PE-TA-VIT: DUAL-RESOLUTION RESIDUAL FORECASTING OF GLOBAL SURFACE AND NEAR-SURFACE OCEAN STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate and computationally efficient forecasting of global surface and near-surface ocean states remains challenging in contemporary oceanography. High-resolution numerical models can better resolve mesoscale dynamics but incur prohibitive computational costs, whereas coarse-resolution models often suffer from biases associated with unresolved subgrid-scale processes. Here, we propose a dual-resolution deep learning framework for forecasting global surface and near-surface ocean states, with near-surface variables represented at 0.49 m depth. The framework adopts a residual reconstruction paradigm in which a Physics-Enhanced Time-Aware Vision Transformer (PE-TA-ViT), incorporating physics-guided inductive biases and adaptive temporal conditioning, predicts coarse-scale spatiotemporal increments at 1° resolution, which are subsequently interpolated to $1/4^\circ$ and added to the evolving high-resolution background state. Over a 9-day forecast horizon, PE-TA-ViT reduces the average Root Mean Square Error (RMSE) relative to persistence by 6.0% to 24.8% across the evaluated ocean variables. It also extends the useful forecast window (ACC > 0.6) for highly dynamic velocity fields by approximately one day. The resulting forecasts retain transient mesoscale structures and large-scale flow patterns, with the deviation in the global energy cascading fraction remaining below 0.009. This framework enables efficient high-resolution forecasting of global surface and near-surface ocean states.

1 INTRODUCTION AND RELATED WORK

The upper ocean mediates the global exchange of heat and momentum (Wunsch & Ferrari, 2004), making accurate forecasting of upper-ocean states important for understanding broader climate dynamics (Large et al., 1994; Donlon et al., 2012) and marine ecosystem responses (Williamson & Guinder, 2021). At the core of these predictive challenges are mesoscale eddies, which represent major reservoirs of oceanic kinetic energy (Chelton et al., 2011; McWilliams, 2016). By strongly influencing the horizontal transport of heat and mass and modulating local geostrophic balances (Zhang et al., 2014; Dong et al., 2014), these transient structures have become an important criterion for assessing the structural fidelity of contemporary global ocean forecasting systems (GOFS) (Cui et al., 2025; El Aouni et al., 2025).

Contemporary high-resolution global ocean forecasting systems (GOFS), including systems based on NEMO (Madec et al., 2022) and HYCOM (Chassignet et al., 2007), increasingly employ eddy-permitting or eddy-resolving grids (e.g., around $1/12^\circ$) to represent mesoscale structures. However, the associated computational overhead can become prohibitive for long-term climate integrations and operational ensemble forecasting (Li et al., 2026; Fox-Kemper et al., 2019). At coarser resolutions, such as approximately 1° configurations used in computationally constrained or climate-scale ocean modeling, unresolved mesoscale processes must instead be represented through empirical subgrid parameterizations for eddy-induced mixing and transport (Gent & McWilliams, 1990; Hewitt et al., 2020; Zanna & Bolton, 2020). This structural compromise can introduce systematic biases and numerical diffusion, reducing the fidelity of simulated mesoscale impacts and limiting the predictive utility of coarse-grid models in highly dynamic regions (Griffies, 2004; Griffies et al., 2000; Ilıcak et al., 2012; Marchesiello et al., 2009).

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.

Emerging data-driven architectures have sought to alleviate the computational burden of conventional numerical ocean modeling by directly emulating high-resolution ocean states (Reichstein et al., 2019; Kashinath et al., 2021; Pathak et al., 2022). Inspired by the success of AI in atmospheric weather forecasting (notably Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), and FuXi (Chen et al., 2023)), oceanic models such as XiHe (Wang et al., 2024), WenHai (Cui et al., 2025), and GLONET (El Aouni et al., 2025) have demonstrated competitive skill in representing mesoscale structures over short-to-medium forecast horizons. However, directly predicting absolute physical fields at high resolution can introduce additional optimization and dynamical challenges (Chattopadhyay et al., 2024; Stengel et al., 2020). Furthermore, full-depth 3D neural emulators distribute prediction objectives across a broad range of depths, potentially making it more challenging to preserve sharp surface frontal structures that are particularly important for air–sea interactions (El Aouni et al., 2025; Cao et al., 2026).

Recognizing these trade-offs, this study moves away from directly predicting absolute fields as the primary learning target, instead modeling coarse-scale spatiotemporal increments that are reconstructed onto a high-resolution background state. This residual reconstruction strategy separates large-scale dynamical adjustment from the preservation of high-resolution spatial structures (Wang & Karimi, 2022; Martin et al., 2024). We focus on the near-surface upper-ocean state represented by the topmost volumetric model level at $0.49\,\mathrm{m}$. This layer is particularly relevant because the ocean surface directly mediates atmospheric momentum and heat fluxes (Large & Yeager, 2009), while the datasets used in this study provide complementary information for characterizing near-surface ocean variability. Specifically, SST is obtained from OSTIA, ocean state variables are provided by GLORYS12, and atmospheric forcing is derived from ERA5 (Donlon et al., 2012; Hersbach et al., 2020).

Contributions. In this work, we propose a dual-resolution residual reconstruction framework based on a Physics-Enhanced Time-Aware Vision Transformer (PE-TA-ViT). The main contributions are as follows:

- We introduce a dual-resolution residual reconstruction paradigm that predicts coarse-scale spatiotemporal increments at a 1° resolution, without explicitly predicting fine-scale turbulent increments, and reconstructs them on a $1/4^\circ$ high-resolution background state.
- We develop PE-TA-ViT, which integrates physics-informed multi-scale spatial features, variable-aware cross-variable aggregation, and adaptive calendar-time conditioning into the coarse-resolution increment predictor.
- We conduct comprehensive multi-variable evaluations over a 9-day autoregressive forecast horizon, demonstrating improved forecast accuracy and temporal skill while preserving key mesoscale structures and large-scale flow characteristics across dynamically active regions.

2 Data and Preprocessing

We construct a multivariate global ocean dataset from three complementary sources to support the proposed dual-resolution forecasting framework. Atmospheric forcing is represented by daily 10-m zonal and meridional winds from the ECMWF ERA5 reanalysis at 0.25° resolution (Bell et al., 2021). High-resolution sea surface temperature (SST) is sourced from the OSTIA dataset at approximately $1/20^\circ$ resolution (Blockley et al., 2014; Donlon et al., 2012). Oceanic thermodynamic variables (salinity and potential temperature) and kinematic variables (zonal and meridional velocities) are extracted from the CMEMS GLORYS12v1 reanalysis (Lellouche et al., 2021) at a depth of $0.49\mathrm{m}$, corresponding to the center of the uppermost volumetric grid cell in the NEMO configuration (Madec et al., 2022; Rio et al., 2014). The eight input channels jointly represent atmospheric forcing and near-surface ocean states.

All variables are spatially regridded to two common target resolutions: a $1^\circ \times 1^\circ$ coarse grid used as the model input resolution, and a $1/4^\circ \times 1/4^\circ$ high-resolution grid used for residual reconstruction and evaluation. The complete variable configuration and data sources are summarized in Appendix A.1, Table 5.

Data preprocessing consists of two steps: **Land-Sea Masking**, which identifies terrestrial grid points and sets them to 0, and **Min-Max Normalization**, which linearly scales each variable to $[0, 1]$ to place variables with different physical units on a comparable numerical scale.

3 METHODOLOGY

The proposed predictive framework consists of two coupled stages: generating the large-scale dynamic increment via the **Physics-Enhanced Time-Aware Vision Transformer (PE-TA-ViT)** at 1° resolution, followed by a **Dual-Resolution Residual Reconstruction** mechanism to project these dynamics onto the $1/4^\circ$ scale (Figure 1).

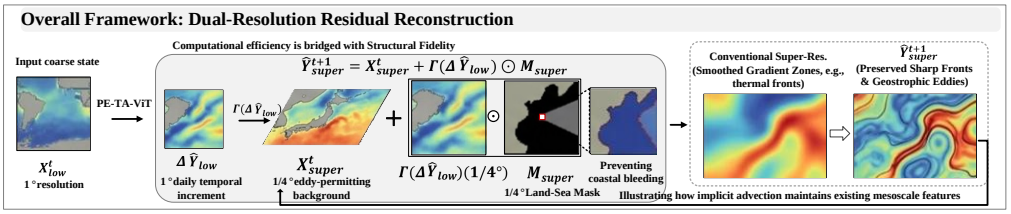

Figure 1: **Overall architecture of the proposed Dual-Resolution Residual Reconstruction framework.** The low-resolution macro-scale temporal increment predicted by the PE-TA-ViT is upsampled, superimposed onto the detail-rich eddy-permitting background, and masked to produce the final state.

3.1 DUAL-RESOLUTION RESIDUAL RECONSTRUCTION

Directly predicting absolute ocean states at coarse resolution and subsequently reconstructing them at higher resolution can attenuate sharp gradients and fine-scale structures during spatial reconstruction (Martin et al., 2024; Wang & Karimi, 2022; Martin et al., 2023). To bridge computational efficiency and high-resolution structural fidelity, we formulate forecasting as residual prediction at coarse resolution followed by high-resolution residual reconstruction.

Given the coarse-resolution (1°) input state $\mathbf{X}^t_{low} \in \mathbb{R}^{C\times H_{low}\times W_{low}}$ and the corresponding high-resolution ($1/4^\circ$) background state $\mathbf{X}^t_{super} \in \mathbb{R}^{C\times H_{super}\times W_{super}}$, the one-step forecast from t to $t+1$ is formulated as:

$$\Delta\hat{\mathbf{Y}}_{low} = f_\theta(\mathbf{X}^t_{low}), \tag{1}$$

$$\hat{\mathbf{Y}}^{t+1}_{super} = \mathbf{X}^t_{super} + \mathcal{I}(\Delta\hat{\mathbf{Y}}_{low}) \odot \mathbf{M}_{super}, \tag{2}$$

where f_θ denotes the PE-TA-ViT engine that predicts the daily temporal increment, $\mathcal{I}$ denotes bilinear interpolation from 1° to $1/4^\circ$, and $\mathbf{M}_{super}$ is the high-resolution binary land-sea mask.

This formulation separates large-scale temporal evolution from fine-scale spatial representation. Specifically, PE-TA-ViT predicts the temporal increment from the coarse-resolution state, while the high-resolution background provides the fine-scale spatial template. During multi-step forecasting, PE-TA-ViT is deployed autoregressively without using future ground-truth states. The initial ground-truth states $\mathbf{X}^0_{low}$ and $\mathbf{X}^0_{super}$ are provided only at $t = 0$. For each subsequent step ($t > 0$), the coarse-resolution state is updated by the predicted temporal increment, while the previous high-resolution prediction is used as the new high-resolution background:

$$\mathbf{X}^t_{low} = \mathbf{X}^{t-1}_{low} + \Delta\hat{\mathbf{Y}}^{t-1}_{low}, \qquad \mathbf{X}^t_{super} = \hat{\mathbf{Y}}^t_{super}, \qquad t > 0. \tag{3}$$

3.2 PHYSICS-ENHANCED ARCHITECTURES INSIDE f_θ

To introduce physics-informed inductive biases and temporal conditioning into the 1° increment predictor $\Delta\hat{\mathbf{Y}}_{low}$, the PE-TA-ViT backbone incorporates several physics-guided and time-aware components, as illustrated in Figure 2.

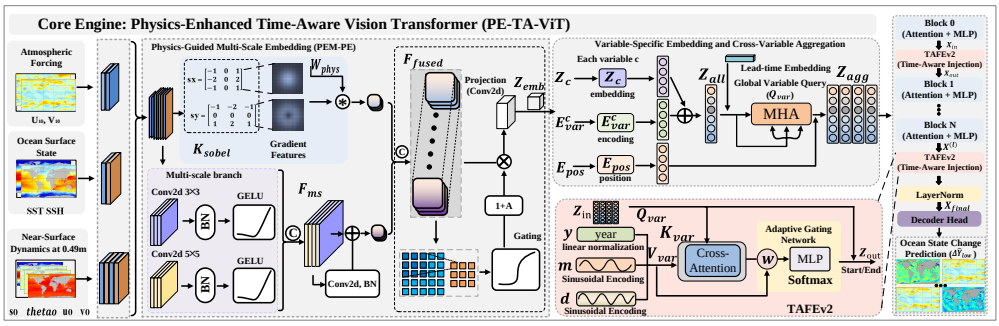

Figure 2: **Architecture of the Physics-Enhanced Time-Aware Vision Transformer (PE-TA-ViT).** (1) PEM-PE module; (2) Cross-Variable Aggregation; and (3) TAFEv2 module.

3.2.1 Physics-Guided Multi-Scale Embedding (PEM-PE)

PEM-PE adopts a dual-branch architecture with multi-scale convolutions and a Sobel-initialized gradient branch, followed by residual spatial gating. First, the Multi-Scale Branch extracts local patterns at distinct receptive fields using parallel 3×3 and 5×5 convolutions. Their outputs are concatenated and refined by a residual convolution to obtain $\mathbf{F}_{ms}$. In parallel, the Physics Branch provides a first-order spatial-gradient prior, yielding the physics feature $\mathbf{F}_{phys}$. This branch is initialized with Sobel operators $\mathbf{K}_{sobel}$. To maintain flexibility, the kernels are adapted during training through a learnable perturbation $\Delta\mathbf{W}$, while a learnable scalar α_{phys} controls the overall contribution of this gradient prior:

$$\mathbf{W}_{phys} = \alpha_{phys} \left(\mathbf{K}_{sobel} + \Delta\mathbf{W}\right), \qquad \Delta\mathbf{W}_{init} = 0. \tag{4}$$

The feature representations from both branches are then concatenated:

$$\mathbf{F}_{fused} = [\mathbf{F}_{ms}; \mathbf{F}_{phys}]. \tag{5}$$

Finally, a residual gating mechanism generates a spatial saliency map $\mathbf{A} \in [0, 1]$ via a Sigmoid activation to enhance the fused representation:

$$\mathbf{Z}_{emb} = \mathrm{Proj}\left(\mathbf{F}_{fused} \odot (1 + \mathbf{A})\right). \tag{6}$$

3.2.2 Variable-Specific Embedding and Cross-Variable Aggregation

The input variables have distinct physical meanings and statistical characteristics. To preserve variable-specific representations before cross-variable interaction, each variable c is processed independently using a dedicated patch-embedding module. A learnable variable embedding $\mathbf{E}^c_{var}$ is then added to its latent representation, yielding variable-aware tokens that are concatenated as

$$\mathbf{Z}_{all} = [\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_C]. \tag{7}$$

The variable-specific tokens are aggregated independently at each spatial location using a shared learnable query embedding $\mathbf{Q}_{var}$:

$$\mathbf{Z}_{agg} = \mathrm{MHA}\left(\mathbf{Q}_{var}, \mathbf{Z}_{all}, \mathbf{Z}_{all}\right). \tag{8}$$

Positional and forecast-lead embeddings are then incorporated into the aggregated representation:

$$\mathbf{X}_{in} = \mathbf{Z}_{agg} + \mathbf{E}_{pos} + \mathbf{E}_{lead}. \tag{9}$$

Here, $\mathbf{E}_{pos}$ encodes spatial token positions, while $\mathbf{E}_{lead}$ provides continuous forecast-horizon information.

3.2.3 Time-Aware Feature Enhancement (TAFEv2)

Month and day are encoded using sinusoidal functions, while the year is continuously normalized; the resulting temporal features are projected into the latent dimension for TAFEv2 conditioning. TAFEv2

employs three independent cross-attention streams. Given the spatial token representation $\mathbf{X}_{in}$ and the temporal embeddings $\mathbf{e}_y$, $\mathbf{e}_m$, and $\mathbf{e}_d$, spatial tokens serve as queries, while each corresponding temporal embedding is used as both key and value:

$$\mathbf{X}_y = \mathrm{MHA}_y(\mathbf{X}_{in}, \mathbf{e}_y, \mathbf{e}_y), \quad \mathbf{X}_m = \mathrm{MHA}_m(\mathbf{X}_{in}, \mathbf{e}_m, \mathbf{e}_m), \quad \mathbf{X}_d = \mathrm{MHA}_d(\mathbf{X}_{in}, \mathbf{e}_d, \mathbf{e}_d). \tag{10}$$

An adaptive gating network predicts normalized weights from the three temporal embeddings:

$$\mathbf{w} = \mathrm{Softmax}\left(\mathrm{MLP}\left([\mathbf{e}_y, \mathbf{e}_m, \mathbf{e}_d]\right)\right), \tag{11}$$

which are used to form a weighted combination of the three temporal features:

$$\mathbf{X}_{enhanced} = w_y \mathbf{X}_y + w_m \mathbf{X}_m + w_d \mathbf{X}_d. \tag{12}$$

The resulting temporal features are injected through a residual projection:

$$\mathbf{X}_{out} = \mathbf{X}_{in} + \mathrm{Proj}\left(\mathbf{X}_{enhanced}\right). \tag{13}$$

TAFEv2 is applied after the first and final Transformer blocks, providing temporal conditioning at both early and late stages of feature processing.

3.2.4 Deep Spatial Modeling and Increment Reconstruction

The variable-aggregated feature sequence is processed by a stack of L standard Vision Transformer blocks ($L = 24$ in our main configuration). Each Transformer block follows the standard ViT formulation, combining layer-normalized multi-head self-attention (MHSA) and a feed-forward multi-layer perceptron (MLP) within residual connections to model long-range spatial dependencies across the global ocean domain:

$$\mathbf{X}^{(l)} = \mathrm{Block}_l\left(\mathbf{X}^{(l-1)}\right), \qquad l = 1, \ldots, L. \tag{14}$$

After the final TAFEv2 module, Layer Normalization produces the refined latent representation

$$\mathbf{X}_{final} \in \mathbb{R}^{B \times N \times D}, \tag{15}$$

where B denotes the batch size, N is the number of spatial patches, and D is the embedding dimension.

The refined token representation is then mapped back to the physical prediction space through a multi-layer MLP prediction head with GELU activations. The prediction head transforms each latent token into a flattened physical patch containing C_{out} predicted ocean-state variables:

$$\hat{\mathbf{X}}_{patch} = \mathrm{MLP}_{head}(\mathbf{X}_{final}) \in \mathbb{R}^{B \times N \times (C_{out} \cdot P^2)}, \tag{16}$$

where C_{out} denotes the number of predicted ocean-state variables and P is the patch size.

Finally, a deterministic unpatchify operation rearranges the flattened patch predictions into their original two-dimensional spatial arrangement, producing the coarse-resolution forecast increment field:

$$\Delta\hat{\mathbf{Y}}_{low} = \mathrm{Unpatchify}(\hat{\mathbf{X}}_{patch}) \in \mathbb{R}^{B \times C_{out} \times H \times W}. \tag{17}$$

3.3 Loss Function

We propose a Dual-Weighted Root Mean Squared Error (DW-RMSE) loss to account for latitude-dependent grid-cell area and magnitude disparities across variables using two normalized weighting factors: a latitude-dependent cosine weight β_i and an inverse-variance weight α_v:

$$\beta_i = \frac{\cos(\theta_i)}{\frac{1}{H}\sum_{k=1}^{H}\cos(\theta_k)}, \qquad \alpha_v = \frac{\sigma_v^{-2}}{\frac{1}{V}\sum_{k=1}^{V}\sigma_k^{-2}}. \tag{18}$$

Here, θ_i is the latitude in radians at the i-th row, H is the total number of latitudinal grids, σ_v is the standard deviation of historical daily variations for the v-th variable, and V is the total number of target variables.

The final DW-RMSE loss $\mathcal{L}$ computes a spatially weighted RMSE over valid grid points Ω_v for each variable and averages the resulting RMSE values across variables:

$$\mathcal{L} = \frac{1}{V}\sum_{v=1}^{V}\sqrt{\frac{1}{N_v}\sum_{(i,j)\in\Omega_v}\alpha_v\beta_i\left(Y_{v,i,j} - \hat{Y}_{v,i,j}\right)^2 + \epsilon}. \tag{19}$$

4 EXPERIMENTS

To comprehensively evaluate the forecast skill, we utilize the Root Mean Square Error (RMSE) to measure deterministic point-wise accuracy, and the Anomaly Correlation Coefficient (ACC) to assess spatiotemporal phase alignment. The rigorous mathematical formulations and variable definitions for these metrics are detailed in **Appendix A.3**.

4.1 HIGH-RESOLUTION DETERMINISTIC ERROR EVALUATION (RMSE)

Table 1: Quantitative comparison of deterministic forecast skill between PE-TA-ViT and the Persistence baseline. The table captures the RMSE evolution at specific lead times and the overall 9-day statistics.

Variable	Model	RMSE Evolution (Lead Time)				Overall Statistics	
		Day 1	Day 4	Day 7	Day 9	Avg RMSE	Std RMSE
SST (°C)	**PE-TA-ViT**	**0.215**	**0.451**	**0.557**	**0.610**	**0.460**	**0.137**
	Persistence	0.238	0.524	0.674	0.754	0.547	0.239
thetao (°C)	**PE-TA-ViT**	**0.161**	**0.412**	**0.542**	**0.604**	**0.431**	**0.160**
	Persistence	0.183	0.454	0.609	0.690	0.482	0.178
SSH (m)	**PE-TA-ViT**	**0.015**	**0.038**	**0.051**	**0.058**	**0.040**	**0.014**
	Persistence	0.021	0.044	0.057	0.064	0.046	0.014
so (PSU)	**PE-TA-ViT**	**0.094**	**0.224**	**0.293**	**0.325**	**0.235**	**0.109**
	Persistence	0.103	0.239	0.310	0.346	0.250	0.132
uo (m/s)	**PE-TA-ViT**	**0.046**	**0.103**	**0.133**	**0.148**	**0.096**	**0.048**
	Persistence	0.065	0.120	0.149	0.163	0.124	0.051
vo (m/s)	**PE-TA-ViT**	**0.047**	**0.110**	**0.140**	**0.156**	**0.100**	**0.052**
	Persistence	0.069	0.127	0.160	0.175	0.133	0.055

To evaluate deterministic forecast skill and error stability, we assess PE-TA-ViT across six prognostic ocean variables over a 9-day forecast horizon using spatially averaged RMSE and its standard deviation. Representative lead-time RMSE values and overall 9-day statistics are summarized in Table 1, while complete variable-wise RMSE trajectories are provided in Appendix E (Figures 10–12).

As shown in Table 1, PE-TA-ViT consistently outperforms persistence across the entire forecast horizon for all target variables. For the thermodynamic fields, the average RMSE is reduced by 15.9% for SST and 10.6% for thetao, yielding 9-day average RMSEs of 0.460°C and 0.431°C, respectively. The corresponding RMSE standard deviations are 0.137°C and 0.160°C, both notably lower than those of persistence, indicating suppressed error variability. For dynamic variables, PE-TA-ViT reduces the average RMSE by 13.0% for SSH, 22.6% for zonal velocity (u_o), and 24.8% for meridional velocity (v_o), maintaining a distinct advantage from Day 1 through Day 9. Salinity also improves, with a 6.0% reduction in average RMSE relative to the baseline. Overall, PE-TA-ViT demonstrates robust and enhanced deterministic forecast skill across the entire suite of evaluated variables. Detailed annual error evolution and 3D spatiotemporal error analyses are provided in Appendices B and C, respectively.

4.2 HIGH-RESOLUTION SPATIAL PATTERN EVALUATION (ACC)

The Anomaly Correlation Coefficient (ACC) evaluates the spatiotemporal phase alignment of climatic anomalies. In this section, we comprehensively assess the ACC performance of the proposed PE-TA-ViT framework against the persistence baseline over a 9-day rollout, utilizing the conventional operational threshold of 0.6 as a benchmark for valid forecasting skill. Representative lead-time ACC values and overall 9-day statistics are summarized in Table 2. (Note: The continuous visual evolution curves corresponding to these metrics are available in Appendix E, Figures 13-15).

As shown in Table 2, PE-TA-ViT achieves higher ACC than persistence at all reported lead times across all six evaluated ocean variables, indicating improved spatiotemporal agreement with the target fields. For SST, thetao, SSH, and salinity, the model achieves mean ACCs of 0.839, 0.863,

Table 2: Quantitative comparison of ACC between PE-TA-ViT and persistence. The table reports ACC at selected lead times and overall 9-day statistics; higher values indicate better alignment, with best results in bold.

Variable	Model	ACC Evolution (Lead Time)				Overall Statistics	
		Day 1	Day 4	Day 7	Day 9	Mean ACC	Std ACC (↓)
SST	**PE-TA-ViT**	**0.967**	**0.857**	**0.783**	**0.736**	**0.839**	**0.080**
	Persistence	0.958	0.817	0.719	0.665	0.794	0.146
thetao	**PE-TA-ViT**	**0.982**	**0.888**	**0.802**	**0.753**	**0.863**	**0.088**
	Persistence	0.975	0.858	0.756	0.698	0.829	0.109
SSH	**PE-TA-ViT**	**0.989**	**0.931**	**0.869**	**0.831**	**0.911**	**0.059**
	Persistence	0.980	0.911	0.845	0.803	0.890	0.062
so	**PE-TA-ViT**	**0.969**	**0.848**	**0.753**	**0.705**	**0.824**	**0.114**
	Persistence	0.963	0.834	0.740	0.688	0.810	0.127
uo	**PE-TA-ViT**	**0.939**	**0.743**	**0.585**	**0.531**	**0.707**	**0.155**
	Persistence	0.896	0.687	0.544	0.471	0.653	0.155
vo	**PE-TA-ViT**	**0.927**	**0.693**	**0.532**	**0.458**	**0.656**	**0.174**
	Persistence	0.869	0.608	0.429	0.348	0.567	0.183

0.911, and 0.824, respectively, compared with 0.794, 0.829, 0.890, and 0.810 for persistence. Their corresponding ACC standard deviations are also lower than those of the baseline, indicating more stable ACC over the reported forecast period. The improvement is particularly evident for the velocity fields. Mean ACC increases from 0.653 to 0.707 for zonal velocity (u_o) and from 0.567 to 0.656 for meridional velocity (v_o). Although ACC decreases with lead time for both models, PE-TA-ViT maintains higher ACC at every reported lead time, with Day-9 values of 0.531 and 0.458 for u_o and v_o, respectively, compared with 0.471 and 0.348 for persistence. Using ACC=0.6 as a reference threshold, the table shows that the velocity forecasts approach or cross this threshold between Days 4 and 7; the exact threshold-crossing times are evaluated from the complete ACC trajectories in Appendix E.

4.3 Physical and Dynamical Diagnostics

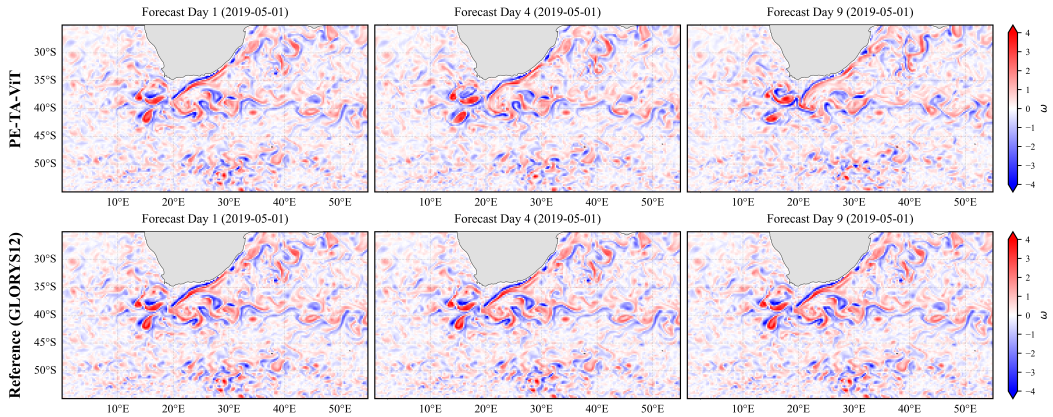

Figure 3: **Surface relative vorticity in the Agulhas Retroflection region (2019-05-01).** Comparison between PE-TA-ViT (top) and GLORYS12 (bottom) across Day 1, 4, and 9.

To evaluate dynamic fidelity, we analyze surface relative vorticity ($\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$) in the Agulhas Retroflection (Figure 3), a region dominated by strong meandering and complex filaments. PE-TA-ViT retains these energetic, fine-scale structures even at Day 9 without pronounced numerical dissipation. The vorticity amplitude ($\omega \in [-4, 4]$) and spatial distribution remain highly comparable to the GLORYS12 reference, indicating preservation of coherent small-scale dynamical structures.

Table 3: **Ablation study of PE-TA-ViT.** Residual versus direct interpolation and the effects of PEM-PE, VarAgg, TAFEv2, and DW-RMSE are evaluated by mean RMSE across eight evaluated variables (lower is better); best results are in bold. † denotes offshore-only evaluation (> 200 km).

Model	Components				Forcing		Surface		Near-Surface			
	PEM-PE	VarAgg	TAFEv2	DW	$\mathbf{u_{10}}$	$\mathbf{v_{10}}$	SST	SSH	so	thetao	uo	vo
Base	×	×	×	×	4.816	5.070	1.677	0.050	0.328	0.509	0.126	0.134
+ PEM-PE	✓	×	×	×	4.816	5.068	1.534	0.049	0.321	0.512	0.125	0.134
+ VarAgg	✓	✓	×	×	4.809	5.053	1.328	0.049	0.332	0.508	0.123	0.133
+ TAFEv2	✓	✓	✓	×	3.404	3.379	0.550	0.046	0.253	0.455	0.116	0.128
Direct	✓	✓	✓	✓	3.659	3.700	33.792	0.222	3.991	2.184	0.292	0.342
Direct†	✓	✓	✓	✓	3.659	3.700	1.274	0.073	1.033	0.728	0.168	0.187
PE-TA-ViT	✓	✓	✓	✓	**3.394**	**3.432**	**0.460**	**0.040**	**0.235**	**0.431**	**0.096**	**0.100**

This structural fidelity is further supported by spectral diagnostics: the maximum divergence in the energy cascading fraction (E_η) remains below 0.009 globally and 0.004 in the dynamically active Agulhas region at Day 9.

As an additional process-oriented diagnostic, geostrophic currents derived from forecasted SSH also exhibit strong spatial agreement with the GLORYS12 reference, with no pronounced gradient-amplified artifacts at the evaluated lead times. Detailed geostrophic-current derivations, quantitative error evolution, and additional regional analyses are provided in Appendix D, together with the extended vorticity and energy-cascading diagnostics.

4.4 Performance, Ablation, and Efficiency Evaluation

Table 3 presents a cumulative ablation of the architectural components and training objective. Adding PEM-PE primarily reduces the SST RMSE from 1.677 to 1.534, while producing only minor changes in the remaining variables; the RMSE of $thetao$ increases slightly from 0.509 to 0.512. The subsequent introduction of VarAgg noticeably further reduces the SST RMSE to 1.328 and yields small reductions for the wind and current variables, whereas the salinity RMSE increases from 0.321 to 0.332, indicating a variable-dependent trade-off. TAFEv2 provides the dominant incremental improvement in the cumulative ablation, reducing the RMSE for all eight variables, with particularly pronounced improvements for u_{10}, v_{10}, and SST. Relative to the TAFEv2 configuration, introducing DW-RMSE further reduces the RMSE for seven of the eight variables, although the v_{10} RMSE increases slightly from 3.379 to 3.432.

The comparison of reconstruction strategies shows that interpolating absolute fields produces substantially larger full-domain errors than residual reconstruction. In particular, direct interpolation results in RMSE values of 33.792 °C for SST, 0.222 m for SSH, and 3.991 PSU for salinity. When the same direct-interpolation outputs are evaluated only at ocean grid points located more than 200 km from the coastline, the corresponding RMSE values decrease to 1.274 °C, 0.073 m, and 1.033 PSU, respectively. These reductions indicate that a substantial fraction of the full-domain direct-interpolation error is associated with near-coastal regions. Under the full-domain evaluation, residual reconstruction achieves lower RMSE than direct interpolation for all eight variables. The results therefore support the use of residual-field interpolation to reduce interpolation-related errors in the reconstruction of high-resolution ocean states.

Table 4 presents a quantitative comparison of PE-TA-ViT with a persistence baseline, representative deep learning forecasting models (U-Net, Swin-Unet, ResNet, ConvLSTM, and MUDDFormer), and GLONET using mean RMSE over a 9-day forecasting horizon across six ocean state variables. PE-TA-ViT achieves the lowest RMSE for all evaluated variables, with values of 0.446 °C, 0.362 °C, 0.031 m, 0.125 PSU, 0.097 m/s, and 0.098 m/s for SST, thetao, SSH, so, uo, and vo, respectively. Compared with the persistence baseline, PE-TA-ViT consistently reduces the RMSE across all six variables, indicating improved forecasting accuracy relative to the state-preserving baseline.

Among the deep learning baselines, Swin-Unet, ConvLSTM, and MUDDFormer generally achieve lower RMSE than U-Net across the evaluated variables, while their relative performance varies across different ocean state variables. For SSH prediction, ConvLSTM, Swin-Unet, MUDDFormer, and

Table 4: Quantitative comparison of mean RMSE over a 9-day forecasting horizon. Lower RMSE indicates better prediction accuracy, and the lowest errors are highlighted in bold.

Model	SST (°C)	thetao (°C)	SSH (m)	so (PSU)	uo (m/s)	vo (m/s)
Persistence baseline						
Persistence	0.515	0.454	0.046	0.234	0.111	0.115
Deep learning forecasting baselines						
U-Net (Ronneberger et al., 2015)	0.863	0.702	0.056	0.380	0.142	0.145
Swin-Unet (Cao et al., 2021)	0.554	0.390	0.042	0.213	0.110	0.106
ResNet (He et al., 2016)	0.588	0.483	0.045	0.269	0.107	0.109
ConvLSTM (Shi et al., 2015)	0.550	0.395	0.041	0.215	0.108	0.104
MUDDFormer (Xiao et al., 2025)	0.525	0.392	0.043	0.217	0.114	0.106
Data-driven ocean forecasting model						
GLONET (El Aouni et al., 2025)	0.716	0.741	0.083	0.740	0.130	0.125
Ours: PE-TA-ViT	**0.446**	**0.362**	**0.031**	**0.125**	**0.097**	**0.098**

Note: For the cross-model comparison in Table 4, all models are evaluated over January–December 2023, using GLORYS12 and OSTIA as reference datasets, with forecasts initialized every Wednesday.

ResNet obtain RMSE values of 0.041 m, 0.042 m, 0.043 m, and 0.045 m, respectively, whereas PE-TA-ViT achieves a lower RMSE of 0.031 m. For the near-surface velocity variables, PE-TA-ViT obtains RMSE values of 0.097 m/s and 0.098 m/s for uo and vo, respectively, which are lower than those of the compared deep learning baselines. PE-TA-ViT also achieves lower RMSE than GLONET across all six evaluated variables, with particularly notable differences in SSH and salinity prediction.

Beyond physical and structural fidelity, a primary bottleneck for global eddy-permitting models is the prohibitive computational overhead. Evaluated on a single consumer-grade NVIDIA RTX 3060 GPU, PE-TA-ViT completes a 9-day global forecast cycle in merely 1.85 seconds (averaging $\sim$205 ms per day) with a peak VRAM utilization of only 0.64 GB. This ultra-lightweight footprint ensures exceptional scientific accessibility, with comprehensive hardware configurations provided in **Appendix A.2**.

5 Conclusion, Limitations and Future Work

This study introduces a dual-resolution residual reconstruction deep learning framework (PE-TA-ViT) for high-resolution forecasting of sea surface and near-surface ocean states. By constraining the predictive task to large-scale temporal increments at 1° resolution and projecting them onto a $1/4^\circ$ eddy-permitting background, the framework combines coarse-scale dynamical prediction with high-resolution structural reconstruction at low inference cost. Evaluated on a single NVIDIA RTX 3060 GPU, PE-TA-ViT completes a full 9-day global rollout in 1.85 seconds with a peak VRAM footprint of 0.64 GB. Comprehensive evaluations show that PE-TA-ViT achieves the lowest mean RMSE among the evaluated forecasting baselines across all six ocean state variables, reducing average RMSE relative to persistence by 6.0%–24.8%.

Ablation results show that PEM-PE and VarAgg provide variable-dependent improvements, while TAFEv2 yields the largest incremental gain across the evaluated variables. The residual reconstruction strategy also substantially reduces the reconstruction errors observed with direct interpolation, particularly in coastal regions. In terms of spatiotemporal phase alignment, PE-TA-ViT extends the useful forecast window for dynamic velocity fields by approximately one day at the ACC=0.6 threshold and maintains higher ACC at all reported lead times, including an ACC of 0.705 for salinity at Day 9. Physical diagnostics further show coherent transient mesoscale vorticity structures and close agreement of the derived geostrophic currents with the GLORYS12 reference, while the maximum global deviation in the energy cascading fraction (E_η) remains below 0.009.

Limitations and Future Work. The model's macro-scale dynamic evolution is currently constrained by its reliance on a 1° coarse-resolution engine. Consequently, it exhibits inherent physical limitations in explicitly triggering spontaneous subgrid mesoscale instabilities, such as new eddy genesis. To address this, future research will integrate generative models into the residual upsampling phase to stochastically parameterize subgrid turbulence.

REFERENCES

B Bell, H Hersbach, A Simmons, P Berrisford, P Dahlgren, A Horányi, J Muñoz-Sabater, J Nicolas, R Radu, D Schepers, et al. The era5 global reanalysis: Preliminary extension to 1950. *Quarterly Journal of the Royal Meteorological Society*, 147(741):4186–4227, 2021. doi: 10.1002/qj.4174.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

E. W. Blockley, M. J. Martin, A. J. McLaren, A. G. Ryan, J. Waters, D. J. Lea, I. Mirouze, K. A. Peterson, A. Sellar, and D. Storkey. Recent development of the met office operational ocean forecasting system: an overview and assessment of the new global foam forecasts. *Geoscientific Model Development*, 7(6):2613–2638, 2014. doi: 10.5194/gmd-7-2613-2014. URL https://gmd.copernicus.org/articles/7/2613/2014/.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Linfei Cao, Jiachen Yang, Meng Xi, Jingyi He, Fei Gao, Jiabao Wen, and Zhijing Wang. M^2f-pinn: A multi-scale frequency-domain multi-physics-informed neural network for ocean forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4z7tMSzQET.

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.

Ashesh Chattopadhyay, Michael Gray, Tianning Wu, Anna B Lowe, and Ruoying He. Oceannet: a principled neural operator-based digital twin for regional oceans. *Scientific Reports*, 14(1):21181, 2024.

Dudley B Chelton, Michael G Schlax, and Roger M Samelson. Global observations of nonlinear mesoscale eddies. *Progress in oceanography*, 91(2):167–216, 2011.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

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.

Changming Dong, James C McWilliams, Yu Liu, and Dake Chen. Global heat and salt transports by eddy movement. *Nature communications*, 5(1):3294, 2014.

Craig J. Donlon, Matthew Martin, John Stark, Jonah Roberts-Jones, Emma Fiedler, and Werenfrid Wimmer. The operational sea surface temperature and sea ice analysis (ostia) system. *Remote Sensing of Environment*, 116:140–158, 2012. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2010.10.017. URL https://www.sciencedirect.com/science/article/pii/S0034425711002197. Advanced Along Track Scanning Radiometer(AATSR) Special Issue.

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, 2025.

Baylor Fox-Kemper, Alistair Adcroft, Claus W Böning, Eric P Chassignet, Enrique Curchitser, Gokhan Danabasoglu, Carsten Eden, Matthew H England, Rüdiger Gerdes, Richard J Greatbatch, et al. Challenges and prospects in ocean circulation models. *Frontiers in Marine Science*, 6:65, 2019.

Peter R Gent and James C McWilliams. Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, 20(1):150–155, 1990.

Stephen M. Griffies. *Fundamentals of Ocean Climate Models*. Princeton University Press, 2004. doi: 10.2307/j.ctv301gzg.

Stephen M Griffies, Ronald C Pacanowski, and Robert W Hallberg. Spurious diapycnal mixing associated with advection in a z-coordinate ocean model. *Monthly Weather Review*, 128(3): 538–564, 2000. doi: 10.1175/1520-0493(2000)128<0538:SDMAWA>2.0.CO;2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Helene T Hewitt, Malcolm Roberts, Pierre Mathiot, Arne Biastoch, Ed Blockley, Eric P Chassignet, Baylor Fox-Kemper, Pat Hyder, David P Marshall, Ekaterina Popova, et al. Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, 6 (4):137–152, 2020.

Mehmet Ilıcak, Alistair J Adcroft, Stephen M Griffies, and Robert W Hallberg. Spurious dianeutral mixing and the role of momentum closure. *Ocean Modelling*, 45:37–58, 2012.

Kartik Kashinath, M Mustafa, A Albert, JL Wu, C Jiang, S Esmaeilzadeh, K Azizzadenesheli, R Wang, A Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194): 20200093, 2021.

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.

William G Large and Stephen G Yeager. The global climatology of an interannually varying air–sea flux data set. *Climate Dynamics*, 33(2-3):341–364, 2009.

William G Large, James C McWilliams, and Scott C Doney. Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Reviews of Geophysics*, 32(4):363–403, 1994. doi: 10.1029/94RG01872.

Jean-Michel Lellouche, Eric Greiner, Romain Bourdallé-Badie, et al. The copernicus global 1/12° oceanic and sea ice glorys12 reanalysis. *Frontiers in Earth Science*, 9:698876, 2021. ISSN 2296-6463. doi: 10.3389/feart.2021.698876.

Xiaofeng Li, Xudong Zhang, Yibin Ren, Yunhe Wang, Nan Yang, Chong Wang, Shanshan Mu, Haoyu Wang, Yuan Zhou, Yingjie Liu, et al. Artificial intelligence for ocean phenomena forecasting. *Science Bulletin*, 2026.

Gurvan Madec, Romain Bourdallé-Badie, Jérôme Chanut, Emanuela Clementi, Andrew Coward, et al. Nemo ocean engine, 2022.

Patrick Marchesiello, Laurent Debreu, and Xavier Couvelard. Spurious diapycnal mixing in terrain-following coordinate models: The problem and a solution. *Ocean Modelling*, 26(3-4):156–169, 2009.

Scott A Martin, Georgy E Manucharyan, and Patrice Klein. Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. *Journal of Advances in Modeling Earth Systems*, 15(5): e2022MS003589, 2023.

Scott A. Martin, Georgy E. Manucharyan, and Patrice Klein. Deep learning improves global satellite observations of ocean eddy dynamics. *Geophysical Research Letters*, 51(17): e2024GL110059, 2024. doi: https://doi.org/10.1029/2024GL110059. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024GL110059. e2024GL110059 2024GL110059.

James C McWilliams. Submesoscale currents in the ocean. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2189):20160117, 2016.

David P. Mulholland, Patrick Laloyaux, Keith Haines, and Magdalena Alonso Balmaseda. Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Monthly Weather Review*, 143(11):4631 – 4644, 2015. doi: 10.1175/MWR-D-15-0076.1. URL https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0076.1.xml.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

M-H Rio, S Mulet, and N Picot. Beyond goce for the ocean circulation estimate: Synergetic use of altimetry, gravimetry, and in situ data provides new insight into geostrophic and ekman currents. *Geophysical Research Letters*, 41(24):8918–8925, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Kai Stengel, Andrew Glaws, Dylan Hettinger, and Ryan N King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29): 16805–16815, 2020.

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.

Yang Wang and Hassan A. Karimi. Perceptual loss function for generating high-resolution climate data. *Applied Computing and Intelligence*, 2(2):152–172, 2022. ISSN 2771-392X. doi: 10.3934/aci.2022009. URL https://www.aimspress.com/article/doi/10.3934/aci.2022009.

Phillip Williamson and Valeria A Guinder. Effect of climate change on marine ecosystems. *The impacts of climate change*, pp. 115–176, 2021.

Carl Wunsch and Raffaele Ferrari. Vertical mixing, energy, and the general circulation of the oceans. *Annual Review of Fluid Mechanics*, 36:281–314, 2004.

Da Xiao, Qingye Meng, Shengping Li, and Xingyuan Yuan. Muddformer: Breaking residual bottlenecks in transformers via multiway dynamic dense connections. *arXiv preprint arXiv:2502.12170*, 2025.

Laure Zanna and Thomas Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020.

Zhengguang Zhang, Wei Wang, and Bo Qiu. Oceanic mass transport by mesoscale eddies. *Science*, 345(6194):322–324, 2014.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and grammatical editing to improve the readability of the manuscript. We have not used generative AI tools for formulating research ideas, writing experimental code, data analysis, or generating scientific claims, and the rest of the required disclosure tasks are not applicable to this work. We have thoroughly reviewed and manually revised all AI-assisted text to ensure scientific accuracy and that the original meaning is strictly preserved. We take full responsibility for the final content of this work, including any text refined with the aid of generative AI.

A EXPERIMENTAL SETUP AND REPRODUCIBILITY DETAILS

A.1 DETAILED VARIABLE CONFIGURATIONS

Table 5 outlines the specific 8-channel input variables utilized in the PE-TA-ViT framework, representing a coupled air-sea boundary layer loop.

Table 5: Configuration of input variables for the data-driven model. The selected variables form a coupled air-sea boundary layer loop.

Ch.	Sym.	Standard Name	Level	Physical Significance	Source
Atmospheric Forcing					
0	u_{10}	10-m Zonal Wind	10 m	Zonal wind forcing	ERA5
1	v_{10}	10-m Merid. Wind	10 m	Meridional wind forcing	ERA5
Oceanic Surface State					
2	SST	Sea Surface Temp.	Surf.	Thermal boundary condition	OSTIA
3	SSH	Sea Surface Height	Surf.	Barotropic pressure anomaly	GLORYS12
Near-Surface Dynamics					
4	so	Salinity	0.49 m	Haline buoyancy component	GLORYS12
5	$thetao$	Potential Temp.	0.49 m	Thermal buoyancy component	GLORYS12
6	uo	Zonal Velocity	0.49 m	Zonal advection	GLORYS12
7	vo	Merid. Velocity	0.49 m	Meridional advection	GLORYS12

A.2 COMPUTATIONAL EFFICIENCY AND IMPLEMENTATION DETAILS

A.2.1 COMPUTATIONAL EFFICIENCY

We evaluate the computational efficiency of PE-TA-ViT in terms of model size, inference latency, and GPU memory consumption. The proposed model contains 70.092 million trainable parameters under the default configuration.

For training, experiments were conducted using four NVIDIA V100 GPUs (32GB). During inference benchmarking, the model was evaluated on a single NVIDIA RTX 3060 GPU with 12GB memory using the original input resolution of 170×360. The inference latency was measured with a batch size of one and only included the model forward propagation.

For a complete 9-day forecasting cycle, PE-TA-ViT requires 1848.61 ms inference time, corresponding to approximately 205 ms per forecast step. The peak GPU memory consumption during inference is 0.64 GB. These results demonstrate the computational efficiency of PE-TA-ViT under the evaluated forecasting setting.

A.2.2 IMPLEMENTATION DETAILS

To facilitate reproducibility, we provide detailed descriptions of the dataset partitioning, preprocessing procedures, model configurations, optimization strategies, and forecasting protocols used in our experiments.

A.2.3 Dataset Partitioning and Normalization

The dataset was partitioned chronologically to simulate a realistic out-of-sample forecasting scenario. Specifically, data from 1993 to 2017 were used for model training, 2018 was used for validation, and 2019 was reserved as the independent test period. The validation set was used for model selection and hyperparameter monitoring, while the test set was only used for final performance evaluation. This temporal partition prevents future information leakage and ensures that the reported results reflect generalization to unseen temporal periods.

All normalization statistics were computed exclusively from the training period (1993–2017) and subsequently applied to the validation and testing sets. Therefore, no information from future periods was involved during preprocessing.

A.2.4 Model Configuration

The proposed PE-TA-ViT adopts a Vision Transformer-based architecture for multivariate ocean state prediction. The detailed architectural configurations, including input specifications, Transformer block depths, attention dimensions, and total trainable parameters, are summarized in Table 6.

Table 6: Detailed architectural configurations of the proposed PE-TA-ViT.

Configuration	**Value**
Input resolution	170×360
Input variables	8
Patch size	8
Embedding dimension	512
Encoder depth	24
Decoder depth	4
Attention heads	16
MLP ratio	4
Trainable parameters	70.092M

A.2.5 Training Configuration

All models were trained using the same dataset partition and forecasting protocol to ensure fair comparison. PE-TA-ViT was optimized using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, together with a weight decay of 1×10^{-5}. The model was trained using a latitude-weighted RMSE objective.

The initial learning rate was set to 1×10^{-4} and decayed following a cosine annealing schedule over 100 epochs. A total batch size of 48 was used with distributed training on four NVIDIA V100 GPUs (32 GB). The checkpoint achieving the lowest validation error was selected for final evaluation.

For baseline models implemented within the same framework, identical training data partitions, optimization settings, and evaluation protocols were adopted to ensure a fair comparison. All experiments were conducted using fixed random seeds for reproducibility.

A.2.6 Autoregressive Forecasting Protocol

PE-TA-ViT performs multi-step forecasting through an autoregressive residual rollout strategy. Given the initial ocean state at time t, the model predicts the low-resolution temporal increment:

$$\Delta \hat{X}^t_{low} = f_\theta(X^t_{low}).$$

The predicted increment is projected onto the high-resolution grid and added to the initial high-resolution state:

$$\hat{X}^{t+1}_{super} = X^t_{super} + \mathcal{I}(\Delta \hat{X}^t_{low}).$$

For subsequent forecasting steps, the previously predicted ocean state is recursively used to construct the input condition for the next prediction. The temporal increment and state update are formulated as:

$$\Delta \hat{X}_{low}^{t+k} = f_\theta(\hat{X}_{low}^{t+k-1}),$$

$$\hat{X}_{super}^{t+k} = \hat{X}_{super}^{t+k-1} + \mathcal{I}(\Delta \hat{X}_{low}^{t+k}).$$

During inference, the model operates in a free-running mode without accessing future ground-truth observations or applying external corrections. The complete 9-day prediction sequence is generated through this autoregressive residual rollout.

A.2.7 Implementation Environment

The proposed model was implemented in PyTorch with CUDA acceleration. Training was performed using distributed training on four NVIDIA V100 GPUs (32 GB), while inference efficiency was benchmarked on a single NVIDIA RTX 3060 GPU (12 GB) to evaluate practical deployment performance. All compared models followed the same dataset partition, preprocessing pipeline, and evaluation protocol to ensure a fair comparison.

A.3 Evaluation Metrics Formulation

Following the IV-TT Class 4 evaluation framework, we utilize two primary statistical metrics to assess the forecast skill: the Root Mean Square Error (RMSE) for deterministic accuracy and the Anomaly Correlation Coefficient (ACC) for spatiotemporal phase alignment. The metrics are mathematically formulated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \tag{20}$$

$$\text{ACC} = \frac{\sum_{i=1}^{N}(F_i - C_i)(O_i - C_i)}{\sqrt{\sum_{i=1}^{N}(F_i - C_i)^2}\sqrt{\sum_{i=1}^{N}(O_i - C_i)^2}} \tag{21}$$

where N denotes the total number of valid oceanic grid points in the evaluated domain. It is important to note that all statistical computations are strictly confined to the effective ocean area, with landmasses and undefined regions explicitly masked out. Here, F_i represents the forecast value generated by our framework at the i-th grid point, O_i denotes the corresponding GLORYS12 ground truth observation, and C_i is the climatological baseline of the target variable.

B Long-Term Temporal Evolution of Forecast Errors

Figure 4 shows the temporal evolution of RMSE for six major variables during the 9-day prediction period evaluated over a full year. For the first-day predictions, all six variables exhibit stable baselines with minimal fluctuations. However, they display occasional sharp error spikes during the Day 2-Day 9 prediction periods. It is observed that periodic error accumulations occasionally align with cross-month transitions, a phenomenon likely reflecting the inherent assimilation shocks within the cyclical update windows of the reference reanalysis datasets (Lellouche et al., 2021; Mulholland et al., 2015). These transient error surges are evident in the initial forecast stage and systematically accumulate across subsequent prediction stages.

Specifically, the RMSE of SST and $thetao$ exhibits an elevated error level between June and September. For most of the year, their RMSE values primarily fluctuate within the ranges of 0.15 - 0.75 °C and 0.15 - 0.65 °C, respectively. In extreme cases, the peak RMSE values reach approximately 0.82 °C and 1.4 °C. The long-term error growth for sea surface height (SSH) is relatively slow, with its Day 9 RMSE primarily fluctuating within the 0.04 - 0.07 m range. In extreme cases, the

peak error reaches 0.09 m. For salinity (so), the RMSE experiences severe spikes from February to April. Subsequently, the long-term forecast error gradually declines and stabilizes, predominantly fluctuating within the 0.05 - 0.5 PSU range. Furthermore, the RMSE for the velocity fields (uo, vo) maintains a relatively flat baseline during most periods, generally staying around 0.05 - 0.15 m/s. Following occasional sharp increases in error (with peaks reaching up to 0.55 m/s), the RMSE rapidly drops back to the baseline level, showing no signs of continuous error divergence. Overall, the temporal evolution indicates that the model's forecast errors remain bounded over the long term.

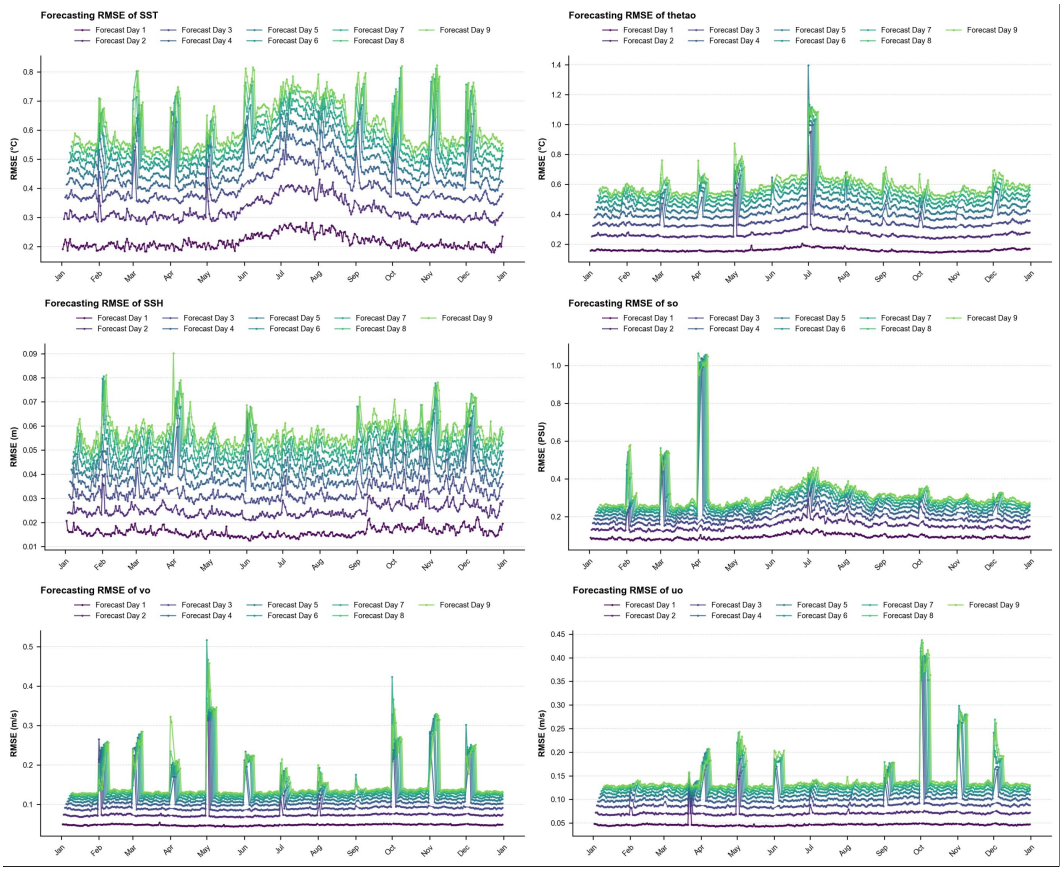

Figure 4: Time series of Root Mean Square Error (RMSE) for various ocean forecasting variables (**SST**, **thetao**, **SSH**, **so**, **vo**, **uo**) at different forecast lead times (Day 1 to Day 9) evaluated over a full year.

C Spatiotemporal Evolution of Global Forecast Errors

To comprehensively evaluate the predictive skill of PE-TA-ViT over the 9-day forecast horizon, we analyze the spatiotemporal evolution of the Root Mean Square Error (RMSE) for key oceanographic variables. Unlike conventional 2D metrics, the 3D stack visualizations (Figure 5) intuitively capture progressive temporal error accumulation and pronounced spatial heterogeneity across different physical fields. The error distribution indicates that variables with shared dynamic characteristics exhibit highly consistent degradation patterns. For the thermal fields (SST and thetao), forecast degradation is concentrated within dynamically active regions—notably the western boundary currents and high-latitude Southern Ocean sectors—with errors approaching $1.00°$C. In contrast, the broader subtropical gyres maintain remarkably high fidelity. Surface velocity predictions (uo and vo) display similar boundary degradation, further characterized by a distinct and continuously elevated error band (up to 0.25 m/s) along the equatorial waveguide. Degradation in sea surface height (SSH) manifests as localized, patchy structures, which visually correspond to the inherent uncertainties

in predicting transient mesoscale eddy phase shifts. Sea surface salinity (so) exhibits the highest overall stability, showing minimal global error propagation. These spatiotemporal patterns confirm that forecast divergence is not driven by random numerical diffusion, but is strictly constrained by underlying ocean physics. PE-TA-ViT effectively sustains large-scale background oceanic states, confining unavoidable forecast uncertainties to regions known to be dominated by intense non-linear advection, equatorial dynamics, and mesoscale instability.

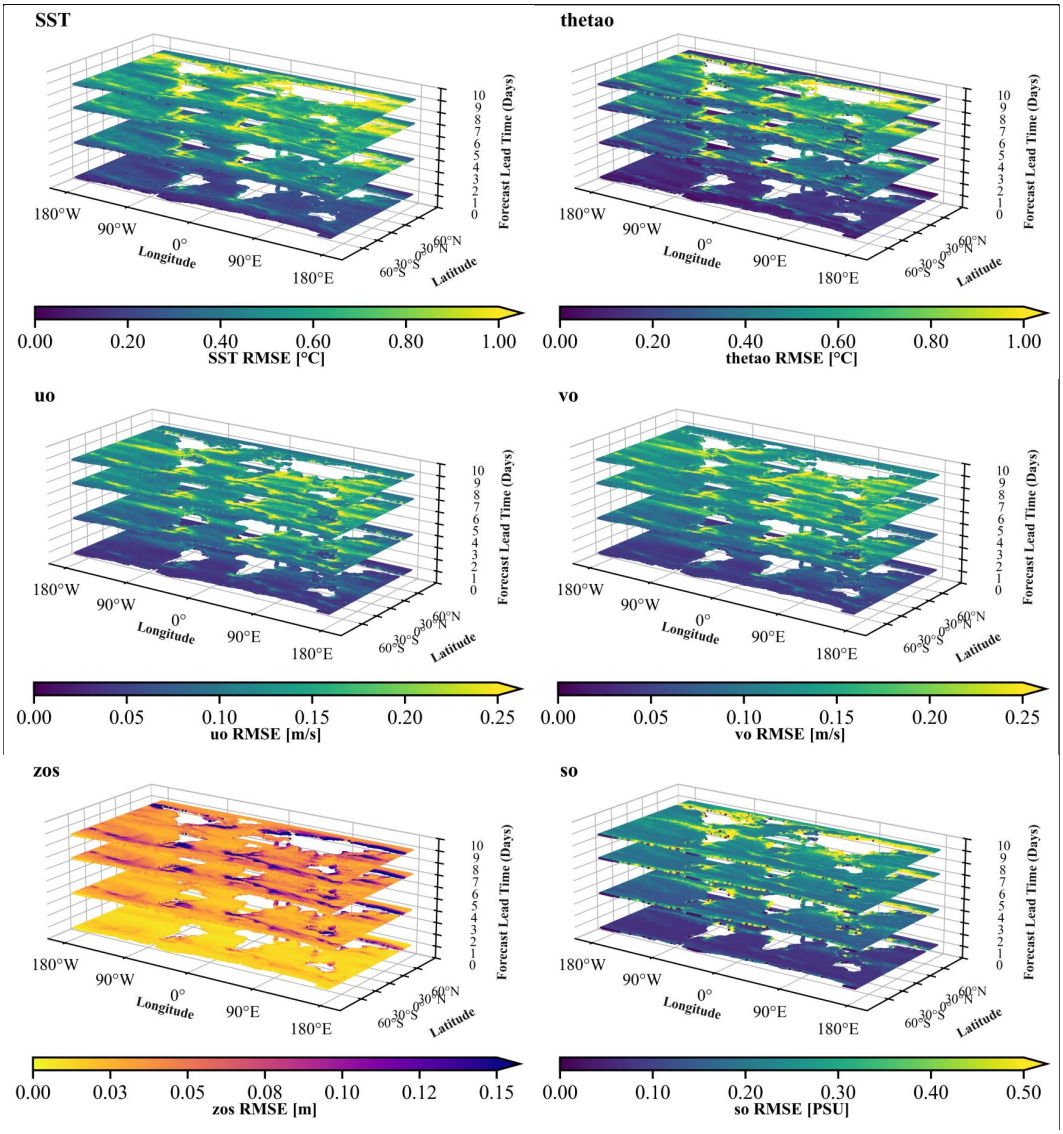

Figure 5: **Spatiotemporal evolution of global forecast errors.** 3D stack visualizations capturing the progressive temporal error accumulation and spatial heterogeneity across different physical variables over the 9-day forecast horizon. (Note: The vertical axis explicitly displays the Day 10 boundary to encompass the complete 9-day integration window.)

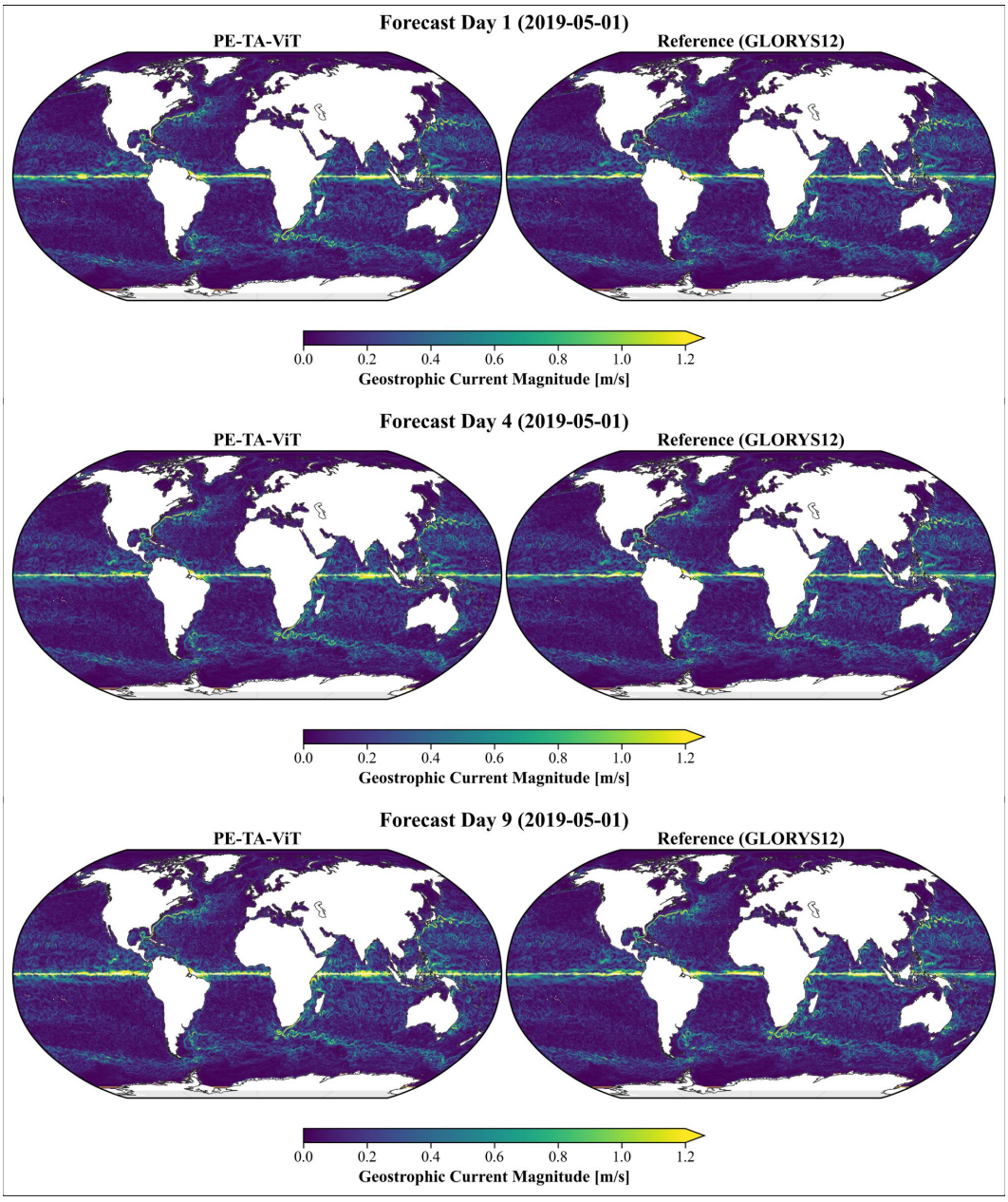

Figure 6: **Spatiotemporal evolution of derived geostrophic current magnitude.** Comparison of PE-TA-ViT forecasts (left) and GLORYS12 reference fields (right) at Days 1, 4, and 9.

D Extended Diagnostic Variables Evaluation

D.1 Geostrophic Current Consistency

Point-wise metrics such as RMSE do not fully assess the spatial-gradient fidelity of predicted ocean fields. Following process-oriented diagnostics adopted in recent neural ocean forecasting studies such as GLONET (El Aouni et al., 2025), we derive geostrophic currents from forecasted SSH to assess the spatial coherence of the predicted fields. Under the traditional geostrophic approximation, the surface geostrophic velocity vector $\mathbf{u}_g = (u_g, v_g)$ is given by

$$\mathbf{u}_g(\phi, \lambda, t) = \frac{g}{f(\phi)} \mathbf{k} \times \nabla_h \eta(\phi, \lambda, t), \tag{22}$$

where g is the gravitational acceleration, $f(\phi)$ is the Coriolis parameter, η denotes SSH, and ∇_h is the horizontal gradient computed in physical-distance coordinates. Because this diagnostic depends directly on SSH gradients, it is sensitive to small-scale noise and spatial discontinuities in the predicted field.

As shown in Figure 6, PE-TA-ViT exhibits strong spatial agreement with GLORYS12 at all reported lead times. The derived current fields retain coherent gradient-driven structures in dynamically active regions, including western boundary currents and mesoscale eddies, without pronounced fragmentation at Day 9. This indicates that the predicted SSH fields retain sufficient spatial-gradient information to recover coherent geostrophic current structures.

Quantitative diagnostics in Appendix D further show gradual error growth in the derived current components, with RMSE increasing from approximately 0.22 m/s at Day 1 to 0.643 m/s for u_g and 0.734 m/s for v_g at Day 9, without abrupt degradation over the evaluated horizon. The observed gradient fidelity is consistent with the gradient-informed inductive bias introduced by PEM-PE and the residual reconstruction strategy, which avoids directly interpolating absolute fields and helps preserve high-resolution spatial structures.

D.2 Annual Geostrophic RMSE Statistics

While Figure 6 in the main text provides a visual snapshot of spatial consistency, we quantitatively evaluate the annual geostrophic accuracy for 2019 in Figure 7. The temporal evolution of the Root Mean Square Error (RMSE) for both the zonal (u_g) and meridional (v_g) velocity components exhibits a monotonic, near-linear increase, reflecting the inherent accumulation of predictive uncertainty. Specifically, the RMSE for u_g rises from 0.225 m/s at Day 1 to 0.643 m/s at Day 9. Similarly, the error for v_g follows a comparable trajectory, increasing from 0.218 m/s to 0.734 m/s. The consistent growth rates across both velocity components suggest that PE-TA-ViT maintains strong dynamical consistency. Furthermore, the absence of abrupt error spikes or exponential growth demonstrates that the model preserves robust numerical stability during extended integration. This rigorous quantitative assessment substantiates the visual findings, confirming controlled error propagation in the derived geostrophic fields throughout the entire 9-day forecast horizon.

D.3 Extended Vorticity Analysis: Formula and Gulf Stream Dynamics

In ocean dynamics, relative vorticity serves as a crucial diagnostic variable to characterize the rotational motion of fluid parcels, offering deep insights into energy distribution, flow stability, and turbulence. Unlike derived quantities assessing inter-variable coherence, vorticity directly probes the system's internal dynamics and cascade-preserving capabilities. The relative vorticity, denoted as ω, is calculated from the zonal (u) and meridional (v) components of the ocean currents:

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{23}$$

where $\frac{\partial v}{\partial x}$ and $\frac{\partial u}{\partial y}$ represent the spatial derivatives of the velocity components with respect to longitude and latitude, respectively. This formulation captures rotational characteristics and acts as a strict diagnostic for multi-scale structure conservation within forecasts.

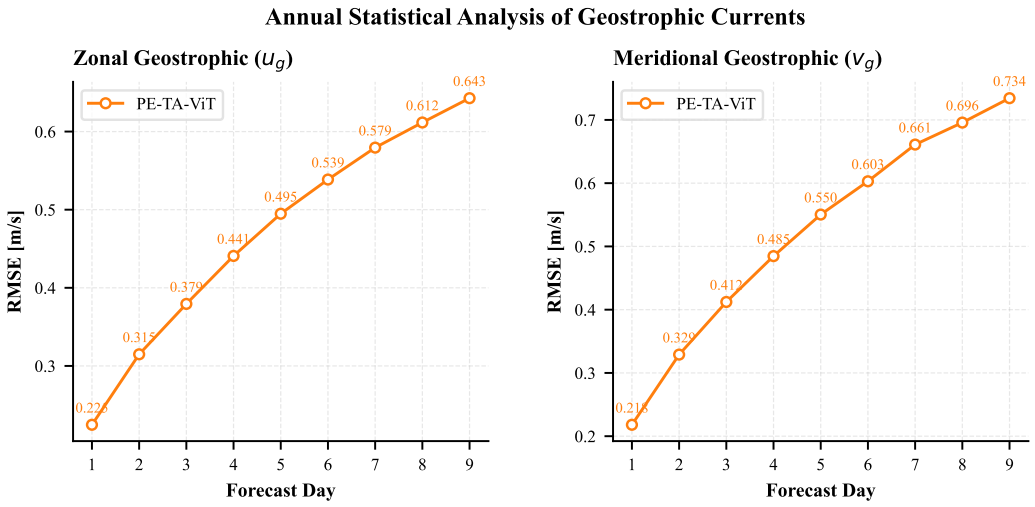

Figure 7: **Annual RMSE statistics of derived geostrophic currents for 2019.** Evolution of RMSE for PE-TA-ViT predicted zonal (u_g, left) and meridional (v_g, right) components over a 9-day forecast.

While Section 4.3 details the model's performance in the Agulhas Retroflection, we provide the corresponding spatiotemporal evolution for the Gulf Stream in Figure 8. The Gulf Stream represents a powerful western boundary current characterized by intense rotational shear and the shedding of mesoscale rings.

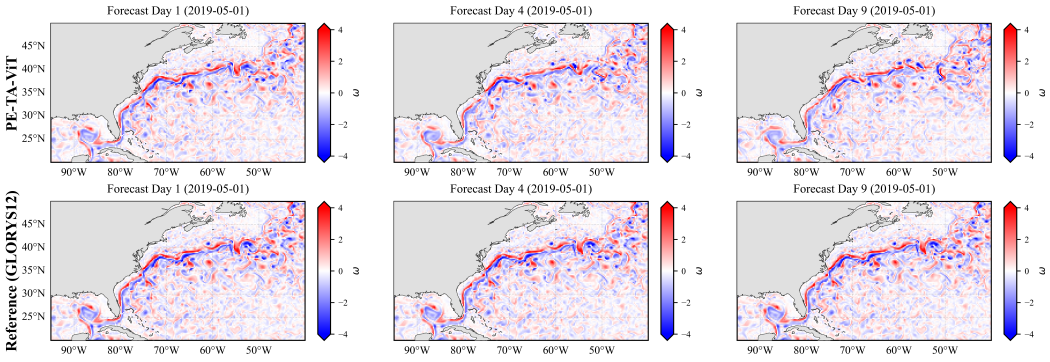

Figure 8: **Spatiotemporal evolution of surface relative vorticity in the Gulf Stream region.** The figure compares the vorticity fields predicted by PE-TA-ViT (top row) against the GLORYS12 reference (bottom row) at Forecast Day 1, Day 4, and Day 9 (initialized on 2019-05-01).

At forecast Days 1 and 4, PE-TA-ViT demonstrates strong visual agreement with the GLORYS12 reference. The model accurately captures the intense shear along the main current axis—manifested as adjacent bands of strong positive and negative vorticity—and distinct cyclonic and anticyclonic eddies.

At the extended 9-day lead time, PE-TA-ViT maintains structural coherence. Purely data-driven models lacking physical constraints often exhibit significant energy cascade attenuation in highly variable regions, leading to a progressive loss of small-scale features and overly smooth fields resembling decaying turbulence. In contrast, PE-TA-ViT retains these energetic, fine-scale rotational structures without undergoing severe numerical dissipation. The spatial variance and amplitude (spanning amplitudes of $\omega \in [-4, 4]$) remain highly comparable to the GLORYS12 reference,

demonstrating that the framework preserves internal dynamic continuity and cross-scale energy transfer irrespective of distinct regional dynamics.

D.4 Energy Cascading Fraction Analysis (E_η)

We also employ a quantitative spectral approach to evaluate the retention of small-scale dynamics. The vorticity fields $\omega(x, y)$ are transformed into the spectral domain using a two-dimensional discrete Fourier transform (DFT) to obtain the power spectrum $|\hat{\omega}(k_x, k_y)|^2$. By radially averaging this spectrum in wavenumber space, we derive the one-dimensional energy spectrum $E(k)$, where $k = \sqrt{k_x^2 + k_y^2}$. To assess the proportion of energy retained across specific spatial scales, we define the energy cascading fraction, E_η, as:

$$E_\eta = \frac{\sum_{k<k_c} E(k)}{\sum_k E(k)} \tag{24}$$

where k_c represents a cutoff wavenumber corresponding to a physical length scale L_c (e.g., $L_c = 100$ km). This dimensionless fraction quantifies the fidelity of the model in preserving the energy cascade characteristics across different lead times compared to the reference data.

As illustrated in Fig. 9, we analyze the evolution of E_η relative to the GLORYS12 ground truth over a 9-day forecast horizon initialized on May 1, 2019. In the Global domain, PE-TA-ViT maintains a close proximity to the reference, with both models centered around 0.863 initially. The maximum divergence occurs on Day 9 with a minor gap of 0.009, underscoring the model's robust preservation of global energy cascade characteristics. In the Gulf Stream, a more pronounced discrepancy is revealed. While PE-TA-ViT follows a smoother trajectory, it fails to capture the intense high-frequency oscillations of the reference, resulting in a maximum error of 0.021 at Day 9. In the Agulhas Retroflection, the model exhibits high consistency with the truth, as both follow a similar downward trend from approximately 0.871 to 0.856. Overall, PE-TA-ViT achieves its highest fidelity in the Global and Agulhas regions, whereas in highly variable regions like the Gulf Stream, it produces a notably regularized prediction.

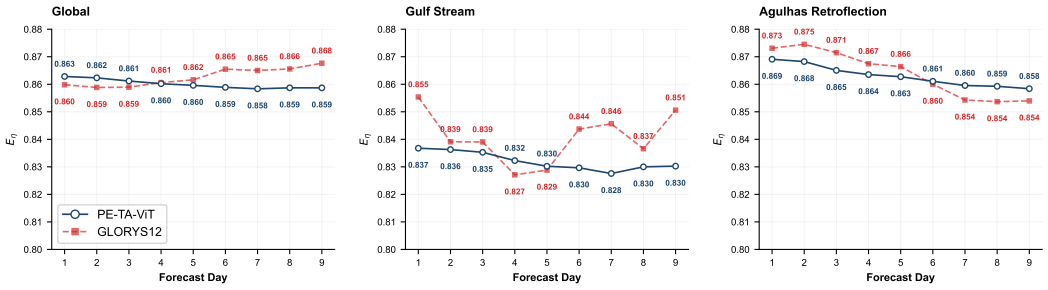

Figure 9: Energy cascading fraction (E_η) evolution across Global, Gulf Stream, and Agulhas Retroflection regions.

E Detailed Visualizations of RMSE and ACC Evolution

This section provides the complete, continuous daily evolution charts for both the Root Mean Square Error (RMSE) and the Anomaly Correlation Coefficient (ACC) across all evaluated near-surface thermodynamic and kinematic variables. The discrete data points extracted from these charts for specific lead times (Day 1, 4, 7, 9) and the corresponding detailed textual analyses are presented in the main manuscript (Tables 1 and 2).

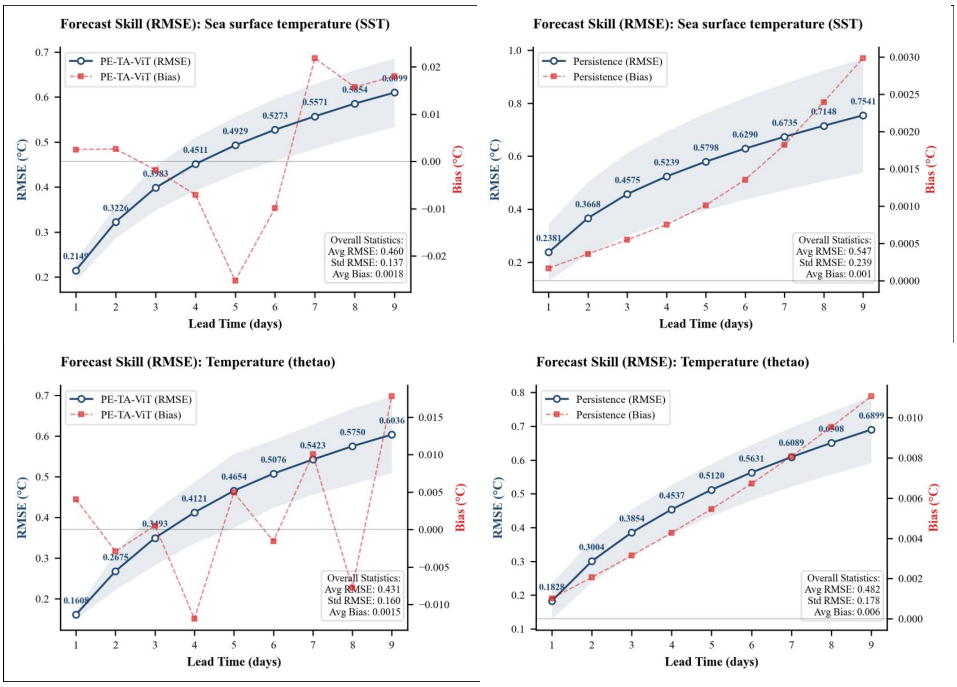

Figure 10: RMSE and Bias evolution for **Sea Surface Temperature (SST)** (left/top) and **Potential Temperature (thetao)** at near-surface depth (0.49 m) (right/bottom). The solid lines trace the spatially averaged RMSE, with the shaded envelopes denoting ± 1 standard deviation.

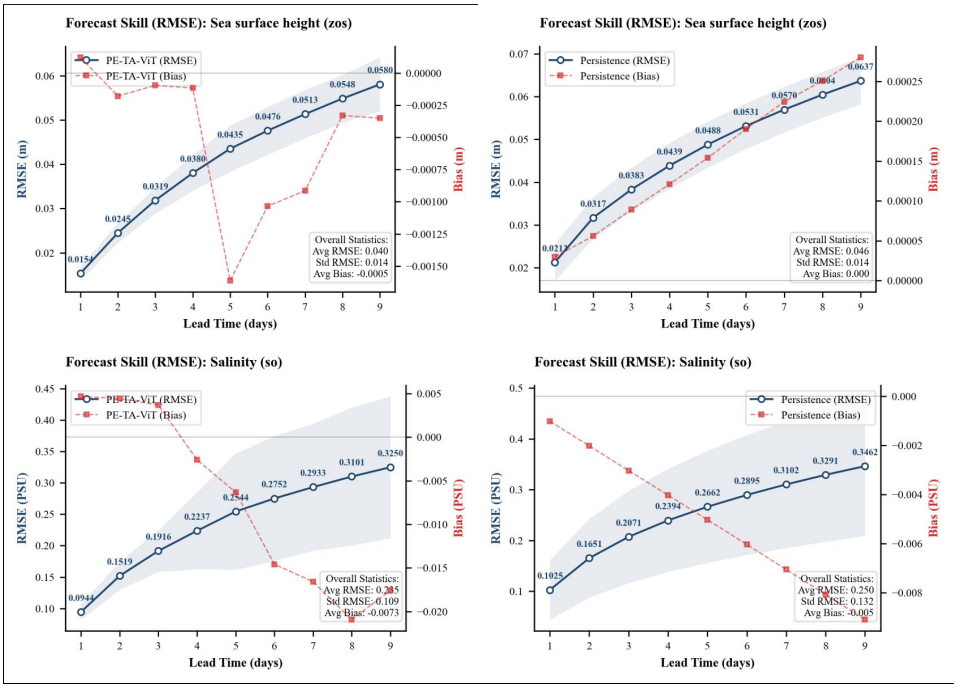

Figure 11: RMSE and Bias evolution for **Sea Surface Height (SSH)** (left/top) and **Near-Surface Salinity (so)** (right/bottom).

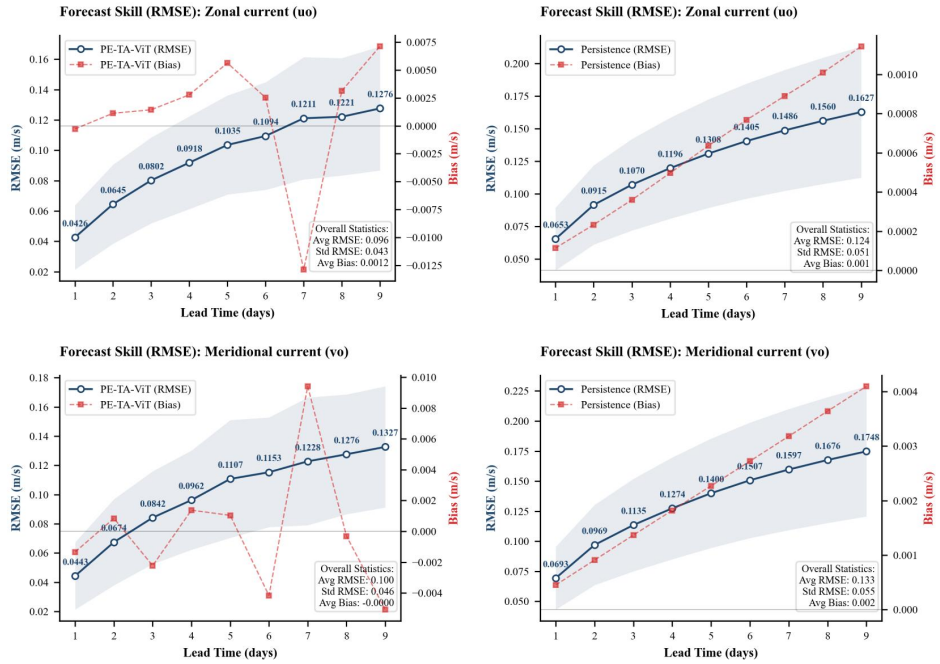

Figure 12: RMSE and Bias evolution for **Zonal Velocity (uo)** (left/top) and **Meridional Velocity (vo)** (right/bottom).

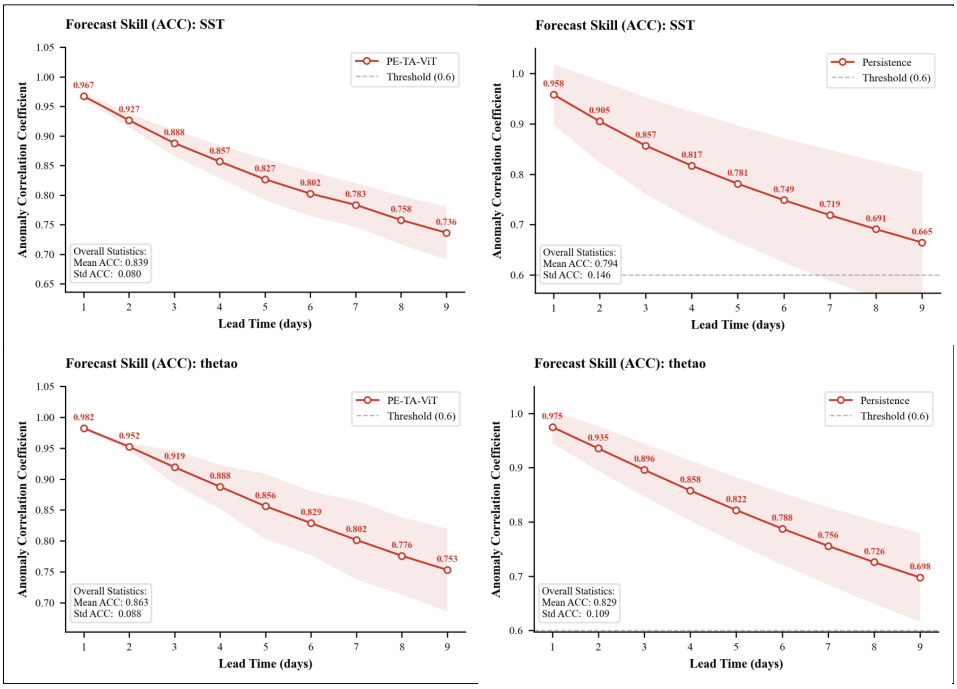

Figure 13: Anomaly Correlation Coefficient (ACC) evolution for **SST** (left/top) and **thetao** (right/bottom). The horizontal dashed line indicates the 0.6 operational threshold.

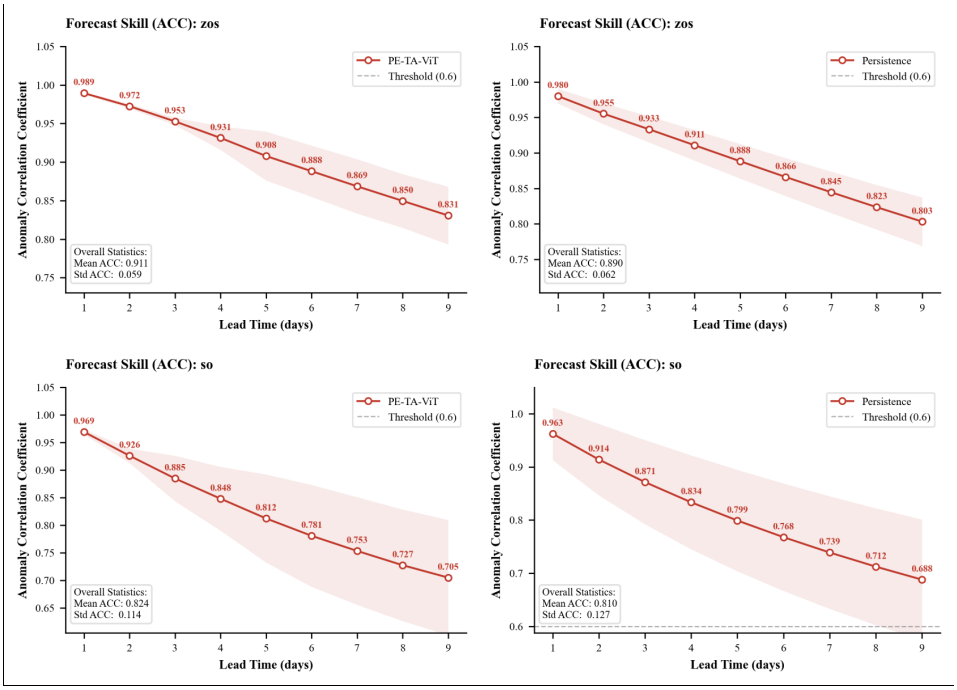

Figure 14: Anomaly Correlation Coefficient (ACC) evolution for **SSH** (left/top) and **so** (right/bottom).

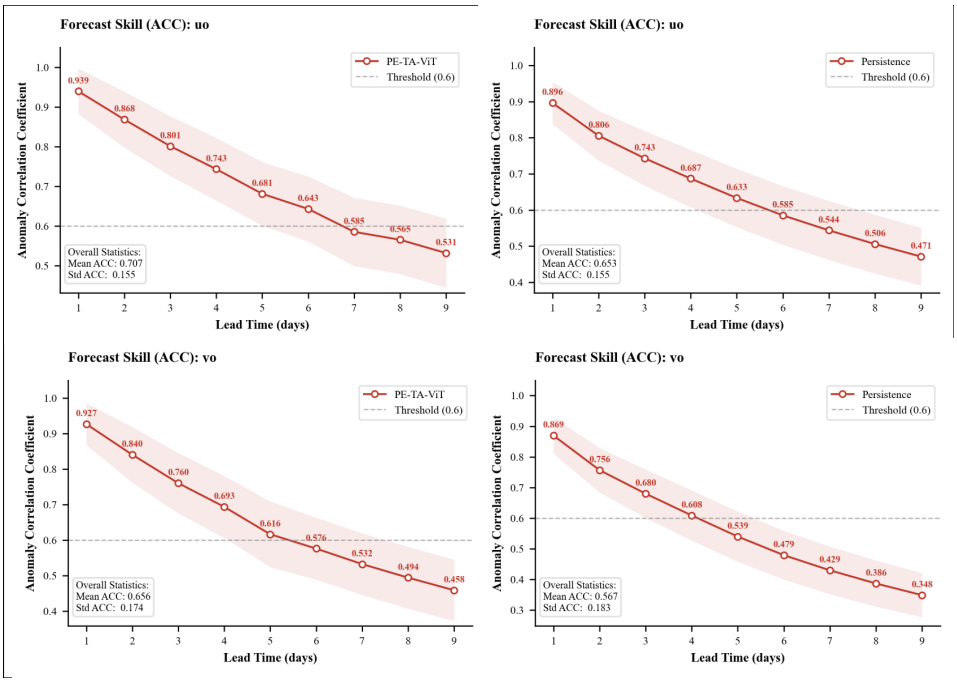

Figure 15: Anomaly Correlation Coefficient (ACC) evolution for **Zonal Velocity (uo)** (left/top) and **Meridional Velocity (vo)** (right/bottom).