulations (El Aouni et al., 2025b; Hirabayashi et al., 2026; Huang et al., 2025; Wang et al., 2024; Yang et al., 2024; Cui et al., 2025; Agarwal et al., 2026), and regional-scale systems (Epicoco et al., 2025; Holmberg et al., 2024). These approaches demonstrate skill in forecasting key ocean state variables, including temperature, salinity, currents, and sea surface height, over short lead times (El Aouni et al., 2025a). By exploiting the underlying spatiotemporal structure of these datasets, they learn efficient approximations of ocean dynamics, enabling fast, high-resolution predictions and suggesting a complementary role to traditional forecasting systems.

However, transferring machine learning atmospheric models to the ocean domain exposes fundamental structural mismatches. Although both systems are governed by geophysical fluid equations,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

their geometric and dynamical properties differ markedly (Talley, 2011). The ocean occupies a highly irregular domain shaped by complex coastlines and heterogeneous bathymetry (Wölfl et al., 2019), yielding a spatially discontinuous computational manifold that challenges standard learning approaches. In addition, unlike the atmosphere's relatively regular vertical structure, the ocean exhibits depth-varying volume and strong stratification. Its dynamics are also highly anisotropic: horizontal transport dominates vertical exchange by orders of magnitude, while variability emerges from the interplay between surface forcing, pressure-driven circulation, stratification, and vertically constrained mixing and propagation processes (Piggott et al., 2009).

These properties introduce failure modes when using standard ML architectures. By implicitly assuming locally homogeneous neighborhoods, they induce systematic land-sea mixing near coastlines, resulting in physically inconsistent representations. Likewise, flattening the vertical structure into generic feature dimensions neglects the distinct scales and transport regimes governing vertical dynamics, often resulting in spurious diffusion, weakened stratification, and inconsistent cross-depth interactions. More critically, unconstrained feature mixing can allow surface forcing signals to propagate directly across depth layers, creating shortcut pathways that bypass physically consistent transport and mixing processes.

In response to these failure modes, a common approach is to incorporate physical constraints via regularization or physics-informed losses (Cai et al., 2021; Cuomo et al., 2022). While partially effective, this remains limited in expressivity, as only a subset of governing constraints can be enforced in tractable form, leaving models free to develop unphysical internal representations as long as output-level constraints are satisfied. Consequently, such methods often require extensive data, careful tuning, and exhibit limited out-of-distribution generalization. This points to a more fundamental perspective: the challenge in ocean machine learning is not only to learn dynamics, but to encode the geometry and inductive biases of the system itself, making it primarily an architectural problem of representation rather than purely a statistical one.

In this work, we propose a physics-aligned neural architecture for ocean forecasting that embeds structural constraints directly into the model design rather than relying on loss-based regularization. Our framework operates exclusively on wet ocean points, eliminating land contamination at the representation level; treats depth as a structured axis with regime-dependent dynamics rather than a flat feature dimension; enforces physically consistent surface-only forcing; and decouples horizontal and vertical operators to reflect the strong anisotropy of ocean transport. By constraining the hypothesis space to align with known physical structure, the proposed architecture improves data efficiency, enhances generalization, and yields more physically faithful spatiotemporal representations of ocean dynamics.

2 Related Work

2.1 Numerical Ocean Models

Operational ocean forecasting remains largely based on numerical ocean general circulation models (OGCMs), such as NEMO ocean model (Madec et al., 2015) and HYCOM (Chassignet et al., 2007), which solve discretized forms of the governing equations. These systems provide physically grounded and consistent representations of ocean dynamics, forming the backbone of forecasting and reanalysis.

However, their formulation imposes intrinsic computational constraints. The integration of nonlinear, stiff PDEs requires tightly coupled solvers and restrictive time stepping, resulting in high computational cost and limited throughput. This restricts the ability to generate large ensembles, iterate rapidly, or adapt models efficiently. In parallel, unresolved processes must be parameterized, introducing structural biases that are difficult to calibrate and often degrade robustness.

More fundamentally, these challenges reflect a lack of scalability in the modeling paradigm: extending physically consistent models while maintaining computational efficiency and flexibility remains difficult, particularly when capturing multiscale interactions and boundary-driven dynamics within a unified framework.

2.2 Data-Driven Ocean Forecasting

Data-driven approaches to ocean forecasting aim to learn surrogate dynamical operators directly from reanalysis and observational data, enabling fast inference once trained. A range of architectures have been explored, including convolutional networks, transformers, neural operators, and graph-based models, with demonstrated skill in predicting key ocean variables such as temperature, salinity, currents, and sea surface height (Epicoco et al., 2025; Wang et al., 2024; El Aouni et al., 2025b; Hirabayashi et al., 2026). Within these frameworks, design strategies such as land masking, temporal embeddings, and autoregressive formulations are commonly used to improve stability and rollout behavior.

Despite these advances, a common pattern emerges: most existing models implicitly inherit architectural inductive biases from atmospheric or vision-based learning systems, where spatial domains are relatively regular and isotropic. When transferred to the ocean, these assumptions become misaligned with the underlying structure of the problem. The ocean is defined on a discontinuous domain with complex coastlines, strong vertical stratification, and highly anisotropic transport dominated by horizontal flows and surface forcing. As a result, standard architectures struggle to consistently represent boundary interactions, vertical coupling, and the asymmetric role of surface forcing in driving interior dynamics.

In practice, these limitations are often obscured by the emphasis on short forecasting horizons, where accumulated errors remain small and the slow evolution of ocean dynamics can mask structural deficiencies. However, such inconsistencies become increasingly apparent over extended autoregressive integrations, particularly through boundary effects, cross-depth coupling errors, and land-sea interactions. Consequently, short-term predictive skill alone may not reliably reflect the physical consistency or stability of the learned dynamics.

3 Method

3.1 Problem Formulation

We consider the problem of learning a stochastic forecasting model for the coupled three-dimensional ocean state. Let x_t denote the full ocean state at time t, comprising circulation, thermodynamics, surface pressure-related variables, sea ice, and biogeochemical tracers, defined over a spatial domain with complex land-ocean geometry and depth-dependent support. Given past states and external forcing f_t, the objective is to learn a model of the form $p_\theta(x_{t+1} \mid x_t, f_t)$, which can be applied autoregressively to generate multi-step forecasts.

A central challenge is that the ocean state does not reside on a regular Euclidean grid with uniform support. Instead, it evolves on a bathymetry-constrained, anisotropic, and boundary-forced domain, where geometry, vertical structure, and physical coupling play a defining role. This motivates a model in which these constraints are encoded directly in the computational structure rather than imposed through the loss or implicitly through auxiliary inputs. While prior work incorporates physical information via regularization terms or additional feature channels, such approaches still rely on the model to learn how to use this information appropriately. In contrast, we restrict the model's degrees of freedom, eliminating unphysical interactions by design. Figure 1 gives an overview of the resulting architecture.

3.2 Anisotropic and Boundary-Constrained Dynamics

Operator factorization Ocean dynamics are governed by strongly anisotropic transport, in which horizontal advection dominates vertical exchange due to stratification and the small aspect ratio of the ocean (Wunsch & Ferrari, 2004). This leads to a natural decomposition of the governing dynamics into lateral transport and vertical flux divergence. We encode this structure by decomposing the learned evolution operator of parameters θ into horizontal and vertical contributions acting on each prognostic variable q:

$$\partial_t q = \mathcal{H}_\theta(q) + \mathcal{V}_\theta(q), \tag{1}$$

where $\mathcal{H}_\theta$ represents horizontal transport and submesoscale variability, and $\mathcal{V}_\theta$ represents vertical redistribution constrained by stratification. In contrast to standard architectures that implicitly assume

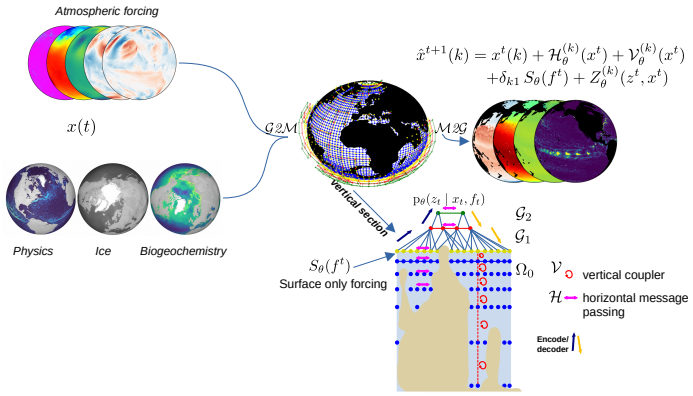


Figure 1: Overview of the proposed physics-aligned forecasting framework. The coupled ocean state $x(t)$, comprising physical, sea ice, and biogeochemical variables, is combined with atmospheric forcing and mapped onto a wet-point graph representation that respects the bathymetry-constrained domain. The evolution operator is factorized into horizontal transport ($\mathcal{H}$) and vertical flux divergence ($\mathcal{V}$), enforcing anisotropic dynamics. Vertical coupling is implemented through interface-based fluxes, while external forcing $S_\theta(f_t)$ is applied exclusively at the surface boundary. A stochastic latent variable z_t models unresolved processes and modulates the dynamics. The full system evolves through a composition of graph-based operators acting across horizontal layers and vertical connections, preserving domain geometry, boundary conditions, and multi-physics coupling.

isotropic interactions, this factorization enforces a separation of transport pathways consistent with the underlying physics, preventing spurious coupling between horizontal and vertical processes.

Interface-based vertical coupling Vertical transport in the ocean is governed by fluxes across layer interfaces, rather than direct mixing between discrete levels. We therefore formulate vertical exchange in conservative flux form. For a state variable q_k at level k (e.g., tracer or momentum component), the discrete update is given by:

$$q_k^{t+1} = q_k^t + \Delta t \left(H_k(q)^t - \frac{F^q_{k+1/2} - F^q_{k-1/2}}{\Delta \zeta_k} \right), \tag{2}$$

where H_k denotes horizontal tendencies and $F^q_{k+1/2}$ is the vertical flux across the interface between depth (ζ) levels k and $k+1$. The fluxes $F^q_{k+1/2}$ are parameterized as functions of the local state and geometric context, including stratification, layer thickness, bathymetry, and proximity to boundaries. This formulation ensures that vertical transport is expressed as a flux divergence, preserving the conservative structure of the underlying equations and avoiding unphysical cross-level shortcuts. The flux form is imposed on the learned operator rather than as a hard constraint on the decoded fields, so closure is approximate; Sec. 5.2 measures it.

Surface-only forcing Air -sea fluxes provide the dominant source of momentum, heat, and tracers in the ocean and act exclusively at the surface boundary. We enforce this boundary condition by restricting external forcing to the top layer:

$$q_k^{t+1} = q_k^t + \Delta t \left[\mathcal{H}_\theta^{(k)}(q^t) + \mathcal{V}_\theta^{(k)}(q^t) + \delta_{k1} S_\theta(f_t) \right], \tag{3}$$

where δ_{k1} is a Kronecker delta that restricts forcing to the surface layer. Interior layers evolve solely through horizontal transport and vertical flux divergence, without direct access to forcing inputs $S_\theta(f_t)$. This enforces a causal structure in which surface perturbations propagate into the interior through physically consistent pathways, rather than through direct injection at depth.

3.3 Coupled Multi-Physics State Representation

The ocean state comprises multiple interacting subsystems, including momentum, thermodynamic variables, sea surface height, sea ice, and biogeochemical tracers, each governed by distinct evolution

equations and defined on different spatial supports. We represent the full system as a coupled state:

$$x_t = \{q_t^{(m)}\}_{m=1}^{M}, \tag{4}$$

where each one of the M components evolve according to:

$$q_k^{(m),t+1} = q_k^{(m),t} + \Delta t \left[\mathcal{H}_\theta^{(m,k)}(x^t) + \mathcal{V}_\theta^{(m,k)}(x^t) + \delta_{k1} S_\theta^{(m)}(f_t)\right], \tag{5}$$

where the dependence on the full state x reflects cross-variable coupling.

Crucially, the spatial support of these variables differs. Three-dimensional tracers and velocity fields are defined over the volumetric ocean domain Ω_o, while surface quantities such as sea surface height and sea ice are defined on the two-dimensional manifold $\partial\Omega_o$. As a result, the coupled system is defined over a union of manifolds with different dimensionality and boundary conditions. To preserve this structure, predictions are parameterized through variable-specific mappings:

$$\hat{q}^{(m)} = \mathcal{D}^{(m)}(h), \tag{6}$$

where h is a shared latent representation and $\mathcal{D}^{(m)}$ maps to the appropriate spatial support and variable space. This formulation ensures that coupling between subsystems arises through the shared latent dynamics and transport operators, while maintaining consistency with the underlying domain geometry and governing equations. In particular, surface-restricted variables are defined on the boundary manifold, while volumetric variables evolve over the interior domain. Interactions between them arise through boundary conditions and flux exchanges, ensuring consistency with the underlying physical coupling mechanisms.

3.4 Stochastic Latent Dynamics

At resolved scales, ocean evolution is influenced by unresolved subgrid processes and chaotic variability, which cannot be captured deterministically. We model this uncertainty through a latent variable z_t that captures stochastic forcing and unresolved dynamics.

The system evolves according to $x_{t+1} \sim p_\theta(x_{t+1} \mid x_t, z_t, f_t)$, with latent dynamics defined by a prior $p_\theta(z_t \mid x_t, f_t)$, and an approximate posterior $q_\phi(z_t \mid x_{t+1}, x_t, f_t)$.

This formulation can be interpreted as a stochastic closure model, in which unresolved processes are represented through latent variables that modulate the resolved dynamics. The latent prior is not trained for calibrated ensemble dispersion, and we make no probabilistic-forecasting claim (Appendix C.2).

3.5 Training

Training is formulated within a stochastic state-space framework, but the optimization procedure departs from standard end-to-end training due to the geometric, physical, and memory constraints of the three-dimensional ocean system. At each time step, the model defines a latent-variable transition:

$$p_\theta(x_{t+1} \mid x_t, z_t, f_t), \tag{7}$$

with latent dynamics specified by a learned prior and posterior. The training objective is variational, combining a reconstruction term with a Kullback-Leibler divergence:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z_t|\cdot)}\left[\ell(x_{t+1}, \hat{x}_{t+1})\right] + \beta\,\mathrm{KL}\left(q_\phi(z_t \mid \cdot) \,\|\, p_\theta(z_t \mid \cdot)\right), \tag{8}$$

where the reconstruction loss decomposes across variables and spatial support, ensuring consistency with the heterogeneous multi-physics state.

The predictive model is implemented as a composition of graph-based operators acting on the wet-domain representation:

$$\hat{x}_{t+1} = \mathcal{F}_\theta(x_t, z_t, f_t), \tag{9}$$

where $\mathcal{F}_\theta$ combines horizontal message passing, vertical flux operators, and hierarchical projections. In this formulation, graph neural network components act as discrete approximations of transport and coupling operators defined over the irregular ocean domain.

However, full end-to-end optimization of the coupled 3D system is intractable at target resolution due to the combined cost of wet-node count, edge density, and hierarchical activations. We therefore adopt a progressive training strategy that constrains the optimization to a sequence of problems with increasing effective complexity.

Formally, let $\mathcal{F}_\theta = \sum_{i=1}^{N} \mathcal{F}_\theta^{(i)}$ denote a decomposition of the full evolution operator into components associated with distinct spatial interactions and coupling pathways. Training proceeds by optimizing the objective with respect to progressively expanding subsets of parameters:

$$\theta^{(k)} \subset \theta, \quad k = 1, \ldots, N, \tag{10}$$

such that the effective hypothesis space increases over the course of training.

At each stage, parameters associated with previously learned components are kept fixed, while newly introduced components are optimized. This defines a sequence of constrained optimization problems in which additional interactions are learned on top of an already stable dynamical backbone. Horizontal and vertical operators are optimized with separate learning rates (10^{-3} and 10^{-4}, respectively), reflecting their distinct dynamical scales.

In practice, this corresponds to progressively activating additional depth-dependent interactions and coupling mechanisms within the graph, while reusing shared hierarchical representations. This staged procedure enables the model to learn regime-specific dynamics prior to enforcing full vertical and multi-physics coupling, while maintaining a consistent global representation.

This approach stabilizes training and controls memory usage by limiting the number of trainable degrees of freedom at each stage. It also imposes a physically meaningful ordering, in which dominant transport and boundary-driven processes are established first, while subsequent stages introduce deeper vertical interactions that progressively refine cross-variable coupling across levels.

A central requirement is stability under long-horizon integration. While short-term predictions may appear accurate due to the slow evolution of ocean dynamics, extended rollouts expose inconsistencies in transport, boundary forcing, and cross-variable coupling. The combination of conservative transport, surface-constrained forcing, and staged operator learning is designed to maintain physically consistent behavior over extended forecast horizons.

4 Experiment

4.1 Data

We construct the training dataset by combining global ocean reanalysis, biogeochemical hindcast simulations, and atmospheric forcing fields to obtain a consistent, coupled ocean state over the period 1993-2025.

Physical ocean variables are derived from the GLORYS12 reanalysis, providing three-dimensional fields of temperature, salinity, and velocity, as well as surface variables including sea surface height and sea ice. Biogeochemical variables are obtained from the BIOMER hindcast simulation and spatially aligned onto the same grid. These include tracers governing biological productivity, oxygen dynamics, and nutrient cycles. The data are interpolated to a regular 1° horizontal grid, while preserving the vertical discretization and bathymetry-dependent ocean mask, resulting in a domain with depth-varying support. Atmospheric forcing is provided by ERA5 reanalysis, including near-surface thermodynamic and dynamical variables as well as surface pressure, and is also interpolated onto the same 1° grid. These fields act at the air-sea interface and are used as external inputs f_t.

The full system is represented as the coupled state x_t defined in Sec. 3.3, comprising volumetric variables defined over the ocean domain Ω_o and surface variables defined on the boundary $\partial\Omega_o$. The spatial support of each variable is preserved explicitly.

To ensure consistency with the model formulation, we restrict the data to the wet ocean domain $\Omega_o \subset \Omega$, removing all land points prior to training. This induces an irregular domain consistent with coastlines and bathymetry, on which the graph representation is constructed.

Temporal sequences are extracted at fixed intervals to form training tuples (x_t, x_{t+1}, f_t), without additional masking or interpolation, preserving the physical consistency of the original datasets.

Additional details regarding the source datasets, spatial resolutions, preprocessing pipeline, and vertical discretization are provided in Appendix B.1.

Variables. The coupled state spans ocean physics, biogeochemistry, and sea ice, and is driven by atmospheric forcing. The variables used in this work are summarized in Table 1 (Appendix B.1).

4.2 Implementation

The model is trained on monthly ocean reanalysis/hindcast data over 1993-2009, covering 15 coupled physical and biogeochemical variables across 20 vertical levels. The state is represented exclusively on wet nodes using a hierarchical graph, reducing computational cost while preserving the irregular ocean geometry. Due to memory constraints from the large number of wet nodes, edges, and hierarchical activations, training follows a progressive depth-wise strategy, where shallow layers are learned first and deeper layers are introduced incrementally.

Optimization is performed with the Adam optimizer on a single NVIDIA H100 GPU. Training runs for 300,000 iterations with a batch size of 1, requiring approximately 25 hours. The loss is area-weighted to account for the spherical geometry of the Earth and the non-uniform spatial distribution of nodes. During inference, the model is applied autoregressively, feeding predictions back as inputs to generate multi-step forecasts, while the hierarchical latent structure maintains large-scale coherence across rollouts. Further details regarding the graph construction, operator implementation, and progressive training procedure are provided in Appendix B.2. Implementation details and notation are listed in Tables 2 and 3 (Appendix B.2).

4.3 Evaluation Design

We benchmark the proposed model against GLONET (El Aouni et al., 2025b), a data-driven global ocean forecasting system designed for short-term prediction, retrained on our data for long-horizon integration (GLONET†). We also compare against the 2D GNN of Sec. 5.2, which represents the graph-based encode-process-decode family developed for global weather prediction (Lam et al., 2023), extended to limited-area domains (Oskarsson et al., 2023) and adapted to the ocean (Holmberg et al., 2024; 2026). These models operate on a two-dimensional mesh in which variables and depth levels are stacked as feature channels, so wet-domain meshing removes land at the surface but not at subsurface levels, where land points remain in the feature vector and must be masked. All three models are trained on identical data with the same optimizer, learning rate, iterations, loss and seeds. Skill ($1 - \mathrm{RMSE}/\mathrm{RMSE}_{\mathrm{persistence}}$) is measured on integrations run from January 2010, after training, to December 2025, with persistence and a 1993–2009 climatology as baselines. We also report the conservation residual $R(t) = [I(t+1) - I(t)]/\Delta t - F_{\mathrm{surface}}(t)$ of global inventories I, interannual variance ratios, spectral distances and coastal errors (Appendix C.5).

5 Results

We consider two complementary evaluation settings to assess both long-horizon fidelity and generalization under perturbed forcing. In the first setting, the model is initialized in 2001 and integrated autoregressively for 25 years using historical forcing. While this period partially overlaps with the training data, it includes 16 years of effectively unseen states, and is used here for qualitative illustration; all quantitative scores are computed on the held-out integrations described in Sec. 4.3. In the second setting, we assess out-of-distribution behavior through a perturbed forcing experiment. Starting from January 2025, we construct a modified forcing scenario based on ERA5 in which surface temperature is gradually increased by $+3^{\circ}C$ over a 25-year period. The model is integrated under this perturbed forcing and compared to a baseline rollout obtained using unmodified forcing. Results for this setting are reported in Appendix C.4. Additional analyses, including extended 200-year autoregressive integrations, stochastic variability diagnostics, and vertical coupling evaluations, are provided in Appendix C.7 and C.2.

5.1 Long-term Stability and Spectral Consistency

We evaluate the fidelity of the simulated dynamics using spatiotemporal power spectral density (PSD) diagnostics. The PSD is computed as a function of time and radial wavenumber, providing a joint characterization of temporal variability and spatial structure, and capturing the distribution of energy across spatial scales.

Figure 3 (Appendix C.1) shows representative results for temperature (`thetao`) and sea surface height (`zos`), comparing the reanalysis and the model rollout (initialized in 2001) over the full 25-year period. The model reproduces the large-scale structure of the spectrum with high fidelity, preserving both the distribution of energy across spatial scales and its temporal evolution. The autoregressive rollout remains stable throughout the integration, with no evidence of drift or spectral collapse. At higher wavenumbers, differences reflect the effective resolution of the model, with reduced energy content at scales not represented on the 1° grid relative to the higher-resolution reanalysis. Additional PSD comparisons for other variables, including biogeochemical tracers and sea ice, are provided in the Appendix C.1.

5.2 Ablation Study

To assess the importance of geometry-aware and depth-resolved dynamics, we compare the proposed model against a two-dimensional graph neural network (2D GNN). In this variant, the ocean state is represented on the same irregular wet-node graph, but the vertical dimension is not explicitly resolved. Instead, depth levels and variables are aggregated into a single feature space, in a manner similar to prior approaches such as (El Aouni et al., 2025b; Wang et al., 2024; Lang et al., 2024; Price et al., 2023; Hirabayashi et al., 2026; Cui et al., 2025), where multi-level information is treated as channels rather than structured spatial axes. This formulation removes the explicit representation

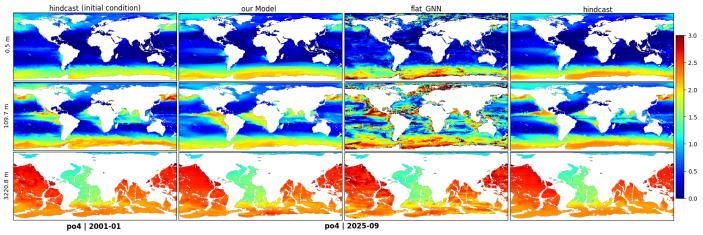


Figure 2: Ablation comparing the proposed model and a 2D GNN for phosphate (po4) at three depths (surface, $\sim$110 m, $\sim$3200 m) after a 25-year rollout initialized in 2001. The 2D GNN shows spurious surface variability and loss of structure with depth, highlighting the importance of explicit depth-resolved dynamics.

of stratification, vertical interfaces, and depth-dependent geometry. As a result, all interactions, both lateral and cross-depth, are learned through generic message passing in feature space, without distinguishing between horizontal transport and vertical exchange.

The comparison is strictly controlled. Both models are trained on the same data using identical optimization settings, including optimizer, learning rate, number of iterations, loss formulation, and random initialization seeds. In particular, the same area-weighted loss is used to account for spherical geometry and the non-uniform distribution of ocean points. The only difference lies in the representation of the state and the structure of the learned operators.

We evaluate both models through identical 25-year autoregressive rollouts starting from the same initial condition. Figure 2 shows the results for a representative biogeochemical tracer (phosphate, `po4`) at three depth levels. The proposed model maintains coherent large-scale structures across all depths, closely matching the reference hindcast even after long integration. In contrast, the 2D GNN exhibits systematic degradation that increases with depth.

At the surface, the 2D GNN introduces spurious small-scale variability and noisy patterns, particularly in dynamically active regions. This reflects the absence of structured transport: without separating lateral advection from vertical processes, the model learns interactions that are not aligned with the dominant physical pathways. At intermediate depth, basin-scale organization deteriorates, with

distorted gradients and weakened large-scale structure. In the deep ocean, the discrepancy becomes more pronounced: the model fails to preserve the large-scale distribution of tracers, producing fragmented and physically inconsistent patterns.

These deficiencies arise from the absence of an explicit depth-resolved representation. In the proposed model, vertical exchanges are mediated through interface-based fluxes and constrained by stratification and bathymetry. In the 2D GNN, cross-depth interactions are implicitly encoded through feature mixing, allowing signals to propagate without respecting the structure of the water column. This leads to spurious coupling between dynamically distinct regimes and a progressive loss of consistency over long rollouts. Further results across variables show the same behavior in preserving spatiotemporal structure; see Appendix C.3 for details.

Overall, this comparison shows that preserving the depth-resolved structure of the ocean system is not a secondary detail. Even under identical training conditions, a model that treats depth as an unstructured feature dimension fails to maintain physically consistent dynamics across vertical regimes and time.

Quantitative comparison. On the held-out integration (Table 7, Appendix C.5), the proposed model retains positive skill against persistence for every variable, depth and lead up to 192 months; the 2D GNN falls below persistence on every field by 192 months, and GLONET† on most fields from the first year. On phosphate, both comparison architectures are below persistence in all nine depth-lead cells (2D GNN -0.057 to -3.75, GLONET† -1.79 to -15.95), whereas the proposed model scores $+0.048$ to $+0.210$. Near coasts, error is 0.85–1.69 times the open-ocean error for the proposed model and 1.43–3.07 times for the 2D GNN (Table 6, Appendix C.5).

Conservation and variance. Salt and the biogeochemical tracers have no surface source, so their residual R should be zero. The proposed model closes every inventory to within 0.35–6.5% per century, while the 2D GNN leaks 65 to 361 times more and systematically creates tracer mass (drift up to $+152\%$; Table 8, Appendix C.5). This closure is not obtained by damping: the proposed model retains 96–100% of reanalysis interannual variance (2D GNN: $0.06\times$ to $13.33\times$), and its kinetic-energy drift matches GLORYS12 (-3.83 vs. $-3.88 \times 10^{-4}\,\mathrm{m^2\,s^{-2}}$ per century; 2D GNN $+0.07$).

6 Conclusion

We present a physics-aligned neural architecture for global ocean forecasting that explicitly encodes the geometric and dynamical structure of the ocean system. By operating on the wet domain, separating horizontal and vertical operators, enforcing interface-based vertical coupling, and incorporating surface boundary conditions for air-sea fluxes, the model avoids common inconsistencies observed in existing data-driven approaches.

The proposed framework produces stable and coherent spatiotemporal dynamics, preserving large-scale structure while maintaining consistent cross-variable interactions. The results further demonstrate that explicitly resolving vertical structure is essential: treating depth as an unstructured feature dimension leads to degraded representations, particularly in deeper layers.

These findings highlight the importance of embedding physical structure directly into model design rather than relying on generic architectures or loss-based constraints. Constraining the hypothesis space to reflect anisotropy, stratification, and boundary-driven dynamics provides a more robust foundation for learning ocean dynamics.

Several directions remain for future work. Extending the framework to higher spatial resolution would enable improved representation of fine-scale processes. Incorporating coupled atmosphere-ocean interactions would further strengthen the ability of the model to capture realistic variability. Finally, developing the stochastic component toward calibrated ensemble forecasting would allow uncertainty to be represented more systematically.

Overall, this work suggests that physically grounded architectural design is a key ingredient for scalable and reliable data-driven ocean forecasting.

Reproducibility Statement

The architecture is specified in Sec. 3, Figure 1 and Algorithm 1 (Appendix B.2). Graph statistics, dimensions, learning rates, training schedule and parameter count are listed in Table 2 (Appendix B.2), and all datasets are public (Appendix B.1). Evaluation protocols, baselines and diagnostic definitions are given in Sec. 4.3.

AI Use Statement

No generative AI tools were used in this work.

References

Niraj Agarwal, Timothy A Smith, Sergey Frolov, and Laura C Slivinski. Skillful global ocean emulation and the role of correlation-aware loss. *arXiv preprint arXiv:2604.18727*, 2026.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D Collins, Michael S Pritchard, and Alexander Keller. Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*, 2025.

Daria Botvynko, Pierre Haslée, Lucile Gaultier, Bertrand Chapron, Clement de Boyer Montégut, Anass El Aouni, Julien Le Sommer, and Ronan Fablet. Neural ocean forecasting from sparse satellite-derived observations: a case-study for ssh dynamics and altimetry data. *arXiv preprint arXiv:2512.22152*, 2025.

Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: A review. *Acta Mechanica Sinica*, 37(12): 1727–1738, 2021.

Eric P Chassignet, Harley E Hurlburt, Ole Martin Smedstad, George R Halliwell, Patrick J Hogan, Alan J Wallcraft, Remy Baraille, and Rainer Bleck. The hycom (hybrid coordinate ocean model) data assimilative system. *Journal of Marine Systems*, 65(1-4):60–83, 2007.

Copernicus Marine Service. Global ocean physics reanalysis (glorys12v1), 2023.

Copernicus Marine Service. Global ocean biogeochemistry hindcast, 2024.

Yingzhe Cui, Ruohan Wu, Xiang Zhang, Ziqi Zhu, Bo Liu, Jun Shi, Junshi Chen, Hailong Liu, Shenghui Zhou, Liang Su, et al. Forecasting the eddying ocean with a deep neural network. *Nature Communications*, 16(1):2268, 2025.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Anass El Aouni, Véronique Garçon, Joël Sudre, Hussein Yahia, Khalid Daoudi, and Khalid Minaoui. Physical and biological satellite observations of the northwest african upwelling: Spatial extent and dynamics. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1409–1421, 2019.

Anass El Aouni, Khalid Daoudi, Khalid Minaoui, and Hussein Yahia. Robust detection of the north-west african upwelling from sst images. *IEEE Geoscience and Remote Sensing Letters*, 18 (4):573–576, 2020.

Anass El Aouni, Ismail Bessa, Karim Hilmi, Hassan Moustahfid, and Avijit Gangopadhyay. Enhanced coastal upwelling indices for moroccan atlantic coast, their force–response variability and sensitivity to extreme events between 1993 and 2021. *Regional studies in marine science*, 77: 103611, 2024.

Anass El Aouni, Quentin Gaudel, Juan Emmanuel Johnson, REGNIER Charly, Julien Le Sommer, Ronan Fablet, Marie Drevillon, Yann Drillet, Pierre Yves Le Traon, et al. Oceanbench: A benchmark for data-driven global ocean forecasting systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a.

Anass El Aouni, Quentin Gaudel, Charly Regnier, Simon Van Gennip, Olivier Le Galloudec, Marie Drevillon, Yann Drillet, and Jean-Michel Lellouche. Glonet: Mercator's end-to-end neural global ocean forecasting system. *Journal of Geophysical Research: Machine Learning and Computation*, 2(3):e2025JH000686, 2025b.

Italo Epicoco, Davide Donno, Gabriele Accarino, Simone Norberti, Alessandro Grandi, Ronan McAdam, Donatello Elia, Emanuela Clementi, Paola Nassisi, Enrico Scoccimarro, et al. Medformer: a data-driven model for forecasting the mediterranean sea. 2025.

Pierre Garcia, Inès Larroche, Amélie Pesnec, Hannah Bull, Théo Archambault, Evangelos Moschos, Alexandre Stegner, Anastase Charantonis, and Dominique Béréziat. Orcast: Operational high-resolution current forecasts. *Artificial Intelligence for the Earth Systems*, 4(4):e250002, 2025.

Hans Hersbach et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Yuta Hirabayashi, Daisuke Matusoka, and Konobu Kimura. Eddy-resolving global ocean forecasting with multi-scale graph neural networks. *arXiv preprint arXiv:2601.12775*, 2026.

Daniel Holmberg, Emanuela Clementi, and Teemu Roos. Regional ocean forecasting with hierarchical graph neural networks. *arXiv preprint arXiv:2410.11807*, 2024.

Daniel Holmberg, Joel Oskarsson, Erik Wikingsson, Fredrik Lindsten, and Teemu Roos. Njord: A probabilistic graph neural network for ensemble ocean forecasting. *arXiv preprint arXiv:2605.15470*, 2026.

Qiusheng Huang, Yuan Niu, Xiaohui Zhong, Anboyu Guo, Lei Chen, Dianjun Zhang, Xuefeng Zhang, and Hao Li. Fuxi-ocean: A global ocean forecasting system with sub-daily resolution. *arXiv preprint arXiv:2506.03210*, 2025.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. Aifs–ecmwf's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1):18, 2026.

Jean-Michel Lellouche et al. The copernicus global 1/12° oceanic and sea ice reanalysis. *Frontiers in Earth Science*, 9:698876, 2021. doi: 10.3389/feart.2021.698876.

Gurvan Madec et al. Nemo ocean engine. 2015.

Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Graph-based neural weather prediction for limited area modeling. *arXiv preprint arXiv:2309.17370*, 2023.

MD Piggott, PE Farrell, CR Wilson, GJ Gorman, and CC Pain. Anisotropic mesh adaptivity for multi-scale ocean modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1907):4591–4611, 2009.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

James F Price, Robert A Weller, and Rebecca R Schudlich. Wind-driven ocean currents and ekman transport. *Science*, 238(4833):1534–1538, 1987.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Lynne D Talley. *Descriptive physical oceanography: an introduction*. Academic press, 2011.

Xiang Wang, Renzhi Wang, Ningzi Hu, Pinqiang Wang, Peng Huo, Guihua Wang, Huizan Wang, Senzhang Wang, Junxing Zhu, Jianbo Xu, et al. Xihe: A data-driven model for global ocean eddy-resolving forecasting. *arXiv preprint arXiv:2402.02995*, 2024.

Anne-Cathrin Wölfl, Helen Snaith, Sam Amirebrahimi, Colin W Devey, Boris Dorschel, Vicki Ferrini, Veerle AI Huvenne, Martin Jakobsson, Jennifer Jencks, Gordon Johnston, et al. Seafloor mapping–the challenge of a truly global ocean bathymetry. *Frontiers in Marine Science*, 6:434383, 2019.

Carl Wunsch and Raffaele Ferrari. Vertical mixing, energy, and the general circulation of the oceans. *Annu. Rev. Fluid Mech.*, 36(1):281–314, 2004.

Nan Yang, Chong Wang, Meihua Zhao, Zimeng Zhao, Huiling Zheng, Bin Zhang, Jianing Wang, and Xiaofeng Li. Langya: Revolutionizing cross-spatiotemporal ocean forecasting. *arXiv preprint arXiv:2412.18097*, 2024.

A Appendix: Limitation and Broader Impact

A.1 Limitation

Despite the strong performance observed across long-horizon rollouts and perturbed forcing experiments, several limitations remain and motivate future work.

First, while the model operates on a coupled ocean state including physical, biogeochemical, and sea ice variables, the training data for biogeochemistry is derived from a hindcast simulation rather than reanalysis products. As a result, the learned dynamics may inherit biases from the underlying numerical model, particularly for biogeochemical processes. Incorporating observational datasets and performing targeted fine-tuning will be important to improve the realism of these components.

second, the evaluation setup relies on prescribed atmospheric forcing, which is not dynamically coupled to the ocean state. In particular, the perturbed warming experiment modifies only surface temperature while keeping other atmospheric variables fixed, whereas in reality such changes would be accompanied by adjustments in winds, pressure, and fluxes. This limitation arises from the current model formulation, where the atmosphere acts as an external driver rather than an interactive component. Extending the framework toward a fully coupled ocean–atmosphere system would enable more physically consistent simulations and allow the study of responses under realistic climate forcing scenarios, including different carbon emission trajectories.

Third, the vertical representation is limited to 20 depth levels. While sufficient for large-scale structure, this discretization constrains the representation of fine-scale vertical processes such as stratification, mixing layers, and deep water mass transformations. Increasing vertical resolution and improving sampling across depth would enhance the fidelity of the learned dynamics.

Further, a closure residual for heat cannot be computed, because the prescribed forcing carries no net air–sea heat flux; heat is therefore reported as drift against reanalysis only (Table 10). The stochastic latent is not trained for calibrated dispersion, and the resulting ensembles are under-dispersive (Appendix C.2). Finally, the figures of Sec. 5 come from integrations initialized in 2001, which overlap the training period, whereas all reported skill, residual and spectral-distance numbers come from post-2009 integrations.

A.2 Broader Impact

This work introduces a data-driven framework for simulating the coupled ocean system, with explicit treatment of geometry, anisotropic transport, and multi-component interactions across physics, sea ice, and biogeochemistry. By embedding physical structure directly into the architecture, the model enables stable long-horizon rollouts and controlled perturbation experiments at a computational cost significantly lower than traditional numerical systems.

One immediate implication is the ability to study the internal consistency of ocean dynamics over extended timescales. Unlike short-term forecasting systems, which primarily optimize predictive skill over limited horizons, the proposed approach allows direct inspection of how energy, tracers, and coupled processes evolve over decades. This can support the analysis of large-scale circulation changes, vertical transport pathways, and cross-component interactions, particularly in regimes where repeated simulation is required.

The efficiency of the model also lowers the barrier to running multi-decadal experiments. This is particularly relevant for sensitivity analyses, where multiple scenarios must be explored to understand the response of the system to external perturbations. In this work, we demonstrate such capability through controlled forcing experiments, highlighting how the model can be used to probe responses outside the training distribution while maintaining coherent coupling across variables.

At the same time, these capabilities introduce important limitations and risks. The model learns from reanalysis and hindcast products, and therefore reflects the assumptions and biases of the underlying numerical systems. This is especially relevant for biogeochemical components, where observational constraints are limited and uncertainty is higher. As a result, simulated responses,particularly under perturbed or out-of-distribution forcing, should not be interpreted as physically accurate projections, but rather as model-consistent scenarios.

More fundamentally, the current framework does not represent a fully coupled Earth system. Atmospheric variables act as prescribed drivers, and do not respond to ocean state changes. This limits the physical consistency of long-term experiments, especially when exploring strong perturbations such as sustained warming. Extending the approach toward fully coupled ocean–atmosphere dynamics is therefore a necessary step for applications involving climate-scale projections or policy-relevant scenarios.

Finally, while the model exhibits stable behavior over long rollouts, its use in operational or decision-making contexts requires careful validation. The ability to generate coherent long-term simulations may create a false sense of reliability if not accompanied by appropriate uncertainty quantification and physical diagnostics. In this sense, the model should be viewed as a complementary tool for analysis and hypothesis testing, rather than a replacement for established numerical forecasting systems.

Overall, this work highlights both the potential and the current boundaries of physics-aligned machine learning for ocean modeling, emphasizing the importance of combining structural inductive biases with careful validation when extending such systems beyond the training regime.

B Appendix: Datasets and Model Architecture

B.1 Datasets

We use a combination of global ocean reanalysis, biogeochemical hindcast, and atmospheric forcing datasets to construct the training data.

Physical ocean variables are derived from the GLORYS12 reanalysis product (Copernicus Marine Service, 2023; Lellouche et al., 2021), which provides a global, eddy-resolving reconstruction of ocean state variables including temperature, salinity, velocity, and sea surface height. GLORYS12 is produced within the Copernicus Marine Service and combines numerical modeling with data assimilation of in situ and satellite observations. The native horizontal resolution is $1/12^\circ$, with 50 vertical levels.

Biogeochemical variables are obtained from the BIOMER dataset (Copernicus Marine Service, 2024), a multi-year global hindcast that includes tracers such as dissolved oxygen, nutrients, and biological productivity. Unlike the physical reanalysis, this product is not directly constrained by observations through data assimilation and therefore reflects the dynamics of the underlying numerical model. The dataset is provided at a native horizontal resolution of $1/4^\circ$.

Atmospheric forcing fields are taken from the ERA5 reanalysis dataset (Hersbach et al., 2020), which provides global, high-resolution estimates of atmospheric variables such as near-surface temperature, winds, and pressure. These fields are used as external drivers of ocean dynamics.

All datasets are interpolated onto a common $1^\circ \times 1^\circ$ horizontal grid to define the model resolution. The vertical discretization is preserved, but restricted to a subset of 20 depth levels spanning the water column from approximately 0.5 m to 3200 m. The selected depth levels (in meters) are:

0.5, 6.4, 11.4, 21.6, 40.3, 77.9, 92.3, 109.7, 130.7, 155.9, 186.1, 266.0, 380.2, 541.1, 763.3, 1062.4, 1452.3, 1941.9, 2533.3, 3220.8.

This results in a depth-resolved ocean state defined on an irregular domain with bathymetry-dependent support, consistent with the wet-point representation used in the model.

B.2 Model Architecture

The overall architecture and training procedure are summarized in Figure 1 and Algorithm 1. The figure provides a conceptual view of the physics-aligned structure, while the algorithm specifies its implementation within a graph-based framework.

The proposed model follows a graph-based encode-process-decode architecture operating on a bathymetry-aware representation of the ocean domain. The full ocean state, defined on a structured grid, is first projected onto a set of wet points using a grid-to-mesh (G2M) mapping. This projection is implemented through level-dependent multilayer perceptrons (MLPs), allowing each depth level to encode distinct physical regimes and variable-specific information.

Category	Variables	Source	Description
Ocean physics	thetao, so, uo, vo, zos	GLORYS12	Momentum, thermodynamic tracers (20 depth levels, 0.49-3220.82 m), and surface pressure-related variable
Biogeochemistry	chl, no3, nppv, o2, po4, si	BIOMER	Nutrients and biological tracers defined over the same 20 depth levels
Sea ice	sithick, siconc, usi, vsi	GLORYS12	Ice thickness, concentration, and velocity (surface only)
Atmospheric forcing	t2m, u10, v10, mslp	ERA5	Near-surface temperature, winds, and pressure

Table 1: Variables used in the coupled ocean forecasting system. Momentum, thermodynamic and biogeochemical variables are defined on 20 vertical levels spanning depths from 0.49 m to 3220.82 m.

The resulting node embeddings define a mesh representation on which the dynamics are learned. The core of the model is a processor composed of graph-based operators that reflect the anisotropic structure of ocean dynamics. Horizontal interactions are modeled through message passing between neighboring nodes using an interaction network, capturing lateral transport processes. Vertical interactions are modeled through local couplers acting between adjacent layers, implemented as MLPs operating on pairs of vertically aligned nodes. These couplers approximate interface-based flux exchanges and enforce structured vertical communication across the water column.

External atmospheric forcing is incorporated through dedicated embeddings and injected exclusively at the surface layer, consistent with the physical role of air-sea interactions. A stochastic latent variable z_t is sampled at each time step and used to modulate the dynamics, providing a representation of unresolved processes and subgrid variability.

The updated mesh representation is mapped back to physical space through a mesh-to-grid (M2G) decoder, producing predictions for all variables on their respective spatial supports.

Training is performed using a progressive depth-wise strategy. The model is optimized over an increasing set of active layers and coupling pathways, while parameters associated with previously learned components are kept fixed. During this process, vertical couplers may operate between layers with heterogeneous levels of optimization, where one side of the interaction corresponds to a representation that is not yet fully adapted. This induces a gradual refinement of cross-layer interactions, allowing the model to stabilize horizontal transport and boundary-driven dynamics before enforcing fully consistent depth-resolved coupling.

From an optimization perspective, this staged procedure reduces memory requirements and improves stability by limiting the number of trainable parameters at each stage. Separate learning rates are used for horizontal and vertical operators to reflect their distinct dynamical scales and convergence properties.

C Appendix: Additional Results

C.1 PSD for Additional Variables

The spectral agreement is quantified on the held-out integration in Table 6. The mean log-spectral L_1 distance to reanalysis is 0.017–0.175 for the proposed model, against 0.94–2.57 for the 2D GNN, and its trend is an order of magnitude or more smaller on every field (at most 0.131 per decade, and 3.9×10^{-4} for surface temperature), whereas that of the 2D GNN grows at $+0.69$ to $+2.01$ per decade. At 110 m, the annual temperature distance of the proposed model rises from 0.036 to a plateau near 0.20 within about six years, whereas that of the 2D GNN keeps growing, reaching 2.11 after 16 years.

We provide additional spatiotemporal power spectral density (PSD) diagnostics for variables not shown in the main text, including biogeochemical tracers and sea ice. These diagnostics follow the same methodology as described in Section 5, evaluating the distribution of energy across spatial scales and its temporal evolution over the 25-year autoregressive rollout.

Algorithm 1 Progressive training

Require: (x_t, x_{t+1}, f_t), wet graph $\mathcal{G}$, stages $\{\mathcal{K}_s\}_{s=1}^{S}$
Require: G2M, M2G, $\mathcal{H}_\theta$, $\mathcal{V}_\theta$, S_θ, $\mathcal{Z}_\theta$
for $s = 1, \ldots, S$ **do**
  activate $\mathcal{K}_s$ and associated vertical couplers
  freeze previously trained components
  for $n = 1, \ldots, N_s$ **do**
    sample (x_t, x_{t+1}, f_t)
    $h_t \leftarrow \mathrm{G2M}(x_t)$
    $e_t^f \leftarrow S_\theta(f_t)$
    $z_t \sim q_\phi(z_t \mid x_{t+1}, x_t, f_t)$
    $h_t^H \leftarrow \mathcal{H}_\theta(h_t; \mathcal{G}, \mathcal{K}_s)$
    $h_t^V \leftarrow \mathcal{V}_\theta(h_t; \mathcal{K}_s)$
    $\hat{h}_{t+1} \leftarrow h_t + \Delta t \left[h_t^H + h_t^V + \delta_{k1} e_t^f + \mathcal{Z}_\theta(z_t, h_t) \right]$
    $\hat{x}_{t+1} \leftarrow \mathrm{M2G}(\hat{h}_{t+1})$
    $\mathcal{L} \leftarrow \ell(x_{t+1}, \hat{x}_{t+1}) + \beta\, \mathrm{KL}(q_\phi(z_t \mid x_{t+1}, x_t, f_t) \,\|\, p_\theta(z_t \mid x_t, f_t))$
    update active parameters
  end for
end for

Table 2: Implementation details of the proposed model.

Quantity	Value
Wet 3D ocean nodes	741,006
Wet surface nodes	41,564
Vertical levels	20
Horizontal grid	170×360
Mesh hierarchy levels	3
Mesh nodes by level	86,070; 539; 55
Same-level horizontal mesh edges	321,127; 1,898; 174
Hierarchy up edges	4,844; 451
Hierarchy down edges	4,844; 451
Adjacent-layer vertical interfaces	699,442
Adjacent-layer vertical messages	1,398,884
Total $\mathcal{H}$+$\mathcal{V}$ message edges	1,732,673
Horizontal connectivity	polar-aware horizontal mesh; north/south polar patching; land-crossing edges rejected
Horizontal radius rule	projection radius uses same-depth horizontal spacing dm; mesh-to-grid accepts same-depth edges with distance $< 4\times$dm
Vertical connectivity	local couplers between vertically aligned adjacent wet layers
Vertical update	interface-based flux divergence with MLP couplers on adjacent-layer node pairs
Message-passing layers	processor=3, encoder=2, prior=2
Hidden dimension	256
Latent dimension	256
KL weight β	1.0
Horizontal operator LR	10^{-3}
Vertical operator LR	10^{-4}
Progressive stages	depth-band stages over the full 0:20 ocean column
Stage length	300-400 epochs per depth-band stage
Total parameters	154,254,076
Surface variables	zos, sithick, siconc, vsi, usi
Volume variables	uo, vo, so, thetao, chl, no3, nppv, o2, po4, si across wet depth levels

Figure 4 shows representative results for dissolved oxygen (o_2) and sea ice thickness (`sithick`). Consistent with the results presented for temperature and sea surface height in Figure 3, the model preserves the large-scale spectral structure and maintains stability over long integration periods.

Table 3: Notation.

Symbol	Meaning
x_t	Coupled ocean state at time t, $x_t = \{q_t^{(m)}\}_{m=1}^{M}$
$q_k^{(m)}$	Prognostic variable m at vertical level k
f_t	Prescribed atmospheric forcing (2 m temperature, 10 m winds, sea-level pressure)
Ω_o, $\partial\Omega_o$	Wet ocean volume and its surface
$\mathcal{H}_\theta$, H_k	Learned horizontal transport operator and its tendency at level k
$\mathcal{V}_\theta$	Learned vertical operator (interface-flux divergence)
$F^q_{k+1/2}$	Flux of q across the interface between levels k and $k+1$
$\Delta\zeta_k$, Δt	Thickness of level k; time step (one month)
$S_\theta(f_t)$, δ_{k1}	Forcing embedding; Kronecker delta restricting it to the surface layer
$\mathcal{D}^{(m)}$, h	Variable-specific decoder; shared latent representation
z_t	Stochastic latent; prior $p_\theta(z_t \mid x_t, f_t)$, posterior $q_\phi(z_t \mid x_{t+1}, x_t, f_t)$
β	Weight of the KL term
$\mathcal{K}_s$	Depth band activated at training stage s
$I(t)$, $R(t)$	Global inventory and its conservation residual (Sec. 4.3)

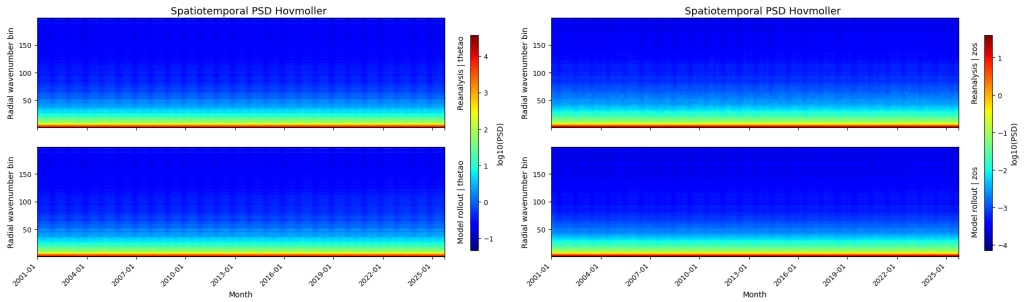


Figure 3: Spatiotemporal power spectral density (PSD) comparison between reanalysis and model rollout for temperature and sea surface height over a 25-year period.

Minor attenuation of high-wavenumber variability is observed, indicating a slight smoothing effect at smaller spatial scales.

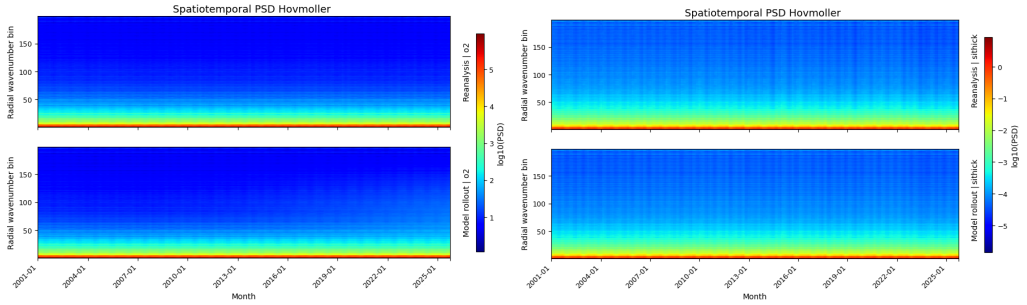


Figure 4: Spatiotemporal power spectral density (PSD) comparison between reanalysis and model rollout (initialized in 2001) for dissolved oxygen and sea ice thickness over a 25-year period.

C.2 Vertical coupling and uncertainties

To further assess the stability and physical consistency of the stochastic formulation, we analyze variability across independent realizations using an upwelling diagnostic. Specifically, we compute a coastal upwelling index based on sea surface temperature (SST) gradients in the Canary upwelling system, defined as the difference between warm offshore waters and colder coastal upwelled waters (El Aouni et al., 2020; 2019).

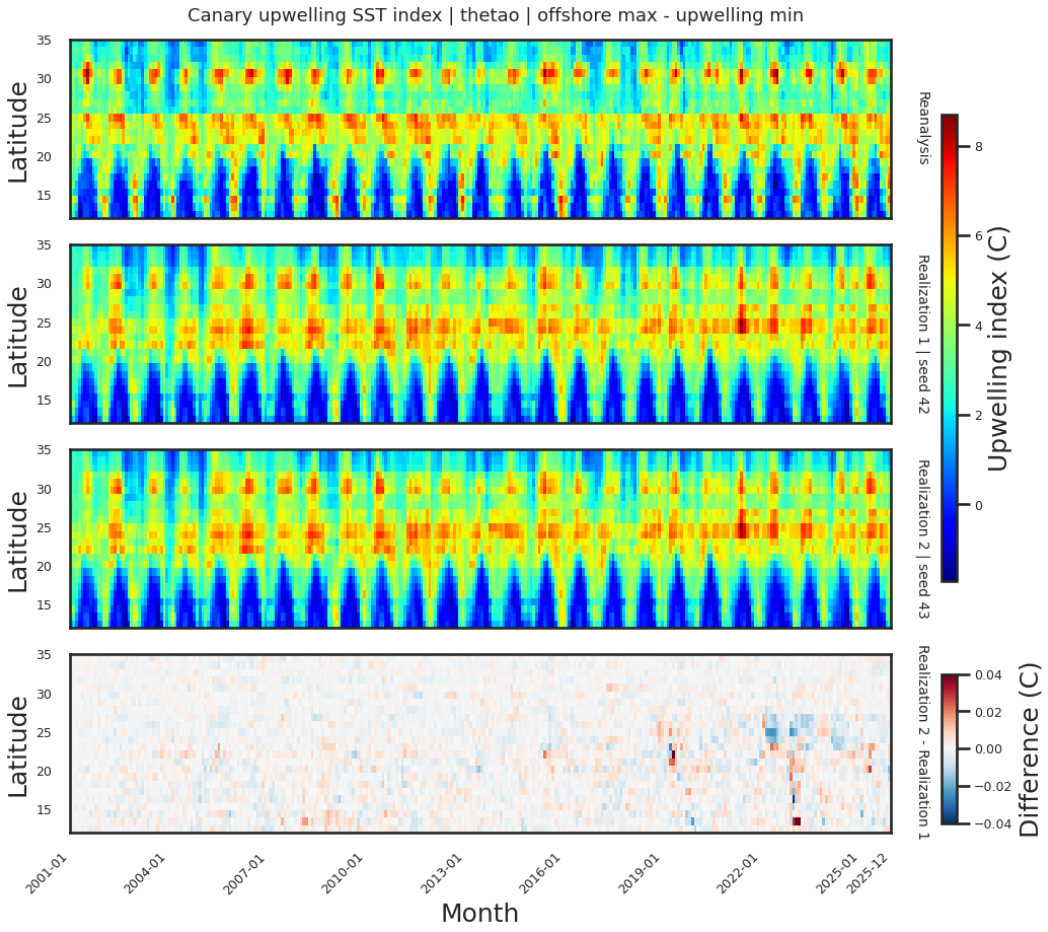


Figure 5: Spatiotemporal evolution of a coastal upwelling index in the Canary region, defined from sea surface temperature as the difference between offshore maximum and coastal minimum values. The top panel shows the reference reanalysis, followed by two stochastic realizations of the model initialized from the same state. Both realizations reproduce the seasonal cycle and latitudinal structure of upwelling, indicating consistent representation of wind-driven Ekman transport and vertical exchange. The bottom panel shows the difference between realizations, highlighting limited variability relative to the signal magnitude and the absence of drift or instability under stochastic integration.

Upwelling systems provide a sensitive test of model dynamics, as they arise from the interaction between wind-driven Ekman transport (Price et al., 1987; El Aouni et al., 2024), coastal divergence, and vertical advection of cold subsurface waters. As such, they reflect both the response to atmospheric forcing and the representation of vertical coupling processes.

Figure 5 shows the spatiotemporal evolution of the upwelling index for the reanalysis and two stochastic realizations of the model. Both realizations reproduce the seasonal cycle and large-scale latitudinal structure of the upwelling signal, indicating that the model consistently captures the underlying physical mechanisms. Variability between realizations remains limited and spatially coherent, with no evidence of divergence or instability over time.

The difference between realizations is small compared to the magnitude of the signal, suggesting that stochasticity does not disrupt the large-scale dynamical balance. At the same time, localized differences indicate the presence of variability associated with unresolved processes, as expected from the stochastic formulation.

We note that the latent prior is not explicitly trained to optimize probabilistic metrics such as CRPS, and therefore does not aim to produce calibrated ensemble dispersion. As such, this experiment should be interpreted as a proof of concept demonstrating stable stochastic dynamics rather than a fully developed probabilistic forecasting system. Future work will explore training strategies that explicitly target ensemble consistency and uncertainty quantification.

To quantify this, we ran ten stochastic members (latent sampled from the prior, seeds 0–9) and one deterministic member (prior mean) of the held-out 2010–2025 integration; all eleven completed the 192 monthly steps. Table 4 summarizes spread and calibration. The ensemble is under-dispersive: the spread-to-skill ratio is below 0.01 at 12 months and 0.14–0.42 at 192 months, against 1 for a calibrated ensemble, and only 1.1–8.1% of verifying values fall strictly inside the ensemble range. The ensemble mean improves on a single member by 0–7% in RMSE.

Table 4: Ten-member ensemble on the held-out integration at leads of 12, 60 and 192 months. Spread/skill ratio is ensemble spread over ensemble-mean RMSE (1 is calibrated). Interior rank fraction is the share of verifying values that fall strictly inside the ensemble range.

	Spread/skill ratio			CRPS			Interior rank
Field	12	60	192	12	60	192	fraction
`zos`, surface	5.43×10^{-4}	0.178	0.206	0.048	0.054	0.060	0.022
`thetao`, surface	3.07×10^{-4}	0.110	0.138	0.370	0.340	0.374	0.023
`thetao`, 110 m	0.003	0.251	0.301	0.936	1.077	1.168	0.067
`po4`, surface	4.758×10^{-4}	0.140	0.162	0.044	0.062	0.071	0.023
`po4`, 110 m	0.010	0.395	0.424	0.097	0.144	0.164	0.081
`siconc`, surface	2.148×10^{-5}	0.072	0.211	0.030	0.030	0.031	0.011

C.3 Additional Ablation

Figure 6 provides a complementary view of model behavior using temperature fields across multiple depths and sea ice concentration, together with their associated spectra. For temperature, the proposed model remains stable over long integrations, retaining coherent patterns and realistic variability throughout the water column. By contrast, the 2D GNN develops increasingly irregular structures, with visible noise near the surface and a marked loss of organization in deeper layers. The spectral diagnostics further indicate a shift toward higher-frequency variability in the 2D GNN, suggesting an imbalance in how energy is distributed across scales when vertical structure is not explicitly represented.

For sea ice concentration, the proposed model preserves the large-scale extent and sharp gradients, maintaining consistent structures in polar regions over time. In contrast, the 2D GNN introduces spurious variability and fragmentation, particularly along the ice edges. The corresponding spectra show enhanced small-scale variability and temporal drift, indicating reduced stability when surface processes are not properly represented.

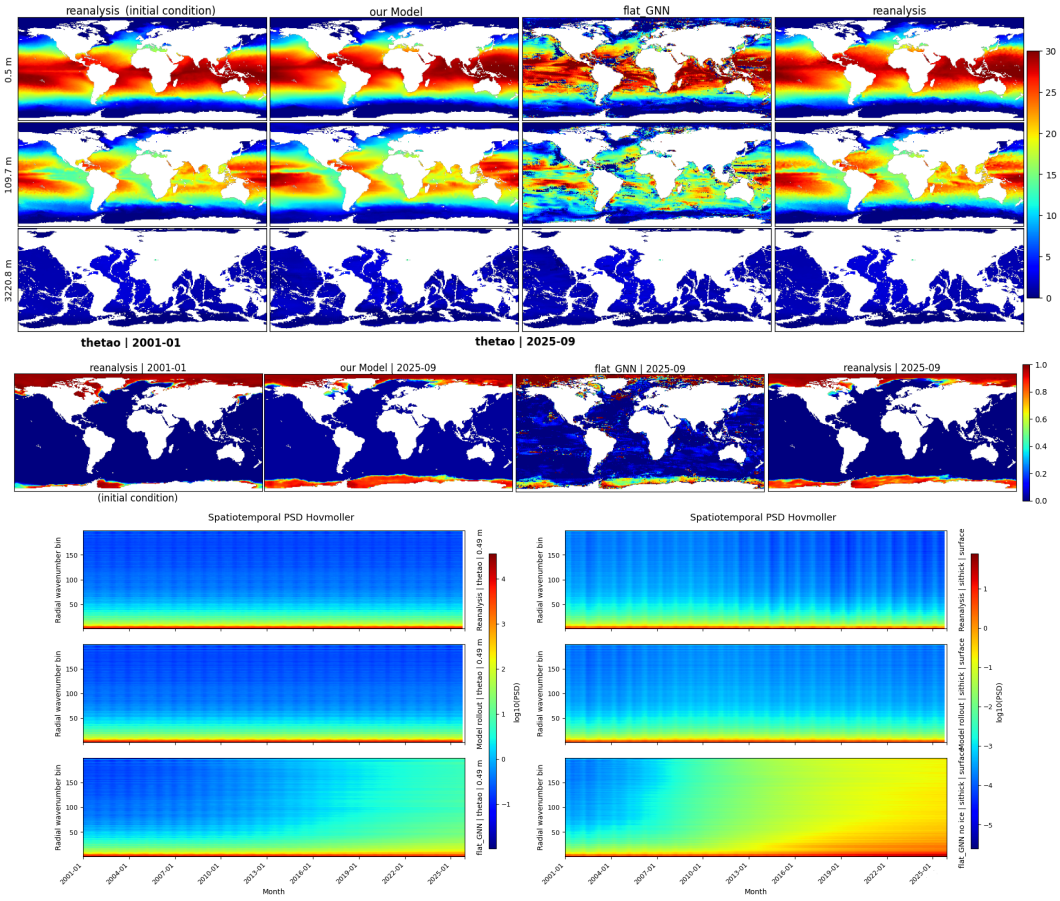


Figure 6: Long-term rollouts (initialized in 2001) of ocean temperature (thetao) at three depths, comparing the proposed model with a 2D GNN, and reanalysis. The proposed model preserves large-scale structures across depths, while the 2D GNN shows increasing noise and degradation, especially at depth. Bottom: spatiotemporal PSD, showing good agreement with reanalysis for the proposed model and excess small-scale energy for the 2D GNN

C.4 RESPONSE TO EXTREME WARMING

We next evaluate the response of the model under an imposed warming scenario. Starting from identical initial conditions, the model is integrated over 25 years using both baseline forcing and a perturbed forcing in which only the surface temperature is increased by $+3^\circ$ over the integration period. This setup is intentionally simplified and does not modify other atmospheric variables, such as wind or pressure, which would co-evolve in a physically consistent climate scenario. The experiment is therefore not designed to be realistic, but rather to probe the model's response to a controlled out-of-distribution perturbation.

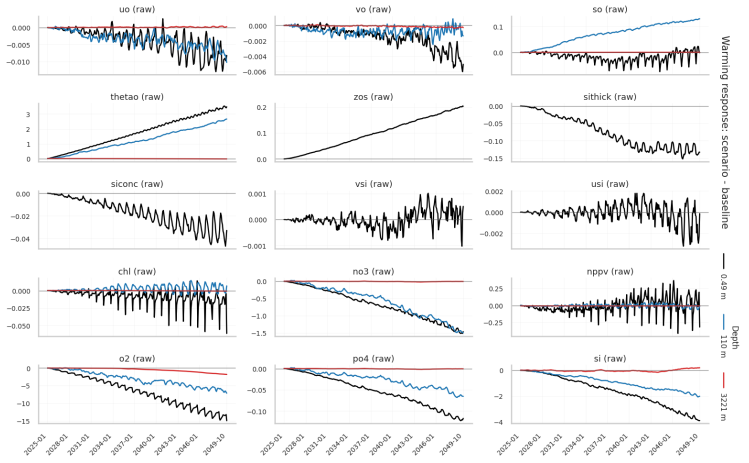


Figure 7: Response of the coupled ocean–sea ice–biogeochemical system to imposed warming, showing coherent large-scale physical adjustments, sea ice decline and increased mobility, and structured biogeochemical changes without loss of stability.

Results are shown in Figure 7. The imposed warming induces a coherent system-wide response. Temperature increases throughout the water column, with a clear attenuation with depth, consistent with vertical propagation of the surface signal, which attenuates with an e-folding depth of 201 m. This is accompanied by adjustments in circulation and sea surface height, indicating large-scale dynamical reorganization.

Sea ice exhibits a pronounced response. Both ice concentration and thickness decrease over time, consistent with sustained warming. Concurrently, ice velocities increase, indicating enhanced mobility associated with a thinner and less consolidated ice cover. This behavior is consistent with reduced mechanical strength and increased deformation, leading to a more dynamic ice regime.

Biogeochemical variables also exhibit structured responses. Oxygen and nutrient concentrations decrease over time, while primary productivity displays increased variability. Chlorophyll shows a depth-dependent response, with localized increases at some levels and decreases at others, consistent with changes in stratification and nutrient availability. Despite the simplified and amplified forcing, the system remains stable over the integration, and interactions across physical, sea ice, and biogeochemical components remain coherent. This indicates that the model preserves physically consistent cross-variable coupling and produces stable responses to perturbations outside the training distribution.

To test how the response scales with the imposed perturbation, we repeated the experiment at $+1$ and $+2\,^\circ$C from the same initial state and protocol (Table 5). Temperature, oxygen and nutrient responses deviate from linear scaling by at most 1.25%, and chlorophyll by 3.25%, so the $+3\,^\circ$C experiment does not sit in a degenerate regime. Sea ice departs from linearity in the physically expected direction, each additional degree removing more ice than the last, and differently in the two hemispheres: -12.9% for Arctic ice area against -2.3% for the Antarctic. Because only surface temperature is perturbed while winds and pressure are unchanged, these experiments characterize the structure of the response, not its magnitude.

Table 5: Response (perturbed minus baseline) to imposed surface warming of $+1$, $+2$ and $+3\,^{\circ}\mathrm{C}$ (native units; ice area in $10^{12}\,\mathrm{m}^2$). Deviation from linear compares the $+2\,^{\circ}\mathrm{C}$ response with the midpoint of the $+1$ and $+3\,^{\circ}\mathrm{C}$ responses; 0 means linear scaling, and a negative value for ice means that each additional degree removes more ice than the last.

Variable	Depth	$+1\,^{\circ}\mathrm{C}$	$+2\,^{\circ}\mathrm{C}$	$+3\,^{\circ}\mathrm{C}$	Dev. from linear [%]
thetao	surface	0.93	1.86	2.78	0.268
thetao	110 m	0.73	1.47	2.205	0.187
thetao	2000 m	−0.035	−0.069	−0.1041	−0.791
siconc	surface	−0.0029	−0.0058	−0.009978	−11
sithick	surface	−0.0065	−0.014	−0.02434	−10.2
o2	surface	−3.5	−7.05	−10.78	−1.25
o2	110 m	−2.1	−4.3	−6.502	−0.0248
no3	surface	−0.32	−0.66	−0.9847	1.16
po4	surface	−0.029	−0.06	−0.08952	1.24
chl	surface	−0.008	−0.0175	−0.02586	3.25
NH ice area	integral	−0.34	−0.73	−1.308	−12.9
SH ice area	integral	−0.7	−1.43	−2.225	−2.29

C.5 Full Held-Out Results

Table 6: Long-horizon statistics on the held-out 2010–2025 integration: interannual variance relative to reanalysis (1 is ideal), log-spectral L_1 distance to reanalysis and its trend (0 is ideal), and ratio of coastal ($<$200 km from land) to open-ocean RMSE.

	Variance ratio		PSD L_1 (mean)		PSD L_1 trend/decade		Coastal/open RMSE	
Field	Ours	2D-GNN	Ours	2D-GNN	Ours	2D-GNN	Ours	2D-GNN
zos, surface	1.00	0.06	0.086	2.569	0.018	2.009	0.85	1.96
thetao, surface	0.99	1.66	0.017	1.155	3.892×10^{-4}	1.083	1.50	1.95
thetao, 110 m	0.96	0.43	0.167	1.625	0.107	1.495	1.13	1.43
po4, surface	0.99	0.94	0.082	1.534	0.044	1.323	1.37	1.88
po4, 110 m	0.97	3.52	0.175	2.366	0.131	1.802	1.00	1.46
siconc, surface	0.99	13.33	0.084	0.941	−0.021	0.688	1.69	3.07

Table 7: Skill relative to persistence ($1-\mathrm{RMSE}/\mathrm{RMSE}_{\mathrm{pers.}}$; >0 beats persistence) on the held-out 2010–2025 integration. Clim.: 1993–2009 climatology. †GLONET architecture retrained on our data for long-horizon integration (Sec. 4.3). Best per row and lead in bold. Full RMSE for all variables, depths and leads in Table 9.

		12-month lead				192-month lead			
Variable	Depth	Ours	2D-GNN	GLONET†	Clim.	Ours	2D-GNN	GLONET†	Clim.
thetao	surface	**0.530**	0.432	−1.63	−1.51	**0.534**	−3.92	−5.39	−1.50
thetao	200 m	**0.167**	0.123	−1.15	0.110	**0.183**	−4.08	−4.03	0.140
thetao	2000 m	**0.202**	0.117	−2.16	0.160	**0.234**	−2.99	−5.99	0.193
o2	surface	**0.313**	0.277	−1.10	−0.993	**0.285**	−5.65	−2.45	−0.080
chl	surface	**0.212**	0.131	−0.348	−0.927	**0.066**	−18.90	−0.150	−0.748
po4	surface	**0.156**	−0.057	−1.89	−0.367	**0.048**	−2.72	−4.04	−0.267
po4	200 m	**0.094**	−0.073	−1.79	0.046	**0.165**	−3.75	−3.33	0.121
po4	2000 m	**0.063**	−0.463	−6.22	0.014	**0.120**	−1.47	−11.48	0.074
siconc	surface	**0.434**	0.248	−0.831	−0.366	**0.393**	−0.903	−1.41	−0.590
sithick	surface	**0.456**	0.394	−0.540	0.212	**0.165**	−1.48	−0.683	0.034

On the held-out integration, surface-temperature RMSE is 0.532, 0.503 and 0.547 °C at 12, 60 and 192 months, against 1.13–1.17 for persistence and 2.81–2.93 for climatology, with an anomaly correlation of 0.988 at both 12 and 192 months (Table 11). Skill does not depend on the start date: over 817 variable–depth–lead cells, the median ratio of 2015- to 2010-initialized RMSE is 0.999, with a median

Table 8: Conservation residual and drift of global inventories over 191 monthly steps of the held-out integration, in % of inventory per century. The expected residual is zero for all five inventories.

	Mean $\lvert R\rvert$ [% inv./century]		Drift [% inv./century]	
Inventory	Ours	2D-GNN	Ours	2D-GNN
Salt	**0.35**	25.1	**−0.08**	−5.32
PO_4	**2.2**	532	**+1.1**	+12.7
NO_3	**2.8**	881	**−1.4**	+152
O_2	**6.5**	425	**−4.8**	+41.5
Si	**3.0**	1084	**−1.0**	+66.4

absolute difference of 1.1% and a 90th percentile of 3.1% (Table 14). At 192 months, vertical-profile errors peak in the thermocline for temperature (1.377 °C at 78 m), at 110 m for phosphate, 11 m for chlorophyll and 380 m for oxygen (Table 12).

All results in this subsection are from the integration initialized in January 2010 and run to December 2025 (Sec. 4.3), except Table 14, which compares it with the January 2015 initialization.

Table 9: RMSE (native units) at leads of 12, 60 and 192 months for all evaluated variables, depths and models. Persistence skill (Table 7) is $1 - \mathrm{RMSE}/\mathrm{RMSE}_{\mathrm{Persistence}}$. †GLONET architecture retrained on our data for long-horizon integration.

		surface			200 m			2000 m		
Variable	Model	12	60	192	12	60	192	12	60	192
thetao	Ours	0.532	0.503	0.547	0.865	1.02	0.887	0.102	0.100	0.102
	2D-GNN	0.643	1.37	5.77	0.911	1.57	5.51	0.113	0.192	0.534
	GLONET†	2.97	4.94	7.50	2.24	3.61	5.46	0.405	0.677	0.935
	Climatology	2.84	2.81	2.93	0.925	1.07	0.933	0.108	0.106	0.108
	Persistence	1.13	1.16	1.17	1.04	1.15	1.09	0.128	0.133	0.134
vo	Ours	0.100	0.105	0.108	0.079	0.077	0.079	0.031	0.031	0.031
	2D-GNN	0.105	0.143	0.408	0.086	0.097	0.186	0.035	0.040	0.085
	GLONET†	0.110	0.120	0.119	0.087	0.091	0.092	0.033	0.033	0.036
	Climatology	0.121	0.119	0.121	0.083	0.081	0.083	0.033	0.032	0.033
	Persistence	0.140	0.134	0.140	0.112	0.109	0.111	0.042	0.042	0.042
o2	Ours	5.56	6.43	8.28	12.48	13.56	17.14	2.48	5.20	9.98
	2D-GNN	5.85	17.03	77.05	13.13	28.36	91.26	2.65	5.86	14.21
	GLONET†	17.03	26.25	39.94	30.70	45.14	60.86	18.18	39.73	46.07
	Climatology	16.14	14.30	12.51	13.35	14.28	18.04	3.13	5.76	10.51
	Persistence	8.10	9.12	11.58	13.08	15.82	18.71	2.54	5.43	10.21
chl	Ours	0.132	0.158	0.183	0.035	0.038	0.039	4.693×10^{-4}	6.904×10^{-4}	8.284×10^{-4}
	2D-GNN	0.145	0.625	3.91	0.044	0.121	0.474	4.94×10^{-4}	0.0020	0.0069
	GLONET†	0.226	0.214	0.226	0.049	0.050	0.059	0.0018	0.0023	0.0027
	Climatology	0.323	0.337	0.343	0.037	0.041	0.041	9.35×10^{-4}	9.74×10^{-4}	8.91×10^{-4}
	Persistence	0.167	0.183	0.196	0.042	0.042	0.050	5.48×10^{-4}	7.07×10^{-4}	8.61×10^{-4}
po4	Ours	0.068	0.092	0.101	0.107	0.120	0.137	0.017	0.021	0.036
	2D-GNN	0.085	0.136	0.393	0.127	0.250	0.782	0.026	0.046	0.101
	GLONET†	0.234	0.348	0.533	0.330	0.537	0.712	0.128	0.453	0.509
	Climatology	0.111	0.117	0.134	0.113	0.126	0.145	0.017	0.022	0.038
	Persistence	0.081	0.097	0.106	0.118	0.143	0.164	0.018	0.027	0.041
si	Ours	1.98	3.27	3.99	2.46	3.39	4.41	1.59	1.91	3.32
	2D-GNN	2.08	5.04	12.09	2.87	5.50	18.82	1.98	3.89	8.05
	GLONET†	11.79	17.30	24.64	10.83	20.59	30.00	9.41	28.21	33.80
	Climatology	3.71	4.07	4.33	2.79	3.57	4.64	1.67	2.01	3.50
	Persistence	2.41	3.45	4.23	2.54	3.97	5.17	1.69	2.43	3.82
sithick	Ours	0.085	0.094	0.151	–	–	–	–	–	–
	2D-GNN	0.094	0.156	0.449	–	–	–	–	–	–
	GLONET†	0.240	0.315	0.305	–	–	–	–	–	–
	Climatology	0.123	0.141	0.175	–	–	–	–	–	–
	Persistence	0.156	0.156	0.181	–	–	–	–	–	–
siconc	Ours	0.044	0.046	0.047	–	–	–	–	–	–
	2D-GNN	0.059	0.079	0.147	–	–	–	–	–	–
	GLONET†	0.143	0.174	0.186	–	–	–	–	–	–
	Climatology	0.107	0.113	0.123	–	–	–	–	–	–
	Persistence	0.078	0.077	0.077	–	–	–	–	–	–
vsi	Ours	0.0072	0.0070	0.010	–	–	–	–	–	–
	2D-GNN	0.0078	0.012	0.030	–	–	–	–	–	–
	GLONET†	0.017	0.015	0.014	–	–	–	–	–	–
	Climatology	0.014	0.014	0.016	–	–	–	–	–	–
	Persistence	0.013	0.014	0.015	–	–	–	–	–	–

Table 10: Drift per century of global budgets over the held-out integration, for GLORYS12/BIOMER and each model. In parentheses: difference from the reanalysis drift over the same window.

Ocean heat content [10^{24} J/century]				
Source	total	0-100 m	100-700 m	700-3220 m
GLORYS	1.73	0.28	0.64	0.72
Ours	2.10 (0.37)	0.28 (-2.04×10^{-3})	0.69 (0.05)	1.05 (0.33)
2D-GNN	0.82 (-0.90)	0.12 (-0.16)	1.92 (1.28)	-1.32 (-2.04)
GLONET†	-5.84 (-7.56)	-0.92 (-1.20)	-2.32 (-2.95)	-2.48 (-3.20)

Salt [10^{16} psu m^3/century], volume-mean KE [$m^2\ s^{-2}$/century]		
Source	Salt content	Volume-mean KE
GLORYS	1.56	-3.88×10^{-4}
Ours	1.49 (-0.07)	-3.83×10^{-4} (4.81×10^{-6})
2D-GNN	1.95 (0.39)	0.07 (0.07)
GLONET†	-59.96 (-61.52)	-1.46×10^{-3} (-1.07×10^{-3})

PO_4 [10^{16}/century]; NO_3, Si [10^{18}/century]; O_2 [10^{19}/century]				
Source	PO4	NO3	O2	Si
GLORYS	0.66	-0.50	1.33	1.91
Ours	1.31 (0.64)	-0.54 (-0.04)	1.09 (-0.25)	1.83 (-0.08)
2D-GNN	-3.14 (-3.80)	-1.25 (-0.75)	1.70 (0.36)	1.31 (-0.60)
GLONET†	5.78 (5.11)	-1.89 (-1.39)	0.51 (-0.83)	3.32 (1.41)

Ice area [10^{12} m^2/century], ice volume [10^{12} m^3/century]				
Source	NH ice area	SH ice area	NH ice volume	SH ice volume
GLORYS	-4.45	-7.50	-41.72	-9.53
Ours	-4.34 (0.11)	-7.35 (0.14)	-41.02 (0.70)	-9.23 (0.30)
2D-GNN	-8.09 (-3.64)	-11.85 (-4.36)	46.49 (88.21)	-4.53 (5.00)
GLONET†	21.01 (25.46)	24.61 (32.11)	-55.78 (-14.07)	-1.16 (8.37)

Table 11: Bias and anomaly correlation coefficient (ACC; anomalies relative to the 1993–2009 climatology) of the proposed model.

	12 months		60 months		192 months	
Field	Bias	ACC	Bias	ACC	Bias	ACC
`zos`, surface	0.002	0.928	0.001	0.921	-0.002	0.915
`thetao`, surface	0.019	0.988	0.012	0.989	0.031	0.988
`thetao`, 110 m	0.046	0.920	0.020	0.929	0.046	0.910
`po4`, surface	-3.24×10^{-5}	0.961	-0.002	0.942	-1.68×10^{-5}	0.935
`po4`, 110 m	0.001	0.922	0.008	0.908	0.008	0.899
`siconc`, surface	0.005	0.982	0.005	0.983	0.006	0.983

Table 12: Vertical-profile RMSE of the proposed model: mean over all valid levels, and the largest level RMSE with its depth.

	12 months		60 months		192 months	
Variable	Mean	Max [depth]	Mean	Max [depth]	Mean	Max [depth]
`thetao`	0.602	1.050 [110 m]	0.735	1.435 [110 m]	0.741	1.377 [78 m]
`so`	0.149	0.411 [0.49 m]	0.154	0.454 [0.49 m]	0.158	0.442 [0.49 m]
`uo`	0.083	0.115 [0.49 m]	0.094	0.142 [6 m]	0.090	0.149 [6 m]
`vo`	0.071	0.100 [0.49 m]	0.073	0.111 [6 m]	0.074	0.113 [6 m]
`po4`	0.073	0.128 [110 m]	0.087	0.154 [110 m]	0.105	0.161 [110 m]
`o2`	7.807	13.449 [92 m]	10.384	17.454 [110 m]	13.661	18.100 [380 m]
`si`	2.110	2.610 [541 m]	2.965	3.392 [6 m]	4.041	4.958 [541 m]
`chl`	0.050	0.141 [11 m]	0.060	0.204 [11 m]	0.063	0.193 [11 m]

Table 13: Sea-ice edge diagnostics of the proposed model (concentration threshold 0.15). IIEE: integrated ice-edge error.

Lead [months]	1	3	6	12	24	60	120	192
IIEE [10^{12} m^2]	0.277	0.338	0.930	0.670	0.796	0.679	0.721	0.802
Ice-area error [10^{12} m^2]	1.746	1.801	1.664	1.747	1.808	1.758	1.811	1.970
Ice-volume error [10^{12} m^3]	0.061	0.118	−0.028	−0.040	0.061	0.055	0.184	0.723

Table 14: Dependence on the initialization date: ratio of RMSE for the January 2015 initialization to the January 2010 initialization over common variables, depths and leads, and absolute relative difference $|\Delta|$.

Lead [months]	Cells	Median RMSE ratio	Median $\lvert\Delta\rvert$	90th pct. $\lvert\Delta\rvert$
12	203	0.994	1.3%	3.1%
24	204	1.000	1.2%	3.1%
60	205	1.000	1.0%	3.0%
120	205	1.000	0.9%	3.1%
all	817	0.999	1.1%	3.1%

C.6 Learned Vertical Coupling

We diagnose an implied effective vertical diffusivity K_z from the learned vertical redistribution in a $+3\,^{\circ}\mathrm{C}$ warming integration (Table 15). The learned coupling is depth-dependent rather than uniform: it is largest in the mixed layer, smallest in the thermocline, and higher again at intermediate depths. We draw no quantitative comparison with observed mixing coefficients from these values.

Table 15: Implied effective vertical diffusivity of the learned vertical operator by depth band.

Depth band	Samples	Median K_z [$\mathrm{m^2\,s^{-1}}$]	Range [$\mathrm{m^2\,s^{-1}}$]
mixed layer 0-50m	96	2.475×10^{-5}	2.054×10^{-5}–2.866×10^{-5}
thermocline 50-200m	144	5.179×10^{-6}	4.262×10^{-6}–5.961×10^{-6}
intermediate 200-700m	72	1.372×10^{-5}	1.145×10^{-5}–1.578×10^{-5}
deep >700m	120	8.167×10^{-6}	6.543×10^{-6}–9.579×10^{-6}

C.7 200 years rollout

To assess the long-term stability and statistical behavior of the model, we perform an extended integration over 200 years using periodically repeated atmospheric forcing. Specifically, ERA5 forcing fields over the period 1993-2025 are cycled and applied repeatedly, while the model is integrated autoregressively from a fixed initial condition.

This setup removes secular trends in the forcing and provides a controlled setting to evaluate whether the learned dynamics remain bounded and statistically consistent over extended time scales. In particular, it allows us to distinguish between physically meaningful variability and spurious drift arising from accumulated model error.

Figure 8 shows the evolution of representative variables across multiple depths and subsystems. The model maintains coherent large-scale structures throughout the integration, with no evidence of drift or degradation. Surface and interior fields remain consistent with the imposed forcing, and cross-variable relationships are preserved over time, indicating stable coupling within the system.

To further characterize this behavior, we analyze the spatiotemporal power spectral density (PSD) over the full simulation (Figure 9). The spectral distribution remains stationary, with energy consistently distributed across spatial scales. No spectral collapse or accumulation at specific wavenumbers is observed, indicating that the model maintains a stable dynamical regime under repeated forcing.

These results show that the model defines a bounded and stable dynamical system under sustained integration, rather than exhibiting error growth or structural drift. This behavior is consistent with the imposed architectural constraints, which limit unphysical interactions and enforce physically meaningful transport pathways.

The 200-year integration corresponds to 2400 autoregressive steps at monthly resolution and is generated in approximately 6 minutes on a single NVIDIA H100 GPU. This highlights the computational efficiency of the proposed formulation, enabling climate-scale simulations at negligible cost compared to traditional numerical systems. Such efficiency makes it feasible to perform long-horizon integrations and ensemble experiments that would otherwise be prohibitively expensive.

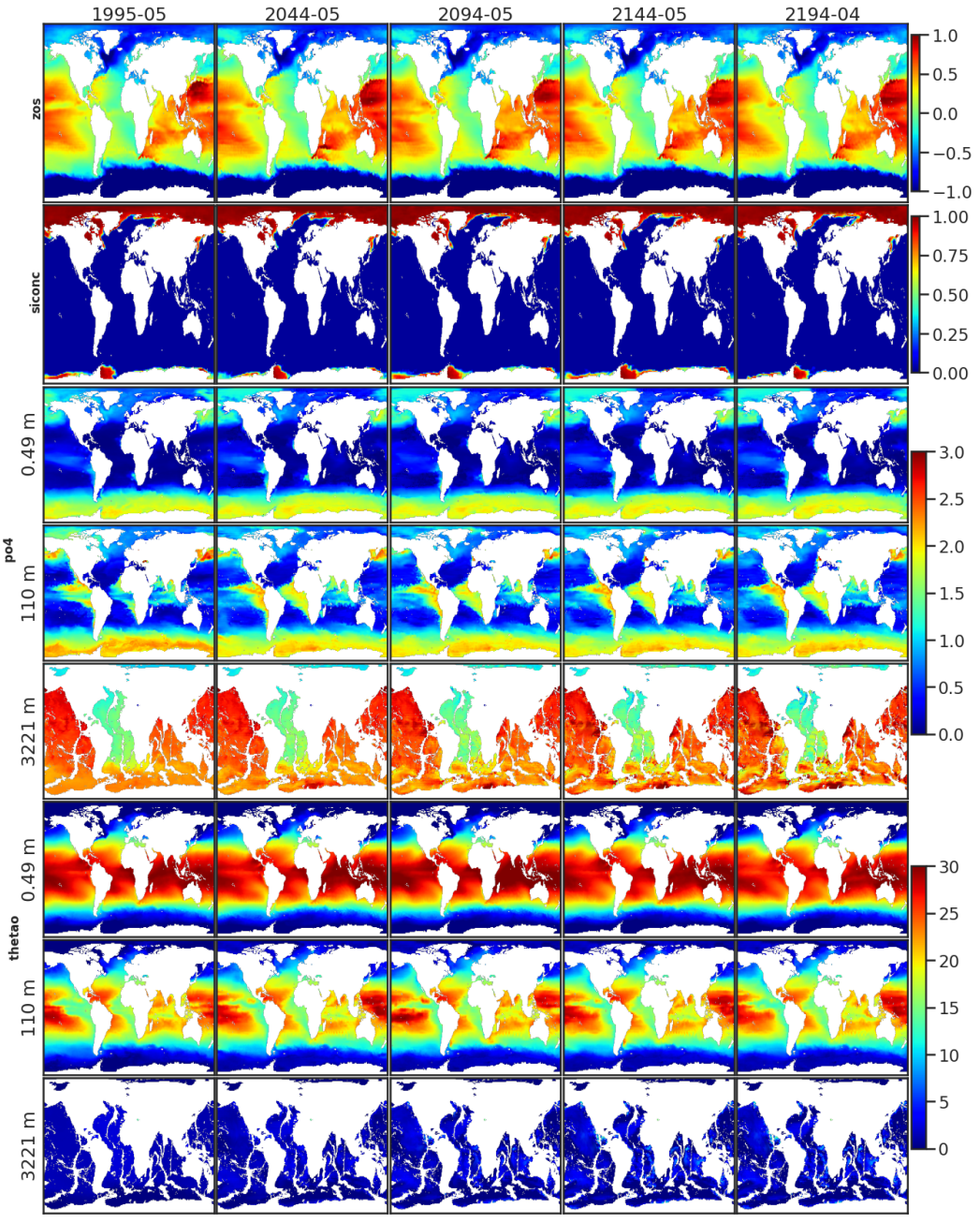


Figure 8: Evolution of the coupled ocean state over a 200-year autoregressive integration under periodically repeated atmospheric forcing. Shown are representative snapshots for sea surface height (`zos`), sea ice concentration (`siconc`), phosphate (`po4`) at multiple depths, and temperature (`thetao`) across the water column. Despite the repeated forcing and extended integration, the model maintains coherent large-scale structures and consistent cross-variable relationships, with no evidence of drift or degradation over time.

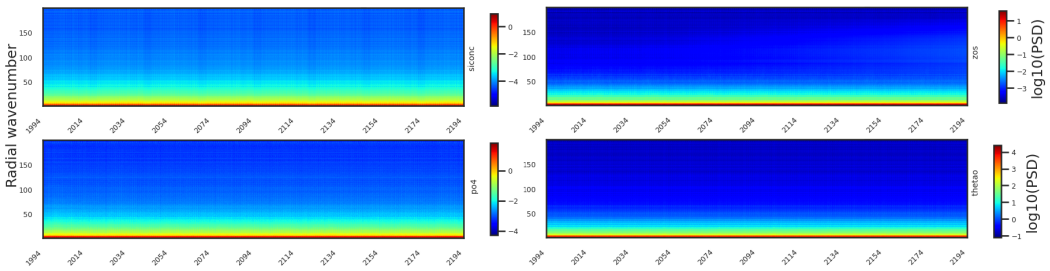


Figure 9: Spatiotemporal power spectral density (PSD) of selected variables over the 200-year integration under periodic forcing. The spectral distribution remains stable over time, with energy consistently distributed across spatial scales and no evidence of spectral collapse or artificial accumulation at specific wavenumbers. This indicates that the model maintains a statistically stationary dynamical regime under sustained autoregressive integration.