EDDYCAST: PHYSICS-AS-BYPASS FOR 3D MESOSCALE EDDY FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing physics-informed ocean forecasting often incorporates physical knowledge through loss penalties or prescribed physical structures, which can limit the flexible integration of physical priors with data-driven representations. We propose Physics-as-Bypass, a physical-embedding paradigm for medium-range 3D mesoscale eddy forecasting. Its zero-initialized residual bypass preserves the data-driven baseline at initialization while enabling adaptive integration of physical priors through a learnable pointwise projection. Based on this paradigm, EddyCast uses fixed analytical operators to provide physically grounded guidance, while EddyCast-VMoD employs learnable vertical modal decomposition to quantify prior-data alignment and modal saturation. We further explore EddyCast-Prob, which equips the shared backbone with an EDM head for probabilistic ensemble forecasting. Experiments show that EddyCast consistently outperforms strong deterministic baselines. Modal ablations and subspace analysis reveal a strong low-rank vertical structure in 3D eddy temperature-salinity fields and show that optimization moves learned representations toward data-aligned physical subspaces. We further explain modal saturation through residual-energy coverage and basis alignment, showing that forecasting-optimal physical dimensionality need not coincide with the effective rank of the data.

1 INTRODUCTION

Mesoscale eddies are a dominant source of ocean variability and play a central role in the transport of heat, momentum, and tracers (Chelton et al., 2007; 2011; Wang et al., 2023). Their evolution involves coupled horizontal and vertical structures across multiple spatial and temporal scales, making accurate 3D forecasting important for both ocean science and practical applications.

Recent advances in deep learning have demonstrated the potential of data-driven models for competitive and computationally efficient ocean forecasting. Convolutional and neural-operator approaches have enabled efficient spatiotemporal prediction (Wang et al., 2024; Pathak et al., 2022), while graph-based, transformer-based, and large-scale forecasting models have substantially extended the forecasting horizon and spatial coverage (Lam et al., 2023; Bi et al., 2023; Nguyen et al., 2023; Kochkov et al., 2024; Chen et al., 2023b;a; Price et al., 2025; Huang et al., 2026). Nevertheless, medium-range 3D forecasting remains challenging. In autoregressive prediction, small transition errors can accumulate over successive lead times, while purely data-driven models only implicitly exploit the physical information contained in observations.

A natural direction is therefore to incorporate physical knowledge into neural forecasting. Existing physics-informed or physics-guided methods introduce physical information through additional input features, physics-based loss terms, explicit constraints, or physics-dependent network structures (Raissi et al., 2019; Karniadakis et al., 2021; Wang et al., 2021). As illustrated in Fig. 1, these approaches couple physical knowledge to the forecasting model at different levels, from the input representation to the training objective and predictive pathway. However, they also raise a more fundamental design question: *how should physical knowledge intervene in a learned forecasting model?* In 3D eddy forecasting, the relevance of physical information can vary across spatial locations, prognostic variables, vertical levels, and dynamical regimes. A physical prior that is informative in one part of the state space may therefore be less useful elsewhere. Treating such a prior as a uniformly active feature or a global training constraint can couple it too tightly to the learned

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.

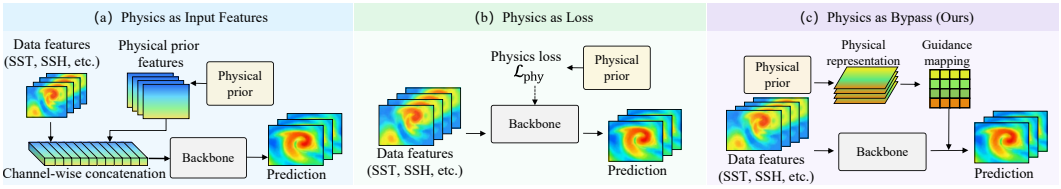


Figure 1: Comparison of three physics-informed learning paradigms for ocean forecasting: (a) Physics as Input Features, (b) Physics as Loss, and (c) Physics as Bypass (Ours).

predictive representation. The challenge is consequently not simply to introduce more physics, but to allow the forecasting model to determine *whether, where, and to what extent* a physical prior should contribute.

We address this problem by proposing *Physics-as-Bypass*, a paradigm for physical embedding in data-driven 3D mesoscale eddy forecasting. Instead of directly coupling a physical prior to the main predictive pathway, we route the prior through an independent bypass branch and introduce its contribution through a zero-initialized projection. At initialization, the zero projection preserves the original data-driven predictor, while optimization can progressively learn the contribution of the physical pathway. In this way, *physical knowledge is available to the predictor without being imposed on it*. The bypass formulation therefore separates the existence of a physical prior from its effective contribution, allowing physical intervention to adapt jointly with the learned data-driven dynamics.

Based on this paradigm, we develop the EddyCast framework and variants for medium-range 3D mesoscale eddy forecasting. The main deterministic model, EddyCast, introduces an analytical Eddy Analytical Guidance (EAG) representation through the physical bypass, providing compact information about local eddy structure and recent evolution without imposing an explicit physical constraint on the predictor. We further develop EddyCast-VMoD, which employs a learnable modal representation over the fixed thermohaline state representation. This formulation enables quantitative comparison between learned modal directions and data-derived dominant structure, allowing us to examine how physical dimensionality and subspace alignment relate to forecasting utility. Finally, we extend the shared forecasting backbone with an EDM-based probabilistic head, yielding EddyCast-Prob for uncertainty-aware prediction.

Our main contributions are summarized as follows:

1. **Physics-as-Bypass.** We propose a physical-embedding paradigm that routes physical priors through an independent, zero-initialized residual pathway, preserving the data-driven predictor at initialization while allowing the effective contribution of physical knowledge to be learned adaptively.
2. **Learnable physical intervention.** We develop complementary analytical and modal physical representations within the bypass pathway, enabling the model to quantify and analyze where and how physical information contributes to the learned forecasting dynamics.
3. **Data-aligned physical dimensionality.** We show that forecasting utility depends jointly on residual-energy coverage and subspace alignment, revealing that the forecasting-optimal physical dimensionality need not coincide with the effective rank of the data.

2 Related Work

2.1 Data-Driven Ocean Forecasting

Deep learning has rapidly advanced data-driven ocean forecasting, evolving from convolutional and recurrent architectures toward transformers and neural operators (Chattopadhyay et al., 2024; Xiong et al., 2023; Xu et al., 2026; Zhang et al., 2024; Huang et al., 2025; Chen et al., 2026). XiHe (Wang et al., 2024) demonstrated high-resolution global ocean forecasting at eddy-resolving scales, establishing the feasibility of learning ocean dynamics directly from reanalysis data. More recent systems have focused on improving long-range stability and 3D consistency. WenHai (Cui et al.,

2025) combines a Swin-Transformer backbone with domain expertise to forecast the global eddying ocean, while NeuralOM (Gao et al., 2026) and AdaptFNO explore neural-operator formulations for learning large-scale ocean dynamics. TSformer (Wang et al., 2025) further adopts a non-autoregressive spatiotemporal Transformer to reduce autoregressive error accumulation and improve extended eddy-resolving forecasting. More recently, LangYa (Yang et al., 2026) introduces explicit lead-time embedding, cross-spatiotemporal atmospheric forcing, and thermocline-aware modeling within a large AI ocean forecasting system. Together, these studies substantially improve the scale, efficiency, and forecast horizon of data-driven ocean prediction.

Despite this progress, most existing methods primarily improve the learned forecasting operator itself, through architectural design, temporal rollout, forcing representation, or operator learning (Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023). The question of how external physical knowledge should be coupled to a strong data-driven predictor remains comparatively less explored. Our work focuses on this representation-level problem rather than introducing another forecasting backbone.

2.2 Physics-Guided and Hybrid Ocean Learning

A parallel line of research incorporates physical knowledge into neural models through physical constraints, process-aware representations, or hybrid architectures (Zhu et al., 2022; Han et al., 2022). Physics-informed learning commonly introduces governing equations or physical residuals into the optimization objective, while physics-guided methods may embed analytical relationships directly into the model (Karniadakis et al., 2021). In ocean modeling, such strategies have been used for learning physical parameterizations and improving physically meaningful predictions (Zhou et al., 2024; Dou et al., 2026). These results demonstrate the value of incorporating domain expertise, but also illustrate a broader design space in which physical information becomes tightly coupled with the forecasting pathway.

Our perspective is different. Rather than imposing a physical constraint or embedding a prior directly into the main representation, we introduce physical knowledge through an independent residual pathway. This formulation makes the strength of physical intervention an explicit and analyzable property of the forecasting model.

3 Method

3.1 Problem Setup

We consider medium-range forecasting of 3D mesoscale ocean states. Let $x_t \in \mathbb{R}^{C\times D\times H\times W}$ denote the ocean state at time t, where C, D, H, and W represent the state variables, vertical levels, and horizontal spatial dimensions, respectively. Given recent states $\mathbf{x}_{t-k+1:t}$, auxiliary environmental conditions $\mathbf{s}_{t+\tau}$, and the normalized forecast lead e_τ, we learn a conditional state-transition operator

$$\hat{x}_{t+\tau} = f_\theta\left(\mathbf{x}_{t-k+1:t}, \mathbf{s}_{t+\tau}, e_\tau\right), \tag{1}$$

where τ denotes the forecast lead.

For medium-range prediction, the learned operator is recursively applied to generate a multi-step trajectory,

$$x_t \to \hat{x}_{t+\tau_1} \to \hat{x}_{t+\tau_2} \to \cdots \to \hat{x}_{t+\tau_n}. \tag{2}$$

This formulation provides the common forecasting setting for the deterministic and probabilistic models developed in the following sections.

3.2 Overview

We propose *Physics-as-Bypass*, a unified physical-embedding paradigm for medium-range 3D mesoscale eddy forecasting. Instead of hard-coding physical priors into the main forecasting pathway, an independent bypass branch maps a physical prior P_ϕ into learnable features and injects them through a zero-initialized pointwise projection. Thus, the model starts from the original data-driven solution and progressively learns how physical priors contribute to forecasting, enabling soft physical embedding without altering the initial data-driven mapping. Based on this paradigm, we

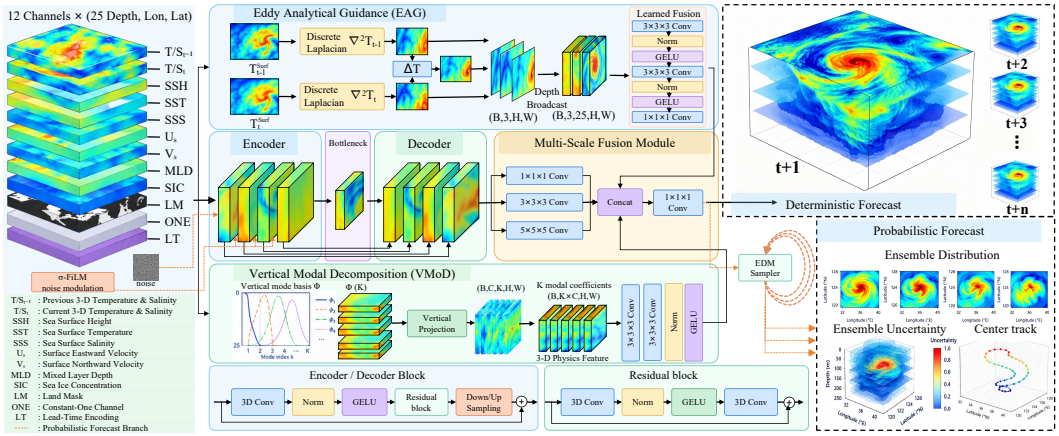


Figure 2: Overview of the proposed EddyCast framework and its variants. Built upon the unified Physics-as-Bypass paradigm, the framework incorporates analytical physical guidance for efficient deterministic forecasting, a learnable vertical-modal branch for physical-subspace analysis, and an EDM-based probabilistic extension for uncertainty-aware multi-day forecasting.

develop three model instances with distinct roles, as illustrated in Fig. 2. *EddyCast* is the deterministic forecasting foundation, incorporating a fixed analytical EAG operator into a compact data-driven backbone for lightweight multi-day 3D mesoscale eddy forecasting. *EddyCast-VMoD* incorporates a learnable vertical-modal branch as an interpretable variant, enabling quantitative analysis of the intrinsic vertical physical subspace, its alignment with data-supported structures, and modal saturation. Finally, *EddyCast-Prob* extends the shared EddyCast backbone with an EDM-based probabilistic head for uncertainty-aware multi-member multi-day forecasting.

3.3 Physics-as-Bypass: A General Physical-Embedding Paradigm

Physics-enhanced forecasting commonly incorporates prior knowledge through either loss-level physical regularization or architecture-level hard-coded constraints. While these strategies can improve physical awareness, they often couple the physical prior tightly with the data-driven representation, making it difficult to control the strength of physical intervention or to accommodate different forms of prior knowledge. We instead formulate physical knowledge as an independent bypass pathway that complements, rather than replaces, the data-driven forecasting backbone. Given an input state x, a general Physics-as-Bypass formulation is defined as

$$\hat{y} = f_\theta(x) + W_0\left(g_\psi(P_\phi(x))\right), \tag{3}$$

where f_θ denotes the data-driven backbone, P_ϕ is a physical-prior operator, g_ψ transforms the resulting physical representation, and W_0 is a learnable pointwise projection initialized to zero. The bypass therefore provides a separate pathway through which physical information can be introduced into the data-driven representation.

The zero initialization of W_0 gives the proposed paradigm a baseline-preserving initialization. Specifically, at initialization,

$$W_0 = 0 \quad \Longrightarrow \quad \hat{y} = f_\theta(x), \tag{4}$$

so that introducing the physical pathway does not alter the initial mapping of the pure data-driven backbone. During optimization, W_0 is updated jointly with the other parameters, allowing the model to gradually learn how strongly the physical representation should contribute to the final prediction. This design separates the existence of a physical prior from its effective contribution, rather than imposing a fixed physical correction throughout the network.

To further characterize this adaptive intervention, consider the scalar abstraction

$$\hat{y}_\alpha = f_\theta(x) + \alpha h_\psi(x), \qquad \alpha_0 = 0, \tag{5}$$

where $h_\psi(x)$ denotes the transformed physical representation. Under a squared-error objective $\mathcal{L} = \frac{1}{2}\|\hat{y}_\alpha - y\|_2^2$, the initial sensitivity of the objective to the bypass strength is

$$\left.\frac{\partial \mathcal{L}}{\partial \alpha}\right|_{\alpha=0} = \langle f_\theta(x) - y,\, h_\psi(x)\rangle = -\langle r_0,\, h_\psi(x)\rangle, \qquad r_0 = y - f_\theta(x). \tag{6}$$

Thus, the early optimization of the bypass is governed by the alignment between its representation and the residual of the data-driven backbone. This provides an interpretable view of physical intervention: the bypass is not required to replace the learned representation, but is optimized according to whether its information is useful for reducing the current prediction error.

Importantly, Physics-as-Bypass does not require a specific form of physical prior. The operator P_ϕ may represent a fixed analytical physical operator or a learnable physically motivated representation, while multiple bypass pathways can be introduced under the same formulation. In this work, we instantiate the paradigm with an analytical physical guidance pathway in EddyCast and a learnable vertical-modal pathway in EddyCast-VMoD.

3.4 Physics-Inspired Representations for Bypass Guidance

Within the Physics-as-Bypass framework, physical information is introduced as an auxiliary representation and injected into the forecasting backbone through the shared bypass mechanism. We consider two complementary forms of physical representation. The first is a fixed analytical representation, which provides physically motivated guidance for mesoscale eddy evolution. The second is a learnable vertical-modal representation, which provides a structured low-dimensional subspace for analyzing the intrinsic vertical organization of the 3D state representation.

Analytical Physical Guidance. We first incorporate an EAG representation to provide local physical guidance for mesoscale eddy evolution. Let SST_t and SST_{t-1} denote the surface temperature fields at the current and previous time steps. We compute their spatial Laplacians as

$$\zeta_t = \nabla^2 SST_t, \qquad \zeta_{t-1} = \nabla^2 SST_{t-1}, \tag{7}$$

and characterize the recent surface-temperature evolution by

$$\delta_T = SST_t - SST_{t-1}. \tag{8}$$

The resulting 2D guidance fields are broadcast along the vertical dimension to match the 3D feature representation:

$$P^{3D}_{\mathrm{EAG}}(d, u, v) = P^{2D}_{\mathrm{EAG}}(u, v), \qquad d = 1, \ldots, D. \tag{9}$$

The Laplacian terms provide local curvature descriptors of the surface-temperature field, while δ_T captures its recent temporal evolution. Together, these quantities provide compact analytical guidance on local mesoscale structures and their short-term variation. The resulting representation is mapped into the feature space of the forecasting backbone and injected through the shared bypass:

$$\hat{x} = f_\theta(v) + W_0\left(g_\psi\left(P^{3D}_{\mathrm{EAG}}(v)\right)\right). \tag{10}$$

Here, f_θ denotes the data-driven forecasting backbone, g_ψ maps the physical representation into the backbone feature space, and W_0 is the learnable pointwise projection defined in Sec. 3.3. Since W_0 is zero-initialized, the resulting model initially has the same functional mapping as the data-driven backbone, while the analytical guidance can be progressively incorporated during optimization. This configuration forms the deterministic forecasting model, EddyCast.

Learnable Vertical Modal Decomposition. To characterize the low-dimensional vertical structure of the 3D state representation, we further introduce a learnable Vertical Modal Decomposition (VMoD) representation. Rather than prescribing the physical subspace after training, we parameterize it directly by a learnable modal basis, whose dimensionality and orientation can be quantitatively compared with the dominant vertical structure of the data.

Let

$$\Phi = [\phi_1; \phi_2; \ldots; \phi_K] \in \mathbb{R}^{K \times D} \tag{11}$$

denote the learnable vertical-modal basis, where D denotes the dimensionality of the ordered vertical state representation and K is the dimensionality of the modal subspace.

$$\phi_1(d) = 1, \tag{12}$$

and initialize the remaining modes with smooth cosine basis functions,

$$\phi_k(d) = \cos\left((k-1)\pi\tilde{d}\right), \qquad k = 2, \ldots, K. \tag{13}$$

where $\tilde{d}$ denotes the normalized index along the ordered vertical-state dimension. The modal basis is subsequently optimized as a learnable parameter without an explicit orthogonality constraint, allowing the representation to adapt to the vertical structures favored by the forecasting task.

Given a backbone feature field $V \in \mathbb{R}^{C\times D\times H\times W}$, the vertical structure of each feature channel is projected onto the modal basis:

$$a_{k,c}(u, v) = \sum_{d=1}^{D} \phi_k(d) V_c(d, u, v). \tag{14}$$

The resulting modal coefficients are transformed into the backbone feature space and introduced through the shared bypass pathway:

$$\hat{x} = f_\theta(v) + W_0\left(g_\psi\left(P_{\mathrm{VMoD}}(V)\right)\right). \tag{15}$$

Here, P_{VMoD} denotes the learnable vertical-modal representation, while g_ψ and W_0 follow the shared bypass mechanism in Sec. 3.3. This shared formulation places the analytical and learnable representations under the same physics-inspired embedding interface.

3.5 Probabilistic Extension

EddyCast provides deterministic forecasts, while medium-range ocean evolution is inherently uncertain. We therefore extend the shared EddyCast backbone with an EDM head (Ho et al., 2020; Song et al., 2021; Karras et al., 2022) for probabilistic forecasting. During training, the target state is perturbed with Gaussian noise, and the network learns to denoise the perturbed target conditioned on recent ocean states, environmental conditions, and forecast lead time. The noisy target is provided as an additional input channel, while the noise level is encoded by a Fourier embedding and injected into the shared backbone through FiLM modulation. The modulation projections are zero-initialized, preserving the learned deterministic representation at initialization. The analytical EAG pathway operates only on the clean conditioning states and remains decoupled from the noisy target.

Implementation details, the EDM parameterization, noise-level sampling, training and sampling procedures, probabilistic metrics, and additional ablation and calibration results are provided in Appendix A.8.

4 Experiments

4.1 Dataset and Experimental Setup

We use the GLORYS12V1 global ocean physical reanalysis produced by the Copernicus Marine Environment Monitoring Service (CMEMS) (Lellouche et al., 2021). The dataset provides daily ocean fields at a horizontal resolution of $1/12^\circ$ on 50 vertical levels. We focus on the Northwest Pacific (120°E–155°E, 10°N–40°N), a region with abundant mesoscale eddy activity, and use the complete record from 2016 to 2025. The input consists of surface variables together with the upper-ocean temperature and salinity fields, while the prediction target contains 25 3D channels: potential temperature on ten subsampled vertical levels and salinity on 15 levels. The 25-channel thermohaline state does not represent a shared physical depth coordinate across temperature and salinity. Instead, the channel axis indexes the prescribed temperature and salinity levels used in the target representation, while the vertical structure of each variable is handled within this fixed representation. The forecasting task uses two consecutive context frames to predict seven future states. The data are split chronologically into training, validation, and test sets with a 8:1:1 ratio.

Table 1: Overall deterministic forecasting performance and computational efficiency averaged over the test set and seven forecasting leads ($L1$–$L7$). Best and second-best forecasting results are shown in bold and underlined, respectively.

Method	RMSE ↓	Skill ↑	MAE ↓	SSIM ↑	PSNR ↑	ACC ↑	PhysRMSE ↓	Params (M)	FLOPs (G)	Time_7 (ms)
Persistence	0.0162	+0.000	0.0102	0.9569	42.91	0.6849	0.2210	–	–	–
WenHai	0.0118	+0.272	0.0078	0.9724	45.86	0.7690	0.1686	6.823	25.729	89.454
Swin3D	0.0119	+0.265	0.0079	0.9724	45.80	0.7798	0.1692	15.316	53.495	88.928
NeuralOM	0.0139	+0.140	0.0097	0.9686	44.04	0.6870	0.2059	0.024	1.468	27.209
AdaptFNO	0.0142	+0.122	0.0093	0.9689	43.87	0.6943	0.1978	0.470	0.207	6.360
FNO3D	0.0144	+0.109	0.0096	0.9662	43.86	0.7021	0.2054	31.865	1.063	9.952
SimVP	0.0175	-0.081	0.0127	0.9650	42.40	0.6423	0.2629	1.899	1.216	6.003
EddyCast (ours)	**0.0106**	**+0.348**	<u>0.0071</u>	<u>0.9827</u>	**46.83**	<u>0.8452</u>	<u>0.1509</u>	2.984	44.441	25.675
EddyCast-VMoD (K=2)	<u>0.0106</u>	<u>+0.346</u>	**0.0070**	0.9827	<u>46.81</u>	**0.8453**	**0.1495**	3.008	46.025	26.221
EddyCast-VMoD (K=4)	0.0110	+0.323	0.0075	**0.9830**	46.49	0.8379	0.1573	3.040	48.040	26.563

We compare EddyCast with seven representative deterministic baselines, including Persistence, Swin3D (Wang et al., 2025), SimVP (Gao et al., 2022), FNO-3D (Choi et al., 2024), WenHai (Cui et al., 2025), NeuralOM (Gao et al., 2026), and AdaptFNO (Dang et al., 2025), under a unified training and evaluation pipeline. All models are implemented in PyTorch and trained on a server equipped with $2\times$ NVIDIA A100-40GB GPUs.

4.2 Evaluation Metrics.

We evaluate deterministic forecasting performance using seven complementary metrics: RMSE, Skill, MAE, SSIM, PSNR, ACC, and PhysRMSE, which jointly characterize numerical accuracy, improvement over the persistence baseline (El Aouni et al., 2025; Han et al., 2026; Rasp et al., 2020; 2024), structural fidelity, spatial correlation, and physical-domain consistency. Unless otherwise specified, RMSE, MAE, SSIM, and PSNR are computed in the normalized domain $[-1, 1]$, while PhysRMSE is computed after channel-wise physical rescaling and reported as a joint thermohaline physical-scale error. All voxel-wise metrics consistently exclude land regions using the same land–sea mask.

4.3 Comparison with State-of-the-Art Methods

Table 1 summarizes the deterministic forecasting performance on the test set over seven forecasting leads ($L1$–$L7$). Overall, EddyCast provides consistently strong performance across the evaluated metrics, consistently outperforming conventional convolutional, attention-based, and neural-operator baselines. These gains indicate that the proposed Physics-as-Bypass design can improve both numerical prediction accuracy and physical-domain consistency without introducing physics as a hard constraint on the data-driven backbone. EddyCast-VMoD exhibits similarly strong performance. Taken together, the results demonstrate that the proposed framework provides a consistent and effective way to incorporate physical information into data-driven eddy forecasting, while maintaining strong general forecasting performance across multiple evaluation criteria. From the computational perspective, EddyCast maintains a compact model size and low inference latency while retaining strong forecasting performance. It uses 2.984M parameters and 44.441G FLOPs per forecasting step, while requiring 25.675 ms for a seven-step rollout. Compared with the high-capacity attention-based baselines, EddyCast uses 56.3% fewer parameters than WenHai and 80.5% fewer than Swin3D, while reducing seven-step inference latency by approximately 71%. Although models such as AdaptFNO and SimVP achieve lower inference latency, their forecasting errors are higher across the reported metrics. These results show that EddyCast achieves strong forecasting performance with a relatively compact parameterization.

4.4 Ablation study

We conduct controlled ablation experiments on the full test set to evaluate different physical-integration paradigms and core architectural components of EddyCast. For the physics-integration study, we use the same Eddy Analytical Guidance (EAG) representation and forecasting backbone to compare Physics-as-Input, Physics-as-Loss, and EddyCast. Physics-as-Input directly concatenates EAG with the input tensor, Physics-as-Loss incorporates EAG through an auxiliary optimization objective, while EddyCast introduces the same EAG representation through the proposed Physics-as-

Bypass pathway. Thus, these comparisons differ only in how the physical information is integrated, with the backbone and training protocol kept unchanged. For the architectural study, we remove surface forcing or replace the depth-aware 3D backbone with a depth-independent 2D design. All variants follow the same seven-day autoregressive rollout protocol.

Table 2 summarizes the results. EddyCast achieves the lowest RMSE, MAE, and PhysRMSE and the highest PSNR and ACC among the physics-integration variants. Under the same EAG representation and forecasting backbone, EddyCast consistently outperforms both Physics-as-Input and Physics-as-Loss, supporting the benefit of the proposed bypass-based integration mechanism. Among the architectural variants, removing surface forcing causes the largest degradation, while removing the depth-aware 3D design also leads to a measurable performance drop, highlighting the importance of both surface environmental information and depth-aware feature extraction.

Table 2: Ablation study on the test set. All values are 7-day rollout means. Physics-integration variants use the same EAG representation, while architectural variants examine surface forcing and depth-aware 3D feature extraction.

	Variant	RMSE↓	MAE↓	SSIM↑	PSNR↑	ACC↑	PhysRMSE↓	skill↑
Physics	No-physics	0.0124	0.0083	0.9812	45.13	0.7534	0.1769	+0.235
	Physics-as-Input	0.0114	0.0079	0.9821	46.09	0.8343	0.1652	+0.294
	Physics-as-Loss	0.0111	0.0075	0.9823	46.28	0.8329	0.1582	+0.316
	EddyCast (Ours)	**0.0106**	**0.0071**	0.9827	**46.83**	**0.8452**	**0.1509**	**+0.348**
Architecture	EddyCast w/o Surf	0.0126	0.0082	0.9725	45.07	0.7649	0.1765	+0.222
	EddyCast w/o 3D	0.0108	0.0072	0.9832	46.79	0.8496	0.1517	+0.334
	Persistence	0.0162	0.0102	0.9569	42.91	0.6849	0.2210	+0.000

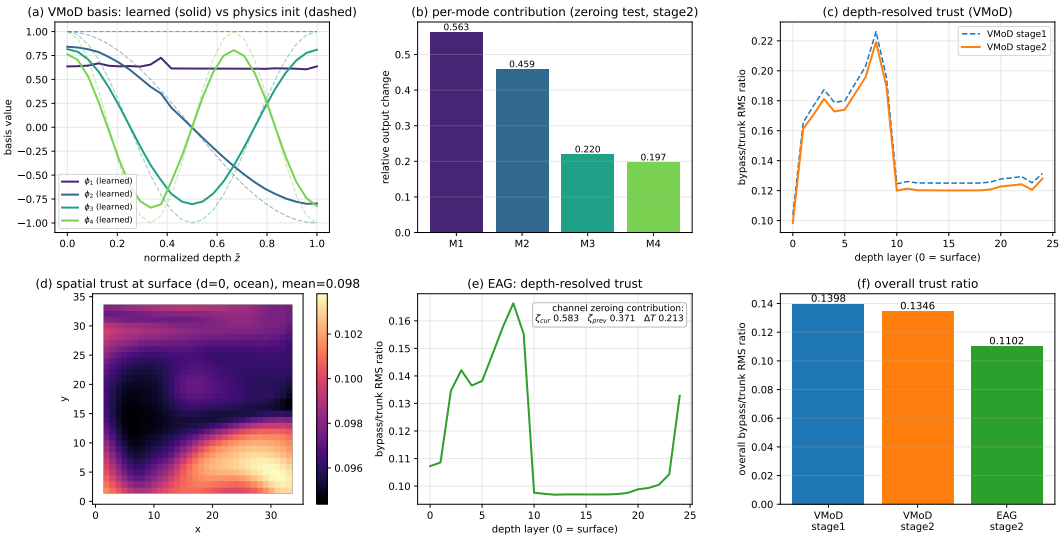


Figure 3: **Learned bypass contributions in the Physics-as-Bypass framework.** Forward-hook probes of the trained physical pathways. (a) Learned VMoD basis compared with the physics-initialized basis. (b) Relative output changes caused by zeroing individual VMoD modes. (c) Depth-resolved bypass-to-trunk RMS ratio for VMoD. (d) Spatial distribution of the learned physical trust at the surface. (e) Depth-resolved trust of the EAG pathway. (f) Overall bypass-to-trunk RMS ratios.

4.5 Learned Bypass Contributions

The central question of Physics-as-Bypass is not whether physical information is available, but whether the network learns to use it when it is forecast-relevant. We therefore probe the trained bypasses directly using forward-hook measurements at the second rollout step. These probes are used as qualitative diagnostics of learned physical trust. Figure 3 shows that the physical pathways

are used selectively rather than uniformly. The learned VMoD basis also deviates from its physics-initialized form, indicating that the modal representation is adapted during optimization rather than remaining fixed. Different VMoD modes and EAG channels make unequal contributions to the prediction, while the spatial and depth-resolved measurements reveal structured variation in the strength of physical corrections. The resulting bypass-to-trunk ratios remain substantially smaller than the main prediction pathway, indicating that the physical branches function primarily as residual corrections rather than replacing the data-driven backbone.

This behavior is consistent with the zero-initialized design in §3.3. At initialization, the physical pathway has no effect on the prediction, and the model therefore starts from the data-driven solution. During optimization, the bypass gradually acquires nonzero influence only when its correction is useful for reducing the forecasting error. The probing results provide empirical evidence that the physical pathways do not simply amplify the injected physical features, but instead exhibit spatially, vertically, and modally varying contributions to the backbone prediction.

4.6 Forecasting Performance across Lead Times

Figure 4 examines the evolution of forecasting quality over the seven-day autoregressive rollout. EddyCast exhibits slower error growth and stronger structural and physical fidelity as the lead time increases. The advantage becomes more evident at longer leads, where autoregressive error accumulation increasingly challenges coherent eddy evolution. These results suggest that the physical bypass contributes not only to short-term correction but also to more stable long-horizon 3D forecasting.

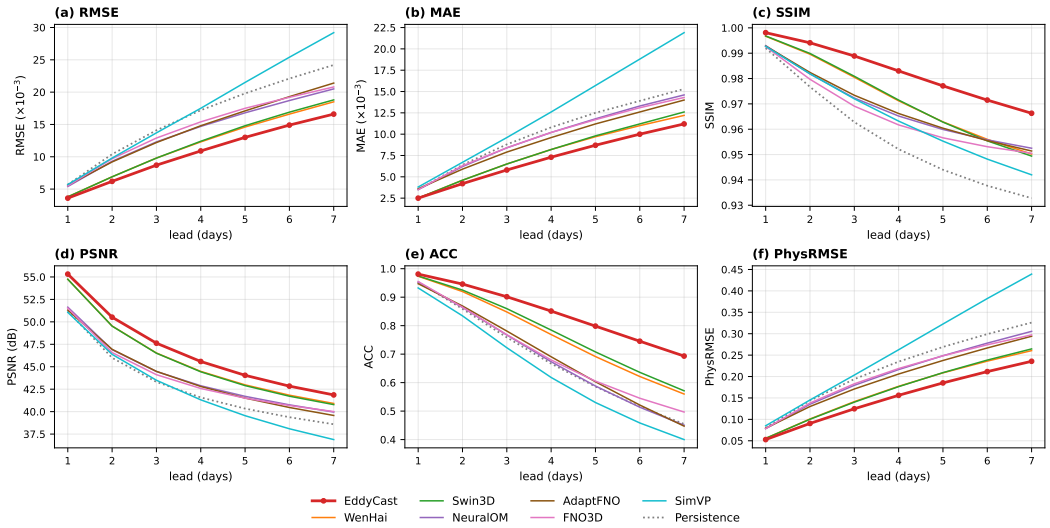


Figure 4: **Forecasting performance across lead times.** Deterministic forecasting performance over the seven-day autoregressive rollout for RMSE, MAE, SSIM, PSNR, ACC, and PhysRMSE. EddyCast exhibits slower error growth and stronger structural and signal fidelity over the forecasting horizon, with the advantage becoming more pronounced at longer leads.

5 Conclusion

We presented EddyCast, a Physics-as-Bypass framework for 3D mesoscale eddy forecasting. By introducing physical knowledge through zero-initialized bypass pathways, EddyCast preserves the flexibility of data-driven forecasting while allowing the model to selectively exploit physical information. Experiments demonstrate improved forecasting accuracy and structural fidelity over representative baselines, while analysis of the learned bypasses shows that physical knowledge contributes primarily as a selective residual correction rather than a hard constraint. These findings support Physics-as-Bypass as a flexible paradigm for integrating physical knowledge into data-driven forecasting.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecasting. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

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*, 14:21181, 2024. doi: 10.1038/s41598-024-72145-0.

Dudley B. Chelton, Michael G. Schlax, Roger M. Samelson, and Roland A. de Szoeke. Global observations of large oceanic eddies. *Geophysical Research Letters*, 34(15):L15606, 2007. doi: 10.1029/2007GL030812.

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

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a.

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, 2023b. doi: 10.1038/s41612-023-00512-1.

Qinghui Chen, Zekai Zhang, Hailong Liu, Jinglin Zhang, and Cong Bai. Kftd: Koopman-fourier time-differentiable network for continuous ocean spatiotemporal forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 94–103. Association for Computing Machinery, 2026. doi: 10.1145/3770854.3780278.

Byoung-Ju Choi, Hong Sung Jin, and Bataa Lkhagvasuren. Applications of the fourier neural operator in a regional ocean modeling and prediction. *Frontiers in Marine Science*, 11:1383997, 2024. doi: 10.3389/fmars.2024.1383997.

Yingzhe Cui, Ruohan Wu, Xiang Zhang, Ziqi Zhu, Bo Liu, Jun Shi, Junshi Chen, Hailong Liu, Shenghui Zhou, Liang Su, Zhao Jing, Hong An, and Lixin Wu. Forecasting the eddying ocean with a deep neural network. *Nature Communications*, 16(1):2268, 2025. doi: 10.1038/s41467-025-57389-2.

Hiep Vo Dang, Bach D. G. Nguyen, Phong C. H. Nguyen, and Truong-Son Hy. Adaptfno: Adaptive fourier neural operator with dynamic spectral modes and multiscale learning for climate modeling. In *NeurIPS 2025 Workshop on Machine Learning and the Physical Sciences*, 2025.

Jiaqi Dou, Huaming Yu, Chunxiao Wang, Ziliang Zhang, Zhenwei Wei, Xin Qi, and Ziyao Wang. Physics-informed deep learning for global ocean surface forecasting by explicitly embedding air-sea bulk-flux formulae. *Ocean Modelling*, 205:102810, 2026. doi: 10.1016/j.ocemod.2026.102810. Published online 29 August 2026.

Anass El Aouni, Quentin Gaudel, Juan Emmanuel Johnson, Charly Regnier, Julien Le Sommer, Simon van Gennip, Ronan Fablet, Marie Drevillon, Yann Drillet, and Pierre-Yves Le Traon. Oceanbench: A benchmark for data-driven global ocean forecasting systems. In *Advances in Neural Information Processing Systems*, 2025.

Yuan Gao, Hao Wu, Fan Xu, Yanfei Xiang, Ruijian Gou, Ruiqi Shu, Qingsong Wen, Xian Wu, Kun Wang, and Xiaomeng Huang. Neuralom: Neural ocean model for subseasonal-to-seasonal simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14756–14764, 2026. doi: 10.1609/aaai.v40i17.38495.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3170–3180, 2022. doi: 10.1109/CVPR52688.2022.00317.

Guoqing Han, Haobin Cen, Jiahan Jiang, Xiaoqian Gao, Xingliang Jiang, Shuyi Zhou, Wenhong Xie, Jinlin Ji, Brandon J Bethel, and Changming Dong. Applying machine learning in devising a parsimonious ocean mixing parameterization scheme. *Deep Sea Research Part II: Topical Studies in Oceanography*, 203:105163, 2022.

Yi Han, Haoming Jia, Xiang Wang, Huizan Wang, Wei Wu, Jianming Zheng, and Peikun Xiao. Oceanforecastbench: A benchmark data set for data-driven global ocean forecasting. *Journal of Geophysical Research: Machine Learning and Computation*, 3(4):e2025JH000838, 2026. doi: 10.1029/2025JH000838.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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.

Qiusheng Huang, Xiaohui Zhong, Anboyu Guo, Ziyi Peng, Lei Chen, and Hao Li. Data-driven ensemble prediction of the global ocean. *arXiv preprint arXiv:2603.19591*, 2026.

George E. Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 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. doi: 10.1126/science.adi2336.

Jean-Michel Lellouche, Eric Greiner, Romain Bourdallé-Badie, Gilles Garric, Angélique Melet, Marie Drévillon, Clément Bricaud, Mathieu Hamon, Olivier Le Galloudec, Charly Regnier, Tony Candela, Charles-Emmanuel Testut, Florent Gasparin, Giovanni Ruggiero, Mounir Benkiran, Yann Drillet, and Pierre-Yves Le Traon. The copernicus global 1/12° oceanic and sea ice glorys12 reanalysis. *Frontiers in Earth Science*, 9:698876, 2021. doi: 10.3389/feart.2021.698876.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 25904–25938, 2023.

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.

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. In *105th Annual AMS Meeting 2025*, volume 105, pp. 449275, 2025.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Souhail Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. 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. doi: 10.1029/2023MS004019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Guosong Wang, Min Hou, Mingyue Qin, Xinrong Wu, Zhigang Gao, Guofang Chao, and Xiaoshuang Zhang. The tsformer: A non-autoregressive spatio-temporal transformers for 30-day ocean eddy-resolving forecasting. *Journal of Marine Science and Engineering*, 13(5):966, 2025. doi: 10.3390/jmse13050966.

Hailin Wang, Bo Qiu, Hanrui Liu, and Zhengguang Zhang. Doubling of surface oceanic meridional heat transport by non-symmetry of mesoscale eddies. *Nature Communications*, 14:5460, 2023. doi: 10.1038/s41467-023-41294-7.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021. doi: 10.1137/20M1318043.

Xiang Wang, Renzhi Wang, Ningzi Hu, Pinqiang Wang, Peng Huo, Guihua Wang, Huizan Wang, Senzhang Wang, Junxing Zhu, Jianbo Xu, Jun Yin, Senliang Bao, Ciqiang Luo, Ziqing Zu, Yi Han, Weimin Zhang, Kaijun Ren, Kefeng Deng, and Junqiang Song. Xihe: A data-driven model for global ocean eddy-resolving forecasting. *arXiv preprint arXiv:2402.02995*, 2024.

Wei Xiong, Yanfei Xiang, Hao Wu, Shuyi Zhou, Yuze Sun, Muyuan Ma, and Xiaomeng Huang. Aigoms: Large ai-driven global ocean modeling system. *arXiv preprint arXiv:2308.03152*, 2023.

Guangjun Xu, Yucheng Shi, Xueming Zhu, Zhao Jing, Shuyi Zhou, Jiexin Xu, Huabing Xu, Guancheng Wang, Dongyang Fu, and Changming Dong. Prediction of three-dimensional ocean temperature, salinity and current fields based on fourier neural operators. *Ocean Modelling*, 200:102674, 2026. doi: 10.1016/j.ocemod.2025.102674.

Nan Yang, Chong Wang, Zimeng Zhao, Meihua Zhao, Huiling Zheng, Bin Zhang, Fan Wang, and Xiaofeng Li. Langya: A large ai model for global ocean forecasting. *Science Bulletin*, 2026. doi: 10.1016/j.scib.2026.08.001.

Ziliang Zhang, Huaming Yu, and Danqin Ren. Oceancastnet: A deep learning ocean wave model with energy conservation. *arXiv preprint arXiv:2406.03848*, 2024.

Shuyi Zhou, Ruizi Shi, Hao Yu, Xueyang Zhang, Jinhui Dai, Xiaomeng Huang, and Fanghua Xu. A physical-informed neural network for improving air-sea turbulent heat flux parameterization. *Journal of Geophysical Research: Atmospheres*, 129(17):e2023JD040603, 2024.

Yuchao Zhu, Rong-Hua Zhang, James N. Moum, Fan Wang, Xiaofeng Li, and Delei Li. Physics-informed deep-learning parameterization of ocean vertical mixing improves climate simulations. *National Science Review*, 9(8):nwac044, 2022. doi: 10.1093/nsr/nwac044.

A Appendix

A.1 Task Definition and Data Details

This subsection provides a self-contained specification of the forecasting task, complementing the problem setup in Sec. 3.1 and the dataset description in Sec. 4.1. Each forecasting step predicts the next daily ocean state from the two most recent daily states; at the first step, the available context state is duplicated so that the operator always receives two context states, and from the second step onward the two most recent states are used. The seven-day forecast is obtained by applying this operator recursively for seven steps, as in the rollout formulation of Sec. 3.1. At every step, the model receives the two context states, the surface environmental variables associated with the target day, and a binary land–sea mask; the lead-aware variants of our backbone additionally receive the normalized forecast lead e_τ. These surface variables are prescribed rather than autoregressively predicted, whereas the 3D context states are updated with the model's own predictions after the first rollout step.

Table 3 summarizes the information of a single forecasting step: the two 25-channel context states, the 7-channel surface environmental variables, and the 1-channel land–sea mask, i.e., 58 input planes in total. The seven surface environmental variables comprise sea-surface height (SSH), sea-surface temperature (SST), sea-surface salinity (SSS), surface eastward and northward currents, mixed-layer thickness, and sea-ice concentration; over the Northwest Pacific domain, the sea-ice concentration channel contains no valid values and is identically zero after missing-value filling.

The prediction target shares the same 25-channel state space as the context states, and its vertical levels are listed in Table 4. The 25 channels follow a fixed variable-specific layout, with the temperature and salinity levels concatenated in channel order; they do not constitute a single shared physical-depth coordinate. Potential temperature is predicted on ten subsampled levels down to $55.76\,\mathrm{m}$ and salinity on all fifteen levels down to $29.45\,\mathrm{m}$, so that the prediction target focuses on the upper-ocean expression of mesoscale eddies within the selected thermohaline representation.

Table 3: Input composition of a single forecasting step.

Block	Channels	Description
$\mathbf{x}_{t-1}$	25	Previous 3-D state: *thetao* on 10 levels and *so* on 15 levels (Table 4)
$\mathbf{x}_t$	25	Most recent 3-D state, same channel layout
$\mathbf{s}_{t+\tau}$	7	Surface environmental variables associated with the target day: SSH, SST, SSS, surface u/v, mixed-layer thickness, and sea-ice concentration
Mask	1	Binary land–sea mask
Total	58	Two context states, surface environmental variables, and land–sea mask

Every channel is min–max normalized to $[-1, 1]$ using per-channel statistics computed from the training split only. During training, the analysis window is subject to a random ± 18-pixel crop jitter, whereas evaluation uses the centered window without jitter. All voxel-wise metrics are computed inside the ocean mask, and predicted land cells are back-filled with the persisted initial state at every rollout step; this protocol is applied identically to all compared methods.

Autoregressive Fine-tuning. The deterministic model is first trained from scratch for 200 epochs and then fine-tuned for 50 epochs from the best Stage-A checkpoint. During fine-tuning, teacher

Table 4: Variable-specific vertical levels of the 25 forecast channels.

Variable	Levels	Depth (m)
thetao (°C)	10	0.49, 2.65, 5.08, 7.93, 11.41, 15.81, 21.60, 29.45, 40.34, 55.76
so (psu)	15	0.49, 1.54, 2.65, 3.82, 5.08, 6.44, 7.93, 9.57, 11.41, 13.47, 15.81, 18.50, 21.60, 25.21, 29.45

forcing and autoregressive inputs are mixed with a constant replacement probability of 0.5, without probability annealing. The model is rolled out over all prediction leads, with the losses averaged uniformly across leads. Recycled predictions are detached before being fed into the next step, so gradients do not propagate across rollout steps. At each rollout step, the replacement decision is sampled once per batch and applied to the batch.

Training Objective To improve multi-step stability, training proceeds in two stages—200 epochs of teacher forcing followed by 50 epochs of fine-tuning with a constant autoregressive replacement probability of 0.5—narrowing the gap between training and autoregressive inference. An exponential moving average (EMA) of the parameters is maintained and used for all evaluations. EddyCast-Prob warm-starts from the deterministic EddyCast checkpoint and is fine-tuned with the EDM denoising objective under the same two-stage recipe: after teacher-forced denoising, we apply a rollout-exposure scheme analogous to scheduled sampling, in which the conditioning frames are replaced, with a fixed probability, by the model's own sampled rollouts.

A.2 Vertical Structure of the Physical Representation

We examine whether the structured channel representation exploited by VMoD aligns with intrinsic properties of the ocean data. Figure 5 shows the leading vertical EOFs, their variance spectrum, and the learned VMoD basis. The variance is highly concentrated in a small number of leading vertical modes, demonstrating strongly low-dimensional vertical structure within the 3D fields. The learned VMoD basis maintains close alignment with the dominant data-derived subspace after optimization. This supports the effectiveness of the learnable modal representation for this forecasting task: rather than independently modeling all depth levels, VMoD provides a compact basis that captures dominant vertical variability while retaining task-oriented flexibility.

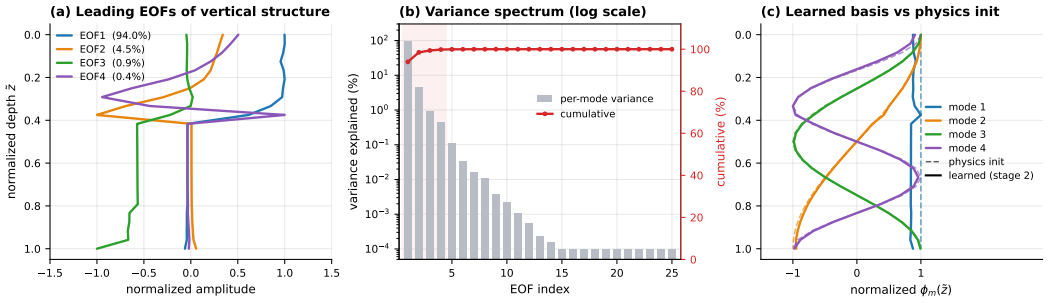


Figure 5: **Vertical structure and learned modal representation.** (a) Leading vertical EOFs of the ocean fields across the 25 depth levels. (b) Variance spectrum and cumulative explained variance, revealing a strongly low-rank vertical structure. (c) Comparison between the physics-initialized and learned VMoD bases. The learned basis remains aligned with the dominant data-derived physical subspace while allowing the individual modes to adapt during training.

A.3 Physical Subspace Analysis

The learnable vertical representation enables explicit subspace-oriented analysis of physical dimensionality. We use this framework to characterize intrinsic vertical complexity of 3D ocean states and quantify compatibility between learned modal representations and data-derived dominant structures.

Intrinsic Vertical Structure. Let Σ_e denote the covariance matrix of vertical state anomalies, with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D. \tag{16}$$

The corresponding leading eigenvectors define the dominant vertical EOF subspace of the data. We measure the fraction of vertical variance captured by the leading K modes as

$$C(K) = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{D} \lambda_i}, \qquad R(K) = 1 - C(K). \tag{17}$$

Here, $C(K)$ measures the cumulative vertical energy captured by the leading K data modes, while $R(K)$ denotes residual energy outside the data-dominant subspace. Rapidly increasing $C(K)$ indicates that 3D ocean states concentrate within a low-dimensional vertical subspace.

Prior–Data Alignment. Variance coverage alone is insufficient to quantify compatibility between physical representations and observed data structures. We compute alignment between the learned modal subspace and leading data EOF subspace. Let $Q_e \in \mathbb{R}^{D\times K}$ contain the leading K data EOFs, and Let $Q_\phi \in \mathbb{R}^{D\times K}$ denote the orthonormal basis obtained by applying reduced QR decomposition to the transpose of the learned modal matrix:

$$\Phi^\top = Q_\phi R, \qquad Q_\phi^\top Q_\phi = I_K. \tag{18}$$

This orthonormalization step is used exclusively for this subspace analysis and is not applied during VMoD model training.

We define

$$A(K) = \frac{1}{K} \left\| Q_\phi^\top Q_e \right\|_F^2 = \frac{1}{K} \sum_{i=1}^{K} \cos^2 \theta_i, \tag{19}$$

where θ_i are principal angles between the two subspaces. Larger $A(K)$ corresponds to higher alignment between learned modal basis and data EOF subspace.

This metric quantifies geometric subspace overlap. Variance contributions of individual directions are captured separately by the EOF coverage metric $C(K)$.

Two-Factor Representational Metric. Residual energy $R(K)$ and alignment score $A(K)$ capture two complementary quantities describing modal representation properties. We define a combined metric to jointly characterize representational properties of the modal subspace:

$$G(K) \propto R(K)A(K) = \big(1 - C(K)\big)A(K). \tag{20}$$

$G(K)$ combines residual unexplained vertical variance and subspace alignment to study modal representation behavior under varying dimensionality.

A.4 Evaluation Metrics

We evaluate the forecasting performance from seven complementary perspectives, including voxel-wise numerical accuracy, forecasting skill, structural similarity, signal fidelity, spatial correlation, and physical-domain consistency. Unless otherwise specified, $\hat{y}$ and y denote the predicted and ground-truth ocean states in the normalized domain $[-1, 1]$. A binary land–sea mask m is consistently applied to exclude land regions from voxel-wise evaluation. All metrics are first computed independently for each test case and forecasting lead, and are then averaged over test cases and seven forecasting leads ($L1$–$L7$).

Root Mean Square Error (RMSE). RMSE measures the overall voxel-wise discrepancy between the predicted and ground-truth fields:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathbf{x}} m(\hat{y} - y)^2}{\sum_{\mathbf{x}} m}}. \tag{21}$$

Lower RMSE indicates more accurate numerical forecasting.

Forecasting Skill. We use Skill to quantify the improvement over the persistence baseline:

$$\text{Skill} = 1 - \frac{\text{RMSE}_{\text{model}}}{\text{RMSE}_{\text{persist}}}, \qquad \hat{y}_{\text{persist}} = x_0, \tag{22}$$

where x_0 denotes the initial ocean state. The persistence forecast directly repeats x_0 at all seven forecasting leads. A positive Skill indicates that the model outperforms persistence, while a larger value indicates a greater relative improvement.

Mean Absolute Error (MAE). MAE measures the average absolute voxel-wise discrepancy:

$$\text{MAE} = \frac{\sum_{\mathbf{x}} m|\hat{y} - y|}{\sum_{\mathbf{x}} m}. \tag{23}$$

MAE is evaluated in the normalized domain, with lower values indicating better forecasting accuracy.

Structural Similarity (SSIM). To evaluate the preservation of mesoscale eddy structures, we employ the structural similarity index (SSIM). SSIM is computed independently on each two-dimensional slice and averaged across depth levels and channels:

$$\text{SSIM} = \frac{1}{|\Omega|} \sum_{\mathbf{x}} m \frac{(2\mu_{\hat{y}}\mu_y + C_1)(2\sigma_{\hat{y}y} + C_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2)}, \tag{24}$$

where $\mu_{\hat{y}}$ and μ_y are the local means, $\sigma_{\hat{y}}^2$ and σ_y^2 are the local variances, and $\sigma_{\hat{y}y}$ is the local covariance. We use a 7×7 uniform window with $C_1 = 0.01^2$ and $C_2 = 0.03^2$. Higher SSIM indicates better preservation of spatial structures.

Peak Signal-to-Noise Ratio (PSNR). PSNR measures the signal fidelity of the predicted field and is computed as

$$\text{PSNR} = 20 \log_{10} \frac{2}{\text{RMSE}}, \tag{25}$$

where 2 is the dynamic range of the normalized data. Higher PSNR corresponds to lower prediction error.

Anomaly Correlation Coefficient (ACC). ACC is the standard measure of spatial agreement between predicted and ground-truth anomaly fields in geophysical forecasting. Let $c_d(\mathbf{x}, k)$ denote the daily climatology, i.e., the day-of-year mean of the training-set target field at grid point $\mathbf{x}$ and vertical channel k, smoothed with a 31-day circular window. The anomaly of a field $\hat{y}, y$ verified on day-of-year d is

$$\hat{y}' = \hat{y} - c_d, \qquad y' = y - c_d. \tag{26}$$

ACC is computed independently for each of the $K = 25$ vertical channels and then averaged:

$$\text{ACC} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{\mathbf{x}} m \left(\hat{y}'_k - \overline{\hat{y}'_k}\right)\left(y'_k - \overline{y'_k}\right)}{\sqrt{\sum_{\mathbf{x}} m \left(\hat{y}'_k - \overline{\hat{y}'_k}\right)^2} \sqrt{\sum_{\mathbf{x}} m \left(y'_k - \overline{y'_k}\right)^2}}, \tag{27}$$

where m is the ocean mask, $\bar{\cdot}$ denotes the spatial mean over the evaluation domain, and the summation $\sum_{\mathbf{x}}$ runs over masked grid points. Since correlation is invariant to per-channel affine transformations, ACC computed on the normalized fields equals that computed on the physical fields. Higher ACC indicates stronger spatial agreement between the predicted and observed anomaly fields; ACC is reported directly in all tables and figures.

Physical-domain RMSE (PhysRMSE). To assess prediction error on the original physical scales, we first compute the RMSE for each target channel in the normalized domain and then convert it using the channel-specific min–max range:

$$\text{RMSE}_c^{\text{phys}} = \text{RMSE}_c^{\text{norm}} \frac{h_c - l_c}{2}, \tag{28}$$

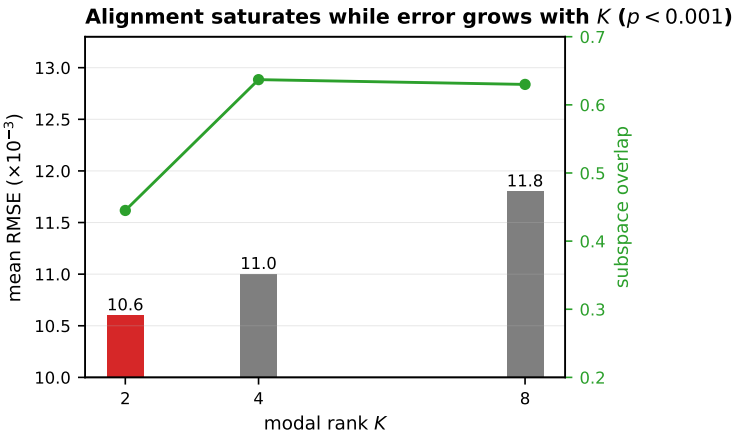


Figure 6: **Modal-rank sweep for VMoD.** Forecasting error and alignment with the data-derived EOF subspace as the number of VMoD modes K varies. Forecasting error increases with additional modal capacity, whereas subspace alignment saturates after the dominant physical structure has been captured. The result distinguishes predictive modal rank from the effective rank of the underlying vertical data distribution.

where h_c and l_c denote the upper and lower bounds of the channel-wise normalization range, respectively. We then report the arithmetic mean across the 25 thermohaline channels:

$$\mathrm{PhysRMSE} = \frac{1}{25}\sum_{c=1}^{25} \mathrm{RMSE}_c^{\mathrm{phys}}. \tag{29}$$

Because the target channels combine potential temperature and practical salinity, this quantity is used as a joint physical-scale error measure.

A.5 Modal-rank Sweep

The number of vertical modes K sets the representational capacity of the VMoD representation. The effective rank estimated from raw data provides one reference for modal capacity. We perform controlled experiments sweeping K while keeping architecture and training settings fixed.

Figure 6 reports forecasting error alongside subspace alignment between learned modal basis and leading EOF subspace. Increasing modal dimension K does not yield monotonic forecasting improvements. Forecasting error rises as additional modes are introduced, while subspace alignment first improves and then saturates. This demonstrates that the modal dimension which best reconstructs raw data variance is not guaranteed to produce optimal forecasting performance.

A.6 Physical-Domain Errors Across Methods and Leads

The lead-time analysis in the main text reports aggregate scores, which can conceal how different methods accumulate error during the autoregressive rollout. Figure 7 therefore disaggregates the deterministic comparison by reporting the physically scaled RMSE (PhysRMSE) of all eight methods at each of the seven leads on the test set, with rows sorted by the mean over leads.

Three key observations emerge from the disaggregated results. First, EddyCast achieves the lowest PhysRMSE at every forecast lead, with its advantage becoming more pronounced at longer lead times. Second, the methods exhibit differing rates of error accumulation during autoregressive rollout. EddyCast increases from 0.053 to 0.236 over the seven-day horizon, whereas SimVP exhibits the steepest degradation, increasing from 0.085 to 0.439 and falling behind persistence from $L5$ onward. Third, lead-wise ranking is broadly consistent with aggregate results in the main text. These results illustrate the effectiveness of the physical bypass for mitigating error accumulation in multi-step forecasting.

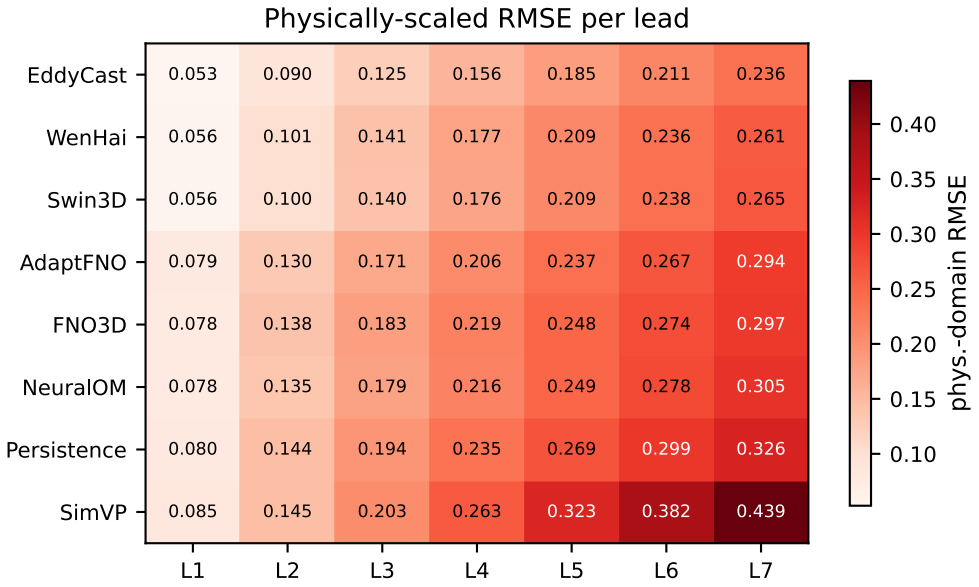


Figure 7: **Physically scaled RMSE (PhysRMSE) at each forecast lead on the test set.** Rows are sorted by the mean PhysRMSE over leads. EddyCast achieves the lowest PhysRMSE at every forecast lead and shows the smallest absolute error growth over the seven-day rollout.

A.7 Additional Qualitative Forecasting Results

To further examine the long-horizon forecasting behavior of EddyCast, we present a qualitative comparison for test case 63 in the northwestern Pacific, covering 136.0°–138.9°E and 23.5°–26.4°N. The case is initialized on 26 August 2025, with daily forecasts from 27 August (day 1) to 2 September 2025. Figures 8–11 cover potential temperature and salinity at multiple vertical levels. Rows correspond to forecast lead times from day 1 to day 7, while columns show the GLORYS12V1 reanalysis (Truth), EddyCast, and six deterministic baselines.

Qualitative inspection shows distinct behaviour across models as rollout proceeds. Baseline models tend to smooth mesoscale features, weaken local gradients and distort coherent eddy structures over successive prediction steps. In comparison, EddyCast maintains closer spatial correspondence with reference fields and preserves coherent mesoscale patterns across multi-step forecasts. This pattern holds for both temperature and salinity variables across multiple depth levels. The physical-aware bypass mechanism supports structural continuity of evolving 3D ocean states during autoregressive forecasting.

The qualitative analysis further investigates vertical prediction quality. Figure 12 examines temperature profiles at three representative locations: eddy core, eddy edge, and domain mean across the full forecast sequence. Vertical eddy evolution exhibits location-dependent structural changes rather than simple uniform depth-wise temperature shifts. EddyCast reproduces reference vertical profiles well for both near-surface and deeper thermal gradients. At day-7 forecast horizon, EddyCast retains the overall shape and turning points of reference profiles, while baseline models show larger deviations in deeper layers. These empirical observations illustrate that our approach preserves not only horizontal mesoscale features but also vertically coherent 3D ocean structures.

A.8 EddyCast-Prob: Additional Details

As an auxiliary extension, we equip the shared EddyCast backbone with an EDM-based diffusion head to enable probabilistic ensemble forecasting. This validates that our Physics-as-Bypass paradigm can be seamlessly combined with stochastic prediction pipelines.

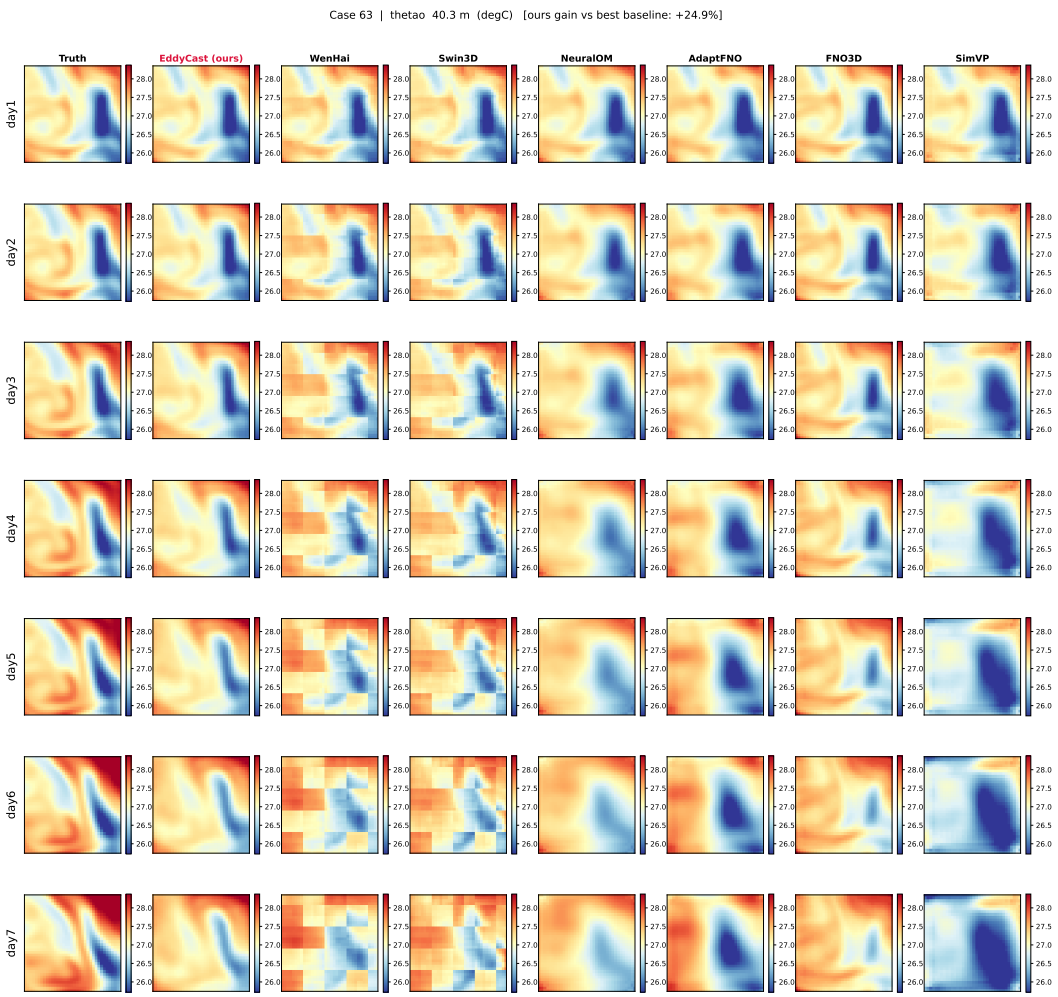


Figure 8: Qualitative comparison for test case 63 at 40.3 m depth ($thetao$, °C).

EddyCast-Prob is warm-started from the pre-trained deterministic EddyCast checkpoint and fine-tuned with the EDM denoising objective. The analytical EAG bypass branch operates on clean conditioning states and remains decoupled from noisy diffusion targets. During inference, ensemble samples are generated via DPM-Solver++(2S). For test-set evaluation we use $M = 4$ ensemble members, 10 sampling levels and churn $\gamma = 0.75$.

Probabilistic forecasting performance is evaluated using CRPS, ensemble spread and spread-to-error ratio. Full EDM parameterization, training objectives and sampling-sensitivity ablation studies are omitted for brevity and will be released together with our open-source code repository.

Table 5 reports lead-wise probabilistic forecasting performance on the test set. Table 6 provides comparisons against alternative denoising backbones and a Gaussian-perturbation baseline under identical sampling budget.

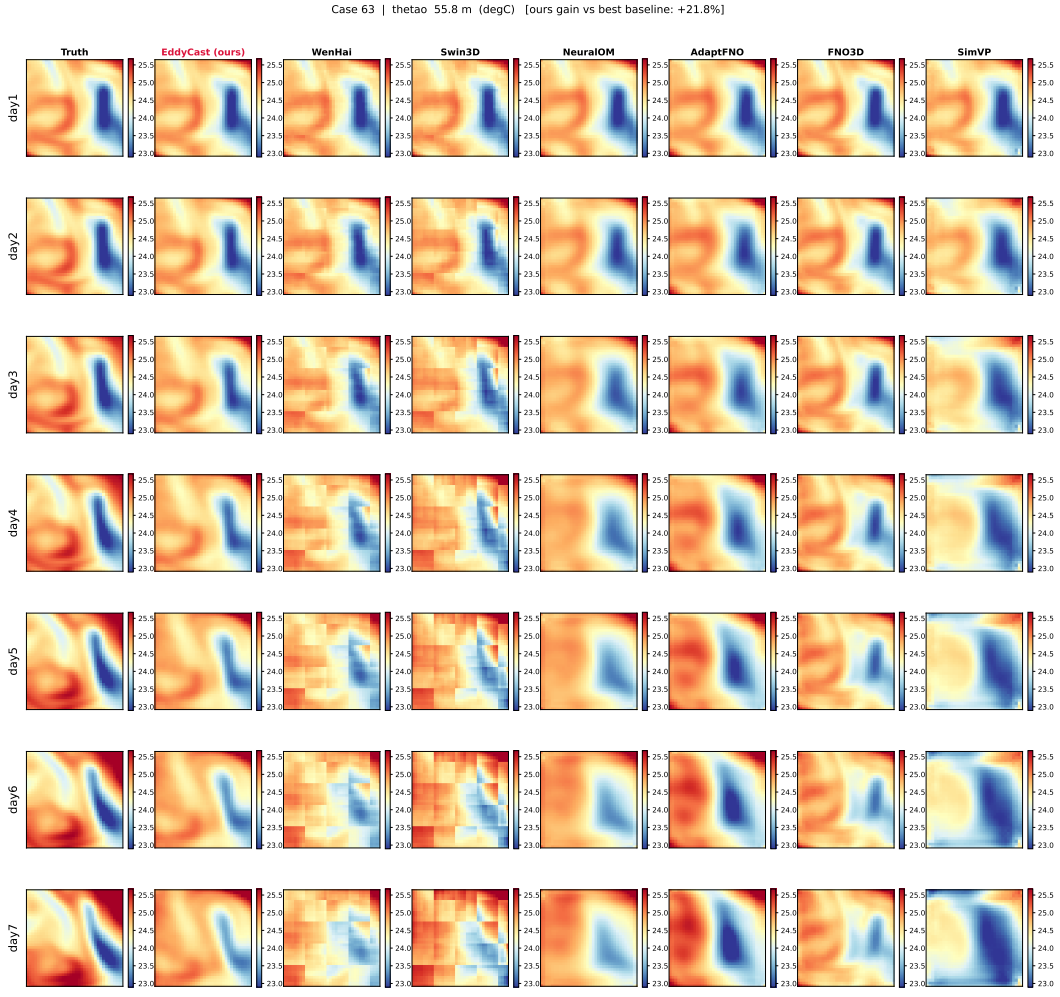


Figure 9: Qualitative comparison for test case 63 at 55.8 m depth ($thetao$, °C).

Table 5: Probabilistic forecasting performance of EddyCast-Prob at each forecast lead on the test set. We use $M = 4$ ensemble members, 10 DPM-Solver++ (2S) sampling levels, churn $\gamma = 0.75$. Metrics are computed within the ocean mask over the normalized $[-1, 1]$ domain.

Lead	RMSE	CRPS	Spread	Spread/RMSE
$L1$	0.0039	0.0017	0.0026	0.66
$L2$	0.0069	0.0033	0.0038	0.55
$L3$	0.0096	0.0044	0.0054	0.57
$L4$	0.0122	0.0060	0.0058	0.48
$L5$	0.0146	0.0074	0.0060	0.41
$L6$	0.0165	0.0081	0.0072	0.44
$L7$	0.0186	0.0092	0.0081	0.44
Mean	0.0117	0.0057	0.0056	0.47

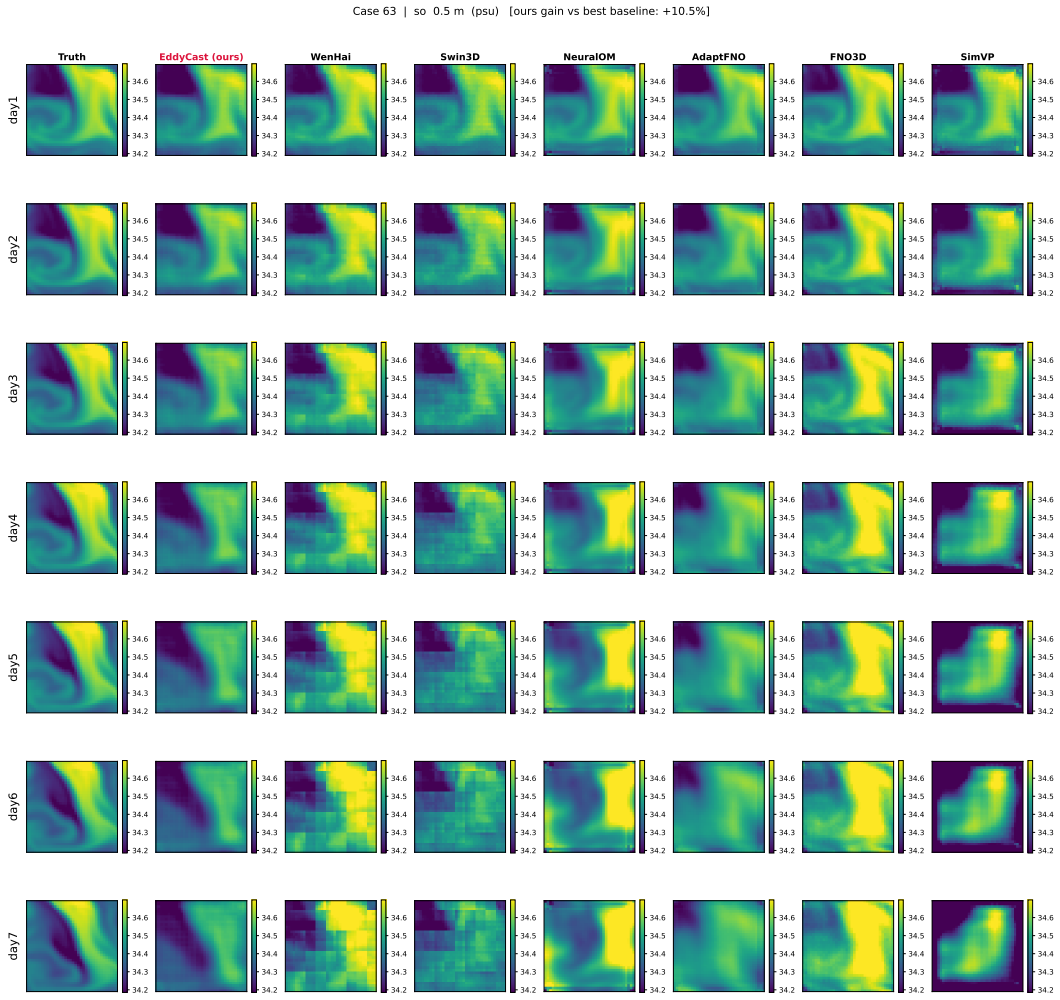


Figure 10: Qualitative comparison for test case 63 at 0.5 m depth (*so*, psu).

Table 6: Comparison of probabilistic denoising backbones and Gaussian-perturbation baseline on test set under identical sampling budget. All metrics are computed over normalized $[-1, 1]$ domain within ocean mask.

Model	Params	RMSE	MAE	SSIM	CRPS
(a) Denoising backbone					
Swin3D (EDM)	15.34M	0.0136	0.0088	0.9667	0.0061
UNet-EAG	3.02M	0.0142	0.0111	0.9762	0.0078
EddyCast-Prob	5.35M	**0.0117**	**0.0081**	**0.9792**	**0.0057**
(b) Gaussian perturbation baseline					
EddyCast-Gauss	2.98M	0.0129	0.0096	0.9460	0.0060

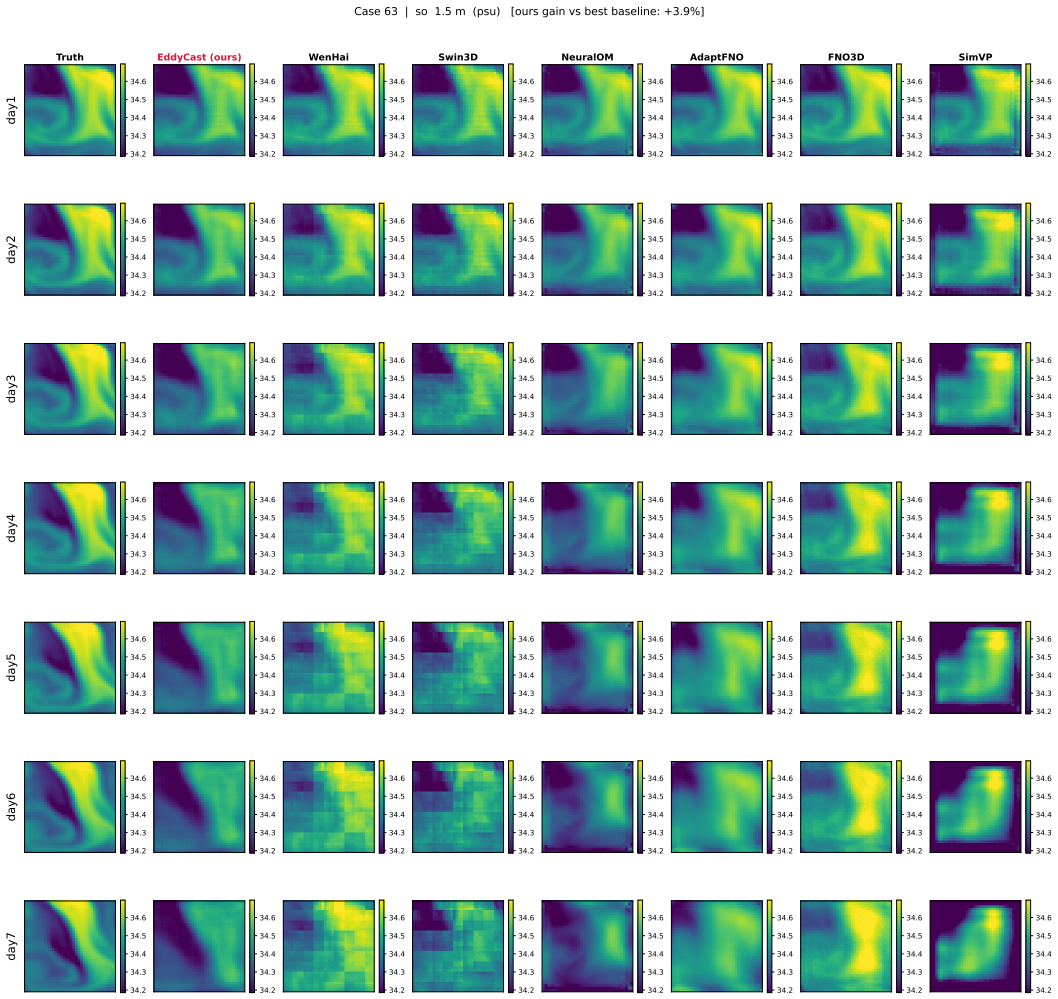


Figure 11: Qualitative comparison for test case 63 at 1.5 m depth (*so*, psu).

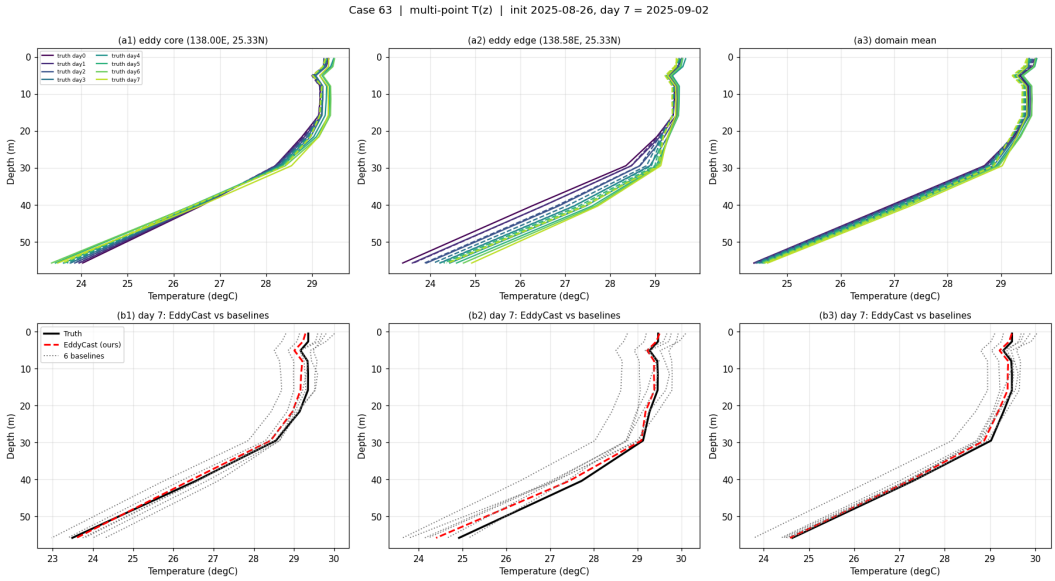


Figure 12: Vertical-profile analysis for test case 63 in the northwestern Pacific, covering 136.0°–138.9°E and 23.5°–26.4°N. The case is initialized on 26 August 2025, with forecasts from day 1 to day 7. The upper row shows the evolution of temperature profiles at the eddy core, eddy edge, and domain mean, while the lower row compares EddyCast with the ground truth and six deterministic baselines at day 7.