OCEANFLOW: COHERENT FLOW MATCHING FOR PROBABILISTIC OCEAN MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic ocean modeling is essential for risk-aware marine operations, climate analysis, and scientific decision-making, where downstream decisions depend on the conditional distribution of future ocean states rather than a single deterministic trajectory. Recent data-driven ocean models have made deterministic global ocean modeling increasingly accurate and efficient, but probabilistic ocean modeling remains underdeveloped. Specifically, perturbation-driven ensembles can create spread without explicitly optimizing the joint distribution of their members, often yielding poor calibration and physical inconsistency. To address these issues, we propose **`OceanFlow`**, a coherent flow matching framework for probabilistic global ocean modeling. **`OceanFlow`** first establishes a sharp, stable mean anchor for the conditional ensemble. Flow matching then learns a coherent state-conditioned pathwise law that jointly reshapes the residual distribution around this anchor at each transition, preserving the predictive center while refining state-dependent uncertainty. We further introduce a Factorized Dynamics Operator (FDO), which combines factorized transitions with a learned state-conditioned response to parameterize the mean anchor and residual-flow velocity. Extensive experiments show that **`OceanFlow`** improves ensemble-mean accuracy, probabilistic reliability, and physical-consistency diagnostics over competitive baselines, supporting coherent flow matching as an effective route toward data-driven probabilistic ocean modeling. Code is available at `https://anonymous.4open.science/r/Code_OceanFlow-25B9/`.

1 INTRODUCTION

Ocean modeling is essential for risk-aware marine operations, climate analysis, and scientific decision-making (Kalnay, 2003; Lellouche et al., 2018; Tan et al., 2025b). Incomplete observations leave uncertainty in the initial ocean state, while nonlinear dynamics can amplify small state errors into substantially different future evolutions (Lorenz, 2017; Leith, 1974). Even an accurate deterministic estimate cannot characterize the range of plausible outcomes or their relative likelihoods. Probabilistic modeling can address this limitation by representing the conditional distribution of future ocean states, providing uncertainty information beyond a single trajectory. Recent data-driven models, including AI-GOMS (Xiong et al., 2023), XiHe (Wang et al., 2024), LangYa (Yang et al., 2024), and WenHai (Cui et al., 2025), have substantially advanced deterministic global ocean modeling, combining competitive predictive accuracy with considerably faster inference than traditional numerical models. These gains make ocean modeling more affordable and provide a strong foundation for efficient ensemble generation. However, despite this progress in deterministic modeling, data-driven probabilistic ocean modeling remains comparatively underexplored.

Similar to initial-condition perturbation in numerical ensemble forecasting, data-driven ocean models can be extended to probabilistic modeling by perturbing their initial states and evolving the resulting members with a pretrained deterministic model (Hao et al., 2025). Prescribed perturbations can generate ensemble spread and improve ensemble-mean accuracy through averaging. However, members evolve independently without joint distributional optimization, potentially yielding distributions that deviate from physically plausible ocean variability and fail to faithfully represent uncertainty in the underlying dynamics. FuXi-ONS (Huang et al., 2026) advances beyond prescribed noise by learning state-dependent, spatially structured perturbations for autoregressive ocean prediction. This provides a more adaptive representation of ensemble variability. Nevertheless, its memberwise prediction

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.

objective leaves ensemble calibration implicit, while member updates do not explicitly account for the evolving ensemble context. These advances motivate a central question: ***how can the ensemble be jointly refined to more faithfully characterize uncertainty in the underlying ocean dynamics while maintaining a controlled predictive center?***

To answer this question, we propose `OceanFlow`, a coherent flow matching framework for probabilistic ocean modeling. As illustrated in Figure 1, `OceanFlow` learns a coherent state-conditioned pathwise law for the residual ensemble around a controlled mean anchor at each transition. This law couples member-specific transport paths through a shared velocity field conditioned on the ocean state and evolving ensemble context. Ensemble diversity is thus jointly refined as a conditional residual distribution, with final recentering preserving the predictive center. We further introduce a Factorized Dynamics Operator (FDO) that parameterizes both the mean anchor and residual-flow velocity through a unified composition of factorized dynamics and learned state-conditioned response. Our contributions can be summarized as:

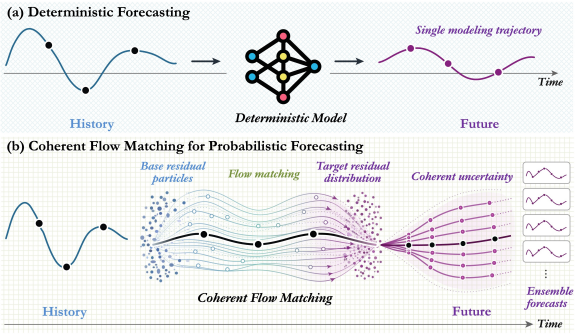

Figure 1: Learning coherent ensemble laws for ocean modeling. (a) Deterministic modeling produces a single trajectory. (b) `OceanFlow` represents ocean uncertainty through a mean-anchored ensemble, jointly refining member residuals under a coherent state-conditioned pathwise law at each transition.

- We propose `OceanFlow`, a coherent flow matching framework for probabilistic ocean modeling that learns a coherent state-conditioned pathwise law around the mean anchor center. By coupling member residuals through the evolving ensemble context, it jointly refines the conditional distribution while preserving a controlled mean anchor.
- We introduce a Factorized Dynamics Operator (FDO), which unifies factorized dynamics and a learned state-conditioned response within a single conditional operator formulation. FDO supports both an accurate deterministic reference for the mean anchor and a structured velocity field for residual transport.
- Extensive experiments on global ocean simulation and forecasting tasks demonstrate that the proposed framework achieves improved ensemble-mean accuracy, probabilistic reliability, and physical consistency.

2 Method

2.1 Problem Formulation

Let $\mathbf{O}_t \in \mathcal{X}$ denote the ocean state and $\mathbf{F}_t, \mathbf{F}_{t+1}$ the external forcing. Simulation uses observed future forcing, whereas prediction uses forcing generated by a separately trained OneForecast atmospheric model (Gao et al., 2025). Both settings follow the same conditional ocean transition. Given $c_t = (\mathbf{O}_t, \mathbf{F}_t, \mathbf{F}_{t+1})$, probabilistic modeling seeks an M-member conditional ensemble,

$$\widehat{\mathbf{O}}_{t+1}^{1:M} \sim Q_\theta^M(\cdot \mid c_t), \qquad Q_\theta^M(\cdot \mid c_t) \in \mathcal{P}(\mathcal{X}^M). \tag{1}$$

As depicted in Figure 2, `OceanFlow` constructs this distribution through a mean anchor and a coherent state-conditioned pathwise law for member anomalies. The anchor controls the predictive center, while ensemble-conditioned residual transport learns the distribution around it. A Factorized Dynamics Operator (FDO) provides a unified parameterization for the deterministic reference underlying the anchor and the residual-flow velocity.

2.2 Coherent Flow Matching

Coherent flow matching couples member residual paths through ensemble-conditioned dynamics while retaining explicit control of their common center. The law acts at each autoregressive transition in residual flow time τ: a deterministic reference advances the member conditions, and the learned transport determines the evaluated conditional ensemble.

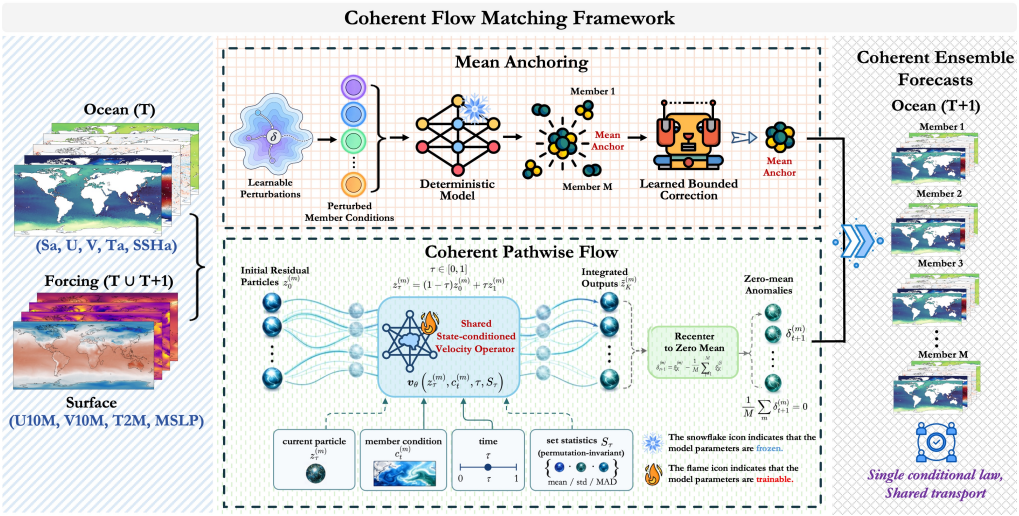

Figure 2: **Overview of `OceanFlow`.** The mean anchor controls the predictive center, while coherent state-conditioned pathwise flow jointly transports member residuals. Learned residuals refine the reference ensemble, while final recentering preserves the mean anchor. FDO parameterizes both the deterministic reference and residual velocity through a unified composition of factorized dynamics and learned state-conditioned response.

Mean anchoring. Mean anchoring separates central-state accuracy from residual uncertainty. A pretrained deterministic FDO instance f_{det} advances member conditions $c_t^{(m)}$, formed by perturbing the ocean state while sharing the same forcing. The reference ensemble can be expressed as:

$$x_{\text{base},t+1}^{(m)} = f_{\text{det}}(c_t^{(m)}), \qquad \mu_{t+1}^{\text{base}} = \frac{1}{M_b}\sum_{m=1}^{M_b} x_{\text{base},t+1}^{(m)}, \tag{2}$$

where M_b is the size of ensemble member and $M_b = M$ at evaluation. A learned correction adapts this reference center,

$$\mu_{t+1} = \mu_{t+1}^{\text{base}} + \mathcal{P}_o\left[\alpha_\mu \tanh\left(\frac{\mathcal{P}_o g_\psi(c_t, \mu_{t+1}^{\text{base}})}{\alpha_\mu}\right)\right], \tag{3}$$

where α_μ controls the correction scale and $\mathcal{P}_o$ removes each channel's spatial mean over valid ocean points. The resulting ensemble separates the mean anchor from zero-mean member anomalies,

$$\hat{x}_{t+1}^{(m)} = \mu_{t+1} + \delta_{t+1}^{(m)}, \qquad \frac{1}{M}\sum_{m=1}^{M} \delta_{t+1}^{(m)} = 0. \tag{4}$$

This decomposition allows the mean branch to learn a central correction without letting residual transport displace the anchor.

Coherent state-conditioned pathwise flow. The residual ensemble evolves under one velocity field whose dependence on symmetric ensemble summaries couples the member paths. Let $Z_0 = \{z_0^{(m)}\}_{m=1}^M$ be smooth normalized residual noise. The training target Z_1 combines a spread target derived from the verifying center error with zero-mean patterns of that error; its construction is given in Appendix E.4. For a shared interpolation time $\tau \sim \mathcal{U}(0,1)$,

$$z_\tau^{(m)} = (1-\tau)z_0^{(m)} + \tau z_1^{(m)}, \qquad u_\tau^{(m)} = z_1^{(m)} - z_0^{(m)}. \tag{5}$$

`OceanFlow` learns a shared residual velocity field v_θ by the flow matching objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{(c_t,y),\Xi,\tau}\left[\frac{1}{M}\sum_{m=1}^{M}\left\| v_\theta\left(z_\tau^{(m)}, c_t^{(m)}, \tau, S_\tau\right) - u_\tau^{(m)}\right\|_{2,o}^2\right], \tag{6}$$

where y is the verifying state, Ξ collects the stochastic member inputs, and $S_\tau = S(Z_\tau)$ contains the ensemble mean, standard deviation, and mean absolute deviation. These summaries are recomputed

from the joint residual state, making each member's velocity responsive to the ensemble context rather than to its own condition alone. Norms use the masked-grid convention in Appendix E.7.

Inference advances the residual ensemble with the same shared velocity,

$$z_{k+1}^{(m)} = \mathcal{P}_r\Big[z_k^{(m)} + \Delta\tau\, v_\theta\Big(z_k^{(m)}, c_t^{(m)}, \tau_k, S_k\Big)\Big], \qquad \tau_k = \big(k+\tfrac{1}{2}\big)\,\Delta\tau, \quad \Delta\tau = N_\tau^{-1}. \tag{7}$$

Here $\mathcal{P}_r$ applies the ocean mask and residual bound. The update defines a coherent member-set transport $\mathcal{T}_\theta^M$; shared parameters and symmetric context make it equivariant to jointly reordering particles and their conditions, as shown in Appendix E.5. Let $\tilde{z}_{N_\tau}^{(m)}$ denote the anomalies obtained by composing the reference-member anomalies with the transported residuals and applying spread calibration and stabilization. The final projection is

$$\delta_{t+1}^{(m)} = \tilde{z}_{N_\tau}^{(m)} - \frac{1}{M}\sum_{j=1}^{M} \tilde{z}_{N_\tau}^{(j)}, \tag{8}$$

which fixes the empirical center at μ_{t+1} for every generated ensemble. Coherence thus resides in the jointly conditioned residual transport, with mean anchoring enforcing the center of the resulting law. Reference composition and calibration are specified in Appendix E.4.

2.3 Factorized Dynamics Operator

The Factorized Dynamics Operator (FDO) is a general-purpose conditional transition operator for mean-anchor and residual-velocity parameterization. It expresses each transition through factorized dynamics composed with a state-conditioned response, giving both instances a unified operator structure.

For operator input $\mathbf{X}$ and context $\mathbf{C}$, a conditional patch encoder establishes the state on which factorized dynamics act,

$$H_0 = \mathcal{E}_\theta^{\text{patch}}(\mathbf{X}; \mathbf{C}). \tag{9}$$

The transition is accumulated through interacting dynamical factors,

$$H_{\ell+1} = H_\ell + \Phi_\ell\Big(H_\ell,\, \mathcal{A}_\ell\Big(\{\mathcal{T}_{\ell,k}\,(\mathcal{P}_{\ell,k}(H_\ell; \mathbf{C}))\}_{k=1}^{K};\, H_\ell, \mathbf{C}\Big); \mathbf{C}\Big), \qquad \ell = 0, \dots, L-1. \tag{10}$$

where $\mathcal{P}_{\ell,k}$ induces factor k, $\mathcal{T}_{\ell,k}$ evolves it, and $\mathcal{A}_\ell$ combines cross-factor interactions with state-dependent selection. The update map Φ_ℓ integrates this collective response into the patch state.

FDO maps this evolution to the target space through a state-conditioned response composition,

$$\widetilde{\mathbf{Y}} = \mathcal{R}_\theta(H_L), \qquad \mathcal{D}_\theta(\mathbf{X}; \mathbf{C}) = [\mathcal{I} + \gamma\mathcal{S}_\theta(\,\cdot \mid \mathbf{X})]\left(\widetilde{\mathbf{Y}}\right) = \widetilde{\mathbf{Y}} + \gamma\mathcal{S}_\theta\left(\widetilde{\mathbf{Y}}, \mathbf{X}\right). \tag{11}$$

Here $\mathcal{R}_\theta$ restores the output field, while $\mathcal{S}_\theta$ conditions its response on both the decoded dynamics and the operator input, with learned coefficient γ. The mean and velocity instances use separate parameters and task-specific inputs; their shared composition is detailed in Appendix E.6.

2.4 Training Losses and Calibration

Training first establishes a deterministic reference and then learns the coherent conditional ensemble around it. For normalized, masked ocean fields, the deterministic FDO instance $f_{\text{det},\vartheta}$ minimizes the channel-averaged relative L_2 risk,

$$\mathcal{L}_{\text{det}}(\vartheta) = \mathbb{E}_{(c_t, y)\sim\mathcal{D}_{\text{train}}}\left[\frac{1}{C_o}\sum_{q=1}^{C_o}\frac{\|M_{o,q}\odot(f_{\text{det},\vartheta}(c_t)_q - y_q)\|_2}{\|M_{o,q}\odot y_q\|_2}\right], \tag{12}$$

where C_o is the number of ocean channels and $M_{o,q}$ the valid-ocean mask for channel q. The pretrained reference is then frozen. With $\widehat{X} = \{\hat{x}^{(m)}\}_{m=1}^{M}$, the probabilistic stage optimizes the joint mean–residual objective

$$\mathcal{L}(\theta) = \mathbb{E}\Big[\lambda_{\text{fm}}\ell_{\text{fm}} + \lambda_{\text{mean}}\ell_{\text{mean}} + \lambda_{\text{prob}}\big(\mathsf{S}_M(\widehat{X}, y) + \eta\ell_{\text{gain}}\big) + \lambda_{\text{reg}}\ell_{\text{reg}}\Big]. \tag{13}$$

The expectation averages training transitions, stochastic member inputs, and flow time. Here ℓ_{fm} is the sample-level integrand of Eq. equation 6; ℓ_{mean} supervises the center and its improvement over the reference; and ℓ_{reg} controls member error and spread calibration. Distributional supervision uses empirical ensemble CRPS,

$$\mathsf{S}_M(\widehat{X}, y) = \frac{1}{M}\sum_m \|\hat{x}^{(m)} - y\|_{1,o} - \frac{1}{2M^2}\sum_{m,n}\|\hat{x}^{(m)} - \hat{x}^{(n)}\|_{1,o}, \tag{14}$$

where $\|\cdot\|_{1,o}$ denotes the masked-grid absolute average. The score balances member–target agreement with ensemble dispersion, while ℓ_{gain} penalizes CRPS exceeding that of the reference ensemble. Final recentering enforces the mean identity independently of optimization. Component definitions are given in Appendix E.7.

2.5 Theoretical Analysis

For transition context c and member conditions $C = (c^{(1)}, \ldots, c^{(M)})$, the continuous-time idealization of coherent transport evolves the residual ensemble through

$$\left[V_\theta^M(Z, C, \tau)\right]^{(m)} = v_\theta\Big(z^{(m)}, c^{(m)}, \tau, S(Z)\Big), \qquad \frac{\mathrm{d}Z_\tau}{\mathrm{d}\tau} = V_\theta^M(Z_\tau, C, \tau), \tag{15}$$

Let $\mathcal{T}_{\theta,C}^M$ denote either the time-one flow or the implemented finite-step transport, $\mathcal{K}_{\theta,c}^M$ the reference composition and calibration, and $\mathcal{R}^M$ the zero-mean projection. For stochastic input $\Xi = (Z_0, C) \sim \rho_0^M(\cdot \mid c)$, the resulting conditional ensemble law is

$$Q_\theta^M(\cdot \mid c) = \left(G_{\theta,c}^M\right)_{\#}\rho_0^M(\cdot \mid c), \qquad G_{\theta,c}^M(\Xi) = \mathbf{1}_M \otimes \mu_\theta(c, \Xi) + \mathcal{R}^M\left(\mathcal{K}_{\theta,c}^M\left(\mathcal{T}_{\theta,C}^M(Z_0), \Xi\right)\right). \tag{16}$$

Lemma 1 (Coherent conditional law)**.** *Under the assumptions in Appendix A, each ensemble-conditioned transport induces its own conditional law Q_θ^M with empirical mean $\mu_\theta(c, \Xi)$ for every stochastic input.*

3 Experiments

We evaluate `OceanFlow` on ocean simulation and forecasting tasks in terms of ensemble-mean accuracy, probabilistic reliability, ocean-specific diagnostics, physical consistency, and model complexity.

3.1 Experimental Setup

❶ **Dataset** We evaluate `OceanFlow` on GLORYS12 reanalysis with 93 ocean variables: salinity, temperature, and two velocity components at 23 depths, plus sea surface height. Four ERA5 forcing variables at two consecutive time steps provide eight atmospheric input channels. Data are processed at daily temporal and 1° spatial resolution, using daily-mean ocean states and forcings sampled at 12:00 UTC. We use 1993-2017 for training, 2018-2019 for validation, and 2020 for testing. Further data details refer to the Appendix.

❷ **Evaluation Metrics** We evaluate ensemble-mean skill with RMSE and ACC, and probabilistic quality with empirical ensemble CRPS, spread–skill ratio (SSR), rank histograms, and coverage calibration curves. RankDev and CalibErr quantify deviations from rank uniformity and nominal coverage, respectively. Metric definitions appear in Appendix C.

❸ **Baselines** We compare `OceanFlow` with a diverse set of baselines: the general spatiotemporal model ***SimVP*** (Tan et al., 2025a); Earth-system architectures ***GraphCast*** (Lam et al., 2023) and ***CirT*** (Liu et al., 2025); and ocean-specific models ***WenHai*** (Cui et al., 2025), ***NeuralOM*** (Gao et al., 2026), and ***DSLCast*** (Wu et al., 2026). We retrain SimVP, GraphCast, and CirT under the same protocol and use official pretrained weights for WenHai, NeuralOM, and DSLCast.

3.2 Probabilistic Ocean Simulation

Table 1 reports ocean simulation results over 20 initial conditions (ICs) under prescribed atmospheric forcing, separating ocean-state uncertainty from atmospheric forecasting error. `OceanFlow` achieves

Table 1: Quantitative ocean simulation results over 93 variables. Lower RMSE/CRPS, higher ACC, and SSR closer to 1 are better. Values are mean$_{\pm \text{std}}$; red/orange/blue mark the best/second-/third-best results. $*$ denotes Perlin-noise ensemble modeling for deterministic baselines.

Method	90-Day				120-Day			
	RMSE ↓	ACC ↑	CRPS ↓	SSR → 1	RMSE ↓	ACC ↑	CRPS ↓	SSR → 1
SimVP	0.932$_{\pm 0.005}$	0.088$_{\pm 0.006}$	—	—	0.932$_{\pm 0.012}$	0.058$_{\pm 0.005}$	—	—
SimVP*	0.811$_{\pm 0.006}$	0.104$_{\pm 0.005}$	0.485$_{\pm 0.004}$	0.582$_{\pm 0.012}$	0.807$_{\pm 0.010}$	0.073$_{\pm 0.006}$	0.489$_{\pm 0.007}$	0.602$_{\pm 0.015}$
GraphCast	0.921$_{\pm 0.005}$	0.198$_{\pm 0.003}$	—	—	1.031$_{\pm 0.004}$	0.139$_{\pm 0.003}$	—	—
GraphCast*	0.854$_{\pm 0.003}$	0.218$_{\pm 0.002}$	0.525$_{\pm 0.002}$	0.452$_{\pm 0.005}$	0.900$_{\pm 0.006}$	0.165$_{\pm 0.003}$	0.547$_{\pm 0.003}$	0.596$_{\pm 0.007}$
CirT	2.823$_{\pm 0.022}$	0.016$_{\pm 0.003}$	—	—	2.902$_{\pm 0.024}$	0.009$_{\pm 0.004}$	—	—
CirT*	2.415$_{\pm 0.036}$	0.019$_{\pm 0.002}$	1.087$_{\pm 0.019}$	0.624$_{\pm 0.026}$	2.455$_{\pm 0.036}$	0.010$_{\pm 0.005}$	1.127$_{\pm 0.018}$	0.648$_{\pm 0.029}$
WenHai	0.875$_{\pm 0.006}$	0.291$_{\pm 0.006}$	—	—	0.939$_{\pm 0.009}$	0.231$_{\pm 0.011}$	—	—
NeuralOM	0.682$_{\pm 0.009}$	0.393$_{\pm 0.009}$	—	—	0.710$_{\pm 0.005}$	0.319$_{\pm 0.007}$	—	—
DSLCast	0.880$_{\pm 0.007}$	0.333$_{\pm 0.009}$	—	—	0.915$_{\pm 0.009}$	0.294$_{\pm 0.013}$	—	—
DSLCast*	0.749$_{\pm 0.005}$	0.385$_{\pm 0.007}$	0.424$_{\pm 0.003}$	0.659$_{\pm 0.005}$	0.737$_{\pm 0.006}$	0.362$_{\pm 0.011}$	0.410$_{\pm 0.003}$	0.779$_{\pm 0.007}$
`OceanFlow`	0.681$_{\pm 0.005}$	0.421$_{\pm 0.005}$	0.370$_{\pm 0.002}$	0.707$_{\pm 0.004}$	0.670$_{\pm 0.003}$	0.400$_{\pm 0.007}$	0.368$_{\pm 0.002}$	0.784$_{\pm 0.005}$

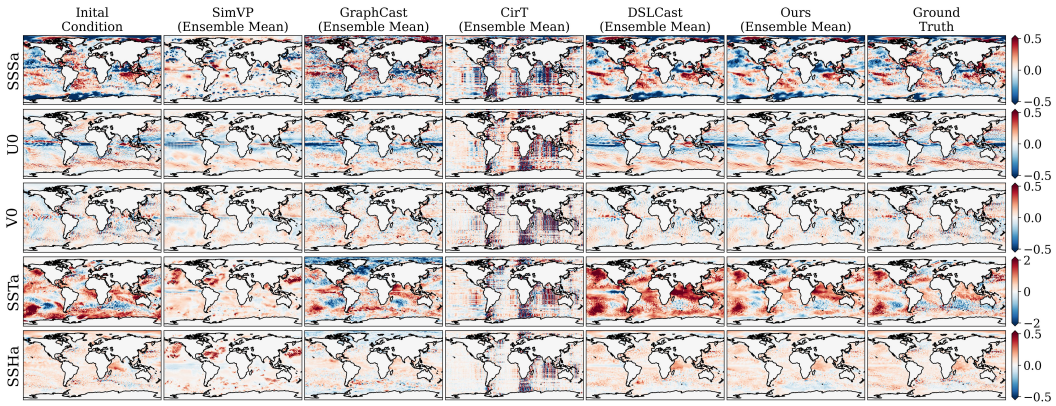

Figure 3: 120-day simulation results for several important surface variables. Our method achieves better physical consistency and yields results that are closest to the ground truth.

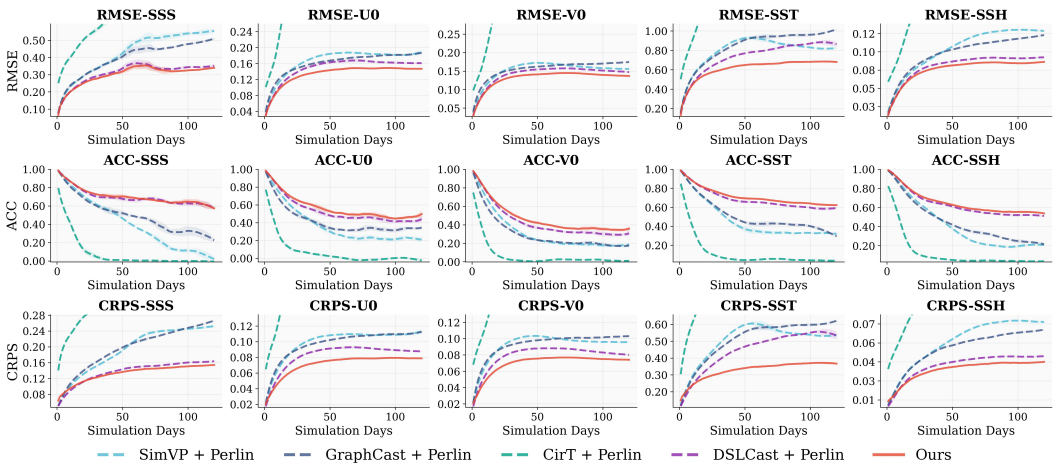

Figure 4: The RMSE, ACC and CRPS results of several important ocean surface variables.

the lowest ensemble-mean RMSE and highest ACC, together with the lowest CRPS, across all lead times. Although Perlin perturbations improve ensemble-mean RMSE over the corresponding deterministic trajectories, substantial gaps in ACC and CRPS remain relative to `OceanFlow`, particularly at longer lead times. This suggests that perturbation-induced spread alone may not adequately represent the state-dependent uncertainty of ocean dynamics. These results support the central premise of coherent flow matching: jointly refining the conditional residual distribution improves probabilistic quality while preserving an accurate ensemble center. Figure 4 shows that the gains extend across important sea-surface variables throughout the simulation range. The 120-day fields in Figure 3 provide

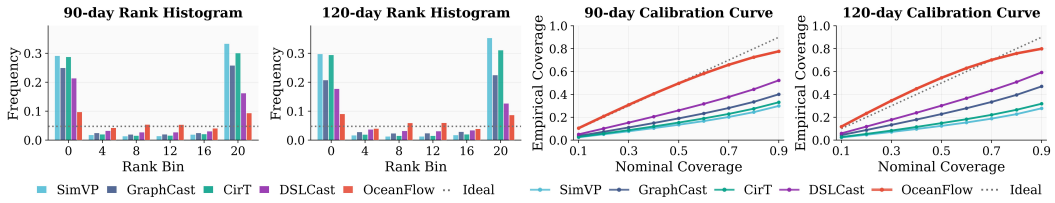

Figure 5: Rank histograms and calibration curves of ocean simulations.

Table 2: Quantitative reliability diagnostics for ocean simulation and forecasting. RankDev is computed as the L2 deviation using all 21 rank bins. Lower values are better.

Method	Simulation				Forecasting			
	90-Day		120-Day		20-Day		30-Day	
	RankDev↓	CalibErr↓	RankDev↓	CalibErr↓	RankDev↓	CalibErr↓	RankDev↓	CalibErr↓
SimVP*	0.396±0.005	0.355±0.003	0.416±0.006	0.366±0.004	0.328±0.002	0.292±0.002	0.338±0.002	0.307±0.002
GraphCast*	0.309±0.002	0.299±0.001	0.254±0.002	0.261±0.002	0.326±0.002	0.294±0.001	0.352±0.002	0.317±0.001
CirT*	0.368±0.010	0.337±0.008	0.381±0.009	0.343±0.009	0.426±0.004	0.372±0.002	0.390±0.003	0.351±0.002
DSLCast*	0.215±0.002	0.230±0.002	0.163±0.003	0.189±0.002	0.288±0.003	0.261±0.003	0.306±0.002	0.281±0.002
`OceanFlow`	0.088±0.001	0.032±0.001	0.093±0.001	0.041±0.001	0.079±0.002	0.104±0.003	0.106±0.002	0.124±0.003

Table 3: Probabilistic ocean forecasting results for 93 ocean variables.

Method	20-Day				30-Day			
	RMSE ↓	ACC ↑	CRPS ↓	SSR → 1	RMSE ↓	ACC ↑	CRPS ↓	SSR → 1
SimVP*	0.667±0.006	0.602±0.004	0.408±0.003	0.221±0.002	0.774±0.008	0.465±0.007	0.475±0.004	0.271±0.005
GraphCast*	0.684±0.005	0.523±0.005	0.414±0.003	0.258±0.002	0.749±0.004	0.395±0.005	0.461±0.002	0.269±0.004
CirT*	1.640±0.020	0.096±0.003	0.787±0.012	0.354±0.008	1.984±0.037	0.044±0.003	0.903±0.016	0.434±0.012
DSLCast*	0.624±0.008	0.641±0.008	0.360±0.004	0.324±0.005	0.712±0.008	0.521±0.007	0.416±0.004	0.357±0.006
`OceanFlow`	0.589±0.007	0.660±0.006	0.311±0.004	0.514±0.006	0.661±0.008	0.550±0.007	0.357±0.003	0.483±0.007

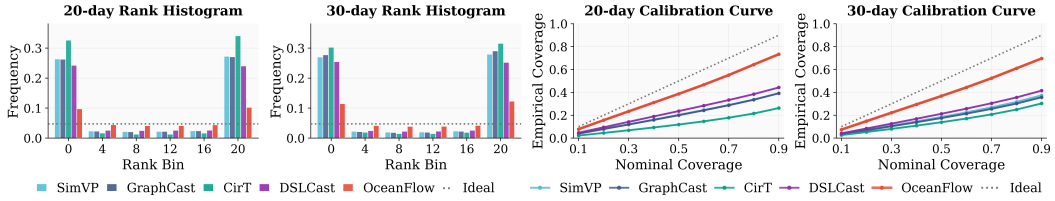

Figure 6: Rank histograms and calibration curves of probabilistic ocean forecasting.

a complementary spatial comparison: the ensemble mean of `OceanFlow` remains sharp and spatially organized, while several baselines lose fine-scale structure or exhibit pronounced spatial distortions. Figure 5 and Table 2 further assess ensemble reliability. `OceanFlow` yields rank distributions closer to uniform and empirical coverage closer to nominal levels, with the lowest RankDev and CalibErr at all simulation horizons.

3.3 Probabilistic Ocean Forecasting

To assess `OceanFlow`'s real-world application ability, we drive the ocean model with atmospheric forecasts produced by OneForecast (Gao et al., 2025). OneForecast is initialized from the atmospheric state paired with each ocean IC and rolled out at 6-hour resolution to provide forcings. The results in Table 3, averaged over 20 ICs, show that `OceanFlow` improves both ensemble-mean skill and CRPS at 20-day and 30-day lead times. The reliability gains observed in simulation persist under predicted forcing: Figure 6 and Table 2 (right) show more uniform rank histograms and better-calibrated uncertainty estimates than the perturbation-based baselines.

3.4 Generalization and ENSO Diagnosis

We evaluate out-of-sample generalization in 2024 using 10 ICs. Global metrics are averaged over 93 variables, while ENSO metrics are computed from the Niño-3.4 SST anomaly index over the

Table 4: Out-of-sample 2024 simulation and ENSO diagnosis. The best results are shown in **bold**.

Lead	Model	Global ocean simulation						Niño-3.4 diagnosis	
		RMSE↓	ACC↑	CRPS↓	RankDev↓	SSR→ 1	CalibErr↓	RMSE↓	CRPS↓
90d	DSLCast*	0.784±0.003	0.399±0.004	0.449±0.001	0.216±0.001	0.665±0.004	0.232±0.001	0.6181	0.5007
90d	`OceanFlow`	**0.698**±0.003	**0.430**±0.003	**0.383**±0.001	**0.095**±0.001	**0.744**±0.001	**0.034**±0.000	**0.1451**	**0.0901**
120d	DSLCast*	0.779±0.003	0.377±0.003	0.435±0.001	0.159±0.001	0.790±0.005	0.188±0.001	0.7977	0.6879
120d	`OceanFlow`	**0.686**±0.003	**0.416**±0.003	**0.382**±0.001	**0.102**±0.001	**0.836**±0.005	**0.054**±0.001	**0.2304**	**0.1496**

Table 5: Out-of-sample 2024 probabilistic ocean forecasting and ENSO prediction results.

Model	Global ocean forecasting						Niño-3.4 diagnosis	
	RMSE↓	ACC↑	CRPS↓	RankDev↓	SSR→ 1	CalibErr↓	RMSE↓	CRPS↓
DSLCast* one-way	0.867±0.027	0.258±0.028	0.495±0.013	0.245±0.009	0.605±0.020	0.253±0.006	0.9736	0.7933
`OceanFlow` one-way	0.796±0.027	0.257±0.033	0.436±0.012	0.134±0.005	0.678±0.019	0.057±0.005	0.9118	0.7173
`OceanFlow` two-way	**0.762**±0.014	**0.327**±0.013	**0.419**±0.006	**0.116**±0.002	**0.699**±0.009	**0.045**±0.001	**0.6898**	**0.4643**

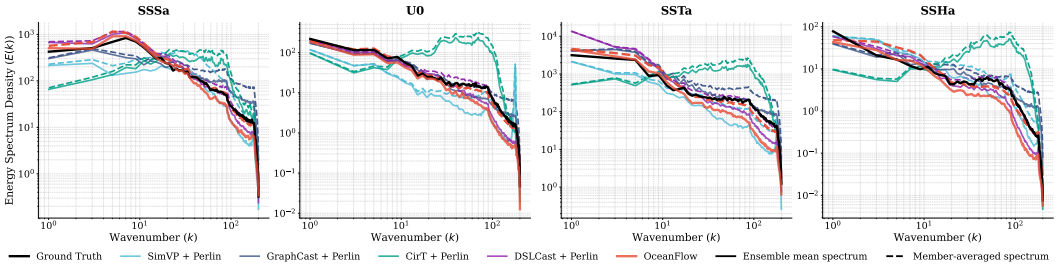

Figure 7: Spectral fidelity of ensemble mean and member-averaged 120-day ocean simulations

same ICs. We compare against DSLCast+Perlin, the strongest perturbation-based ensemble baseline in the main simulation comparison. Table 4 shows improvements across global simulation metrics and Niño-3.4 RMSE and CRPS at both 90-day and 120-day horizons. These gains extend beyond the 2020 test year, supporting generalization in both global ocean states and large-scale tropical Pacific variability. As a preliminary extension toward coupled forecasting, we drive `OceanFlow` with atmospheric states generated by an independently running two-way PnP-Corrector system (Wu et al., 2026). At the 90-day horizon, this configuration improves all reported global and Niño-3.4 metrics relative to the one-way forcing configuration. These findings suggest that end-to-end probabilistic two-way coupling could further improve ocean forecasting.

3.5 Analysis of Physical Consistency

Aggregate scores alone do not fully reveal whether the generated ensemble members are physically plausible. Moreover, in a well-formed ensemble system, the ensemble mean should not be interpreted as a single realized ocean state: as lead time increases, averaging over plausible but phase-shifted members can naturally smooth unpredictable high-wavenumber components(Leith, 1974). We therefore separate ensemble-mean structure from member-level realism in Figure 7. For readability, we visualize representative surface variables covering typical near-surface dynamical and thermodynamic ocean processes. Solid lines show spectra of the ensemble mean, while dashed lines show member-averaged spectra computed by first evaluating each of the 20 ensemble members separately and then averaging their spectra over members and 20 ICs. The member-averaged spectra of `OceanFlow` are overall closer to the ground truth across these variables, indicating that individual ensemble members preserve realistic multiscale ocean variability. Meanwhile, the ensemble-mean spectra show larger deviations from the ground truth at high wavenumbers. This behavior is consistent with a well-calibrated ensemble, whose mean averages out unpredictable spectral components rather than reproducing the full spectrum of a single realization. In addition to the spectral comparison, Figure 8 reports three targeted physical diagnostics at the 120-day simulation horizon, covering SSTa mean tendency consistency, SSH-current multi-

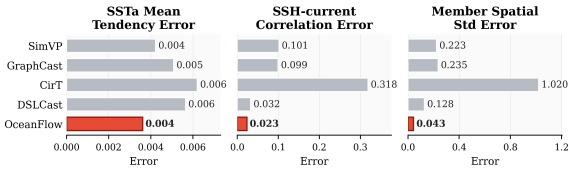

Figure 8: Physical diagnostics of the generated ensemble members and ensemble mean.

Table 7: Ablation results of OceanFlow. The best results are shown in **bold**.

Variants	RMSE↓	ACC↑	CRPS↓	RankDev↓	SSR→ 1	CalibErr↓
w/o Factorized Dynamics	1.054±0.008	0.188±0.009	—	—	—	—
w/o State Response	0.934±0.062	0.072±0.090	—	—	—	—
w/o Adaptive State Response	0.844±0.007	0.322±0.010	—	—	—	—
FDO	0.825±0.009	0.352±0.007	—	—	—	—
FDO+Perlin	0.715±0.005	0.406±0.003	0.403±0.002	0.204±0.001	0.625±0.004	0.225±0.001
w/o mean correction/re-centering	0.692±0.005	0.420±0.003	0.376±0.002	0.101±0.001	0.690±0.003	0.059±0.001
w/o learned pathwise flow	0.697±0.004	0.417±0.002	0.379±0.002	0.109±0.001	0.665±0.004	0.081±0.002
`OceanFlow`	**0.684**±0.005	**0.425**±0.002	**0.371**±0.002	**0.088**±0.000	**0.708**±0.003	**0.032**±0.000

variate consistency, and member-level spatial variability. `OceanFlow` achieves the lowest error on all three diagnostics, further indicating that it better preserves large-scale thermal evolution, dynamical SSH-current relationships, and realistic spatial variability of individual ensemble members.

3.6 Ablation Study

To validate the architectural design of `OceanFlow`, we conduct ablation studies through 90-day simulations over 10 ICs, as reported in Table 7. All variants are trained same as the main method. The deterministic variants are compared against FDO: (1) *w/o Factorized Dynamics* replaces the factorized dynamics block with a plain convolutional block, substantially degrading simulation skill. (2) *w/o State Response* removes the state-conditioned response, leading to poorer accuracy and highlighting its contribution to the factorized transition. (3) *w/o Adaptive State Response* substitutes a fixed persistence addition for the learned response, partially recovering performance but remaining inferior to FDO. This supports adapting the state response rather than prescribing a fixed additive connection. The probabilistic comparisons further isolate the role of coherent residual transport: (4) *FDO+Perlin* uses the same deterministic backbone with prescribed perturbations but yields poorer probabilistic scores than `OceanFlow`, showing that the gains cannot be attributed to the backbone alone. (5) *w/o mean correction/re-centering* degrades both ensemble-mean accuracy and reliability, supporting explicit control of the predictive center. (6) *w/o learned pathwise flow* also reduces probabilistic performance, demonstrating the benefit of learned residual transport beyond the perturbation reference. Together, these results support the complementary roles of FDO, mean anchoring, and coherent residual transport in producing accurate and reliable ocean ensembles.

3.7 Training Protocol and Efficiency Analysis

Table 6: Parameter comparison for retrained models.

Models	SimVP	GraphCast	CirT	OceanFlow
Params (M)	78.54	40.82	61.56	37.45

We train SimVP, GraphCast, CirT, and the deterministic FDO backbone for 50 epochs on 8 NVIDIA A100 GPUs under the same framework. The learning rate starts at 10^{-3} and decays to zero with a cosine schedule. For `OceanFlow`, residual transport is trained for 3 additional epochs with 10 residual particles per sample in the flow matching loss; the perturbation-reference mean is estimated from 20 members. Table 6 restricts the parameter comparison to models retrained under the common protocol; official pretrained baselines use different training configurations, `OceanFlow` has the fewest trainable parameters.

4 Conclusion and Limitations

We present `OceanFlow`, a coherent flow matching framework for probabilistic ocean modeling. By coupling member residuals through the evolving ensemble context, `OceanFlow` learns a coherent state-conditioned pathwise law that refines ensemble diversity around a controlled mean anchor. FDO parameterizes both the mean anchor and residual-flow velocity through a unified composition of factorized dynamics and learned state-conditioned response. Experiments demonstrate improved ensemble-mean accuracy, probabilistic reliability, and physical-consistency diagnostics, supporting coherent flow matching as an effective approach to ocean uncertainty modeling. Current ocean forecasting relies on the forcings produced by external model, extending the framework to end-to-end probabilistic ocean-atmosphere coupling may further improve ocean forecasting accuracy.

AI USE STATEMENT

We use AI to check for grammatical errors in the text; the AI is not involved in ideas, experimental design, or the actual writing of the paper.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Human evaluations were conducted with informed consent from volunteers. All datasets and pretrained models are publicly available and used under their respective licenses. We have carefully designed our experiments to minimize risks of bias or discrimination. We uphold the principles of transparency and integrity throughout the work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document data sources, preprocessing, and splits in Appendix D, and provide architectural details, training objectives, and inference procedures in Appendix E. Evaluation metrics and the theoretical proof are presented in Appendices C and A, respectively. Training settings and hardware are reported in the main text.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Hassan Al Sukhni, Fatma Sakr, Mutasem K Alsmadi, Magdy Abd-Elghany, Ibrahim A Gomaa, Shaimaa Abdallah, et al. Data-driven weather prediction: Integrating deep learning and ensemble models for robust weather forecasting. *Journal of Cybersecurity & Information Management*, 15 (2), 2025.

Roberto Buizza, PL Houtekamer, Gerald Pellerin, Zoltan Toth, Yuejian Zhu, and Mozheng Wei. A comparison of the ecmwf, msc, and ncep global ensemble prediction systems. *Monthly Weather Review*, 133(5):1076–1097, 2005.

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.

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.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5): 10143–10162, 1994.

GAEL Forget, J-M Campin, P Heimbach, CN Hill, RM Ponte, and C Wunsch. Ecco version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, 8(10):3071–3104, 2015.

Yuan Gao, Hao Wu, Ruiqi Shu, huanshuo dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, Jiahao Wu, Li Qing, Hui Xiong, and Xiaomeng Huang. Oneforecast: A universal framework for global and regional weather forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9xGSeVolcN.

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*, volume 40, pp. 14756–14764, 2026.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Rixu Hao, Yuxin Zhao, Shaoqing Zhang, and Xiong Deng. Deep learning for ocean forecasting: A comprehensive review of methods, applications, and datasets. *IEEE Transactions on Cybernetics*, 2025.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

Peter L Houtekamer and Herschel L Mitchell. Data assimilation using an ensemble kalman filter technique. *Monthly weather review*, 126(3):796–811, 1998.

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.

Haoming Jia, Yi Han, Xiang Wang, Huizan Wang, Wei Wu, Jianming Zheng, and Peikun Xiao. Oceanforecastbench: A benchmark dataset for data-driven global ocean forecasting. *arXiv preprint arXiv:2511.18732*, 2025.

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

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.

Erik Larsson, Joel Oskarsson, Tomas Landelius, and Fredrik Lindsten. Crps-lam: Regional ensemble weather forecasting from matching marginals. *arXiv preprint arXiv:2510.09484*, 2025.

Cecil E Leith. Theoretical skill of monte carlo forecasts. *Monthly weather review*, 102(6):409–418, 1974.

Jean-Michel Lellouche, Eric Greiner, Olivier Le Galloudec, Gilles Garric, Charly Regnier, Marie Drevillon, Mounir Benkiran, Charles-Emmanuel Testut, Romain Bourdalle-Badie, Florent Gasparin, et al. Recent updates to the copernicus marine service global ocean monitoring and forecasting real-time 1/ 12 high-resolution system. *Ocean Science*, 14(5):1093–1126, 2018.

Yang Liu, Zinan Zheng, Jiashun Cheng, Fugee Tsung, Deli Zhao, Yu Rong, and Jia Li. Cirt: Global subseasonal-to-seasonal forecasting with geometry-inspired transformer. *arXiv preprint arXiv:2502.19750*, 2025.

Edward N Lorenz. Deterministic nonperiodic flow 1. In *Universality in Chaos, 2nd edition*, pp. 367–378. Routledge, 2017.

Gurvan Madec et al. Nemo ocean engine. 2015.

James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529):73–119, 1996.

Congyi Nai, Xi Chen, Shangshang Yang, Ziniu Xiao, and Baoxiang Pan. Boosting weather forecast via generative superensemble. *npj Climate and Atmospheric Science*, 8(1):377, 2025.

Tim Palmer. The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.

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

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.

Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z Li. Simvpv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Transactions on Multimedia*, 2025a.

Xianchun Tan, Zhe Peng, Yonglong Cheng, Yi Wang, Qingchen Chao, Xiaomeng Huang, Hongshuo Yan, and Deliang Chen. Leveraging artificial intelligence for research and action on climate change: opportunities, challenges, and future directions. *Science Bulletin*, 2025b.

Zoltan Toth and Eugenia Kalnay. Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the american meteorological society*, 74(12):2317–2330, 1993.

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.

Hao Wu, Yuan Gao, Xingjian Shi, Shuaipeng Li, Fan Xu, Fan Zhang, Zhihong Zhu, Weiyan Wang, Xiao Luo, Kun Wang, et al. Spatiotemporal forecasting as planning: A model-based reinforcement learning approach with generative world models. *arXiv preprint arXiv:2510.04020*, 2025.

Hao Wu, Fan Xu, Yuxu Lu, Penghao Zhao, Fan Zhang, Hao Jia, Yuxuan Liang, Ruijian Gou, Qingsong Wen, Xian Wu, Xiaomeng Huang, and Yuan Gao. Pnp-corrector: A universal correction framework for coupled spatiotemporal forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IYtXaU7N6B.

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

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

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Xiaohui Zhong, Lei Chen, Hao Li, Roberto Buizza, Jun Liu, Jie Feng, Zijian Zhu, Xu Fan, Kan Dai, Jing-jia Luo, et al. Fuxi-ens: A machine learning model for efficient and accurate ensemble weather prediction. *Science Advances*, 11(44):eadu2854, 2025.

APPENDIX CONTENTS

A Proof of the Coherent Conditional Law **14**

B Related Work **15**

C Evaluation Metrics **15**

D Data Details **18**

D.1 Dataset 18

D.2 Data Preprocessing 18

E Detailed Method **19**

E.1 Masked Normalization and Member Conditions 19

E.2 Structured Perturbation Reference 19

E.3 Mean-Anchored Residual Decomposition 20

E.4 Coherent Residual Transport 20

E.5 Ensemble-Conditioned Transport 21

E.6 Factorized Dynamics Operator 21

E.7 Training Objectives 22

E.8 Reference-Guided Autoregressive Inference 23

E.9 Ablation Design 24

F Additional Results **24**

A Proof of the Coherent Conditional Law

Proof of Lemma 1. Fix trained parameters and ensemble size M, with $\mathcal{X} = \mathbb{R}^d$ denoting valid ocean-grid coordinates. State and context spaces carry their Borel sigma-algebras, and $\rho_0^M(\cdot \mid c)$ is a probability kernel. Assume joint measurability of the velocity, anchor, reference composition, calibration, and context encoding. For continuous transport, assume continuity in τ and global Lipschitz continuity in Z, uniformly on $[0,1]$ with constant $L < \infty$ for fixed (c, C), including the shared context. With $\|Z\|_M^2 = M^{-1} \sum_m \|z^{(m)}\|_2^2$, the flow satisfies

$$\begin{aligned} \left\| V_\theta^M(Z, C, \tau) - V_\theta^M(Z', C, \tau) \right\|_M &\le L \, \|Z - Z'\|_M \,, \\ \left\| \Phi_{\theta,C,\tau}^M(Z_0) - \Phi_{\theta,C,\tau}^M(Z_0') \right\|_M &\le e^{L\tau} \, \|Z_0 - Z_0'\|_M \,. \end{aligned} \tag{17}$$

Picard–Lindelöf gives existence and uniqueness; Grönwall bounds sensitivity in flow time τ, not autoregressive lead time. For initial law $\nu_0^M(\cdot \mid c, C)$, the joint residual law obeys the weak continuity equation,

$$\rho_{\theta,\tau}^M(\cdot \mid c, C) = \left(\Phi_{\theta,C,\tau}^M\right)_{\#} \nu_0^M(\cdot \mid c, C), \qquad \partial_\tau \rho_{\theta,\tau}^M + \nabla_Z \cdot \left(\rho_{\theta,\tau}^M V_\theta^M\right) = 0. \tag{18}$$

All member paths are components of this joint flow. For $m \neq j$, the chain rule at differentiable points and fixed (c, C) gives

$$D_{z^{(j)}} \left[V_\theta^M(Z, C, \tau)\right]^{(m)} = D_S v_\theta \, D_{z^{(j)}} S(Z). \tag{19}$$

A nonzero cross derivative makes the velocity locally nonseparable: $S(Z)$ enables cross-member interaction beyond parameter sharing, although its learned strength need not be nonzero. The mean, standard deviation, and mean absolute deviation remain Lipschitz even at nondifferentiable points.

Shared parameters and symmetric summaries give $V_\theta^M(\Pi Z, \Pi C, \tau) = \Pi V_\theta^M(Z, C, \tau)$; uniqueness transfers this joint-relabeling symmetry to the flow.

Center control follows from the projection itself. For $U = (u^{(1)}, \ldots, u^{(M)})$ and $\bar{u} = M^{-1} \sum_m u^{(m)}$,

$$\begin{aligned} \mathcal{R}^M(U)^{(m)} &= u^{(m)} - \bar{u}, \qquad \frac{1}{M} \sum_{m=1}^M \left[\mathcal{R}^M(U)\right]^{(m)} = 0, \\ \left[\mathcal{R}^M(U)\right]^{(m)} - \left[\mathcal{R}^M(U)\right]^{(j)} &= u^{(m)} - u^{(j)}. \end{aligned} \tag{20}$$

Centering therefore preserves the pairwise contrasts of its input. Taking $U = \mathcal{K}_{\theta,c}^M(\mathcal{T}_{\theta,C}^M(Z_0), \Xi)$ after reference composition, calibration, and stabilization yields

$$\frac{1}{M} \sum_{m=1}^M \left[G_{\theta,c}^M(\Xi)\right]^{(m)} = \mu_\theta(c, \Xi) + \frac{1}{M} \sum_{m=1}^M \left[\mathcal{R}^M(U)\right]^{(m)} = \mu_\theta(c, \Xi). \tag{21}$$

The identity holds for every Ξ, even when the anchor varies with the stochastic reference. Joint measurability makes $G_{\theta,c}^M$ measurable in (c, Ξ), so its pushforward of $\rho_0^M(\cdot \mid c)$ is a probability kernel.

For continuous transport, $\mathcal{T}_{\theta,C}^M = \Phi_{\theta,C,1}^M$. Each implemented update in Eq. (7), including masking and clipping, is measurable and permutation equivariant. Finite composition therefore induces its own conditional law with the same mean identity, without ODE regularity or equality to the continuous law.

Finally, condition on a reference rollout r and its forcing, with remaining stochastic inputs Ω. Composing the transition-wise generators and context encodings defines $\mathcal{G}_{\theta,r}^M$ with

$$Q_{\theta,0:T}^M(\cdot \mid r) = \left(\mathcal{G}_{\theta,r}^M\right)_{\#} \rho_\Omega^M(\cdot \mid r), \qquad \frac{1}{M} \sum_{m=1}^M \widehat{\mathbf{O}}_t^{(m)} = \mu_{\theta,t}, \quad t = 1, \ldots, T. \tag{22}$$

This law describes reference-guided ensembles, not flow-driven autoregressive particles. Coherence concerns jointly conditioned residual paths and exact center control at each transition; the result neither requires persistent memory nor establishes calibration or long-horizon skill. □

B Related Work

Deterministic ocean modeling. Numerical circulation models such as HYCOM (Chassignet et al., 2007) and NEMO (Madec et al., 2015), together with state-estimation and reanalysis systems such as ECCO (Forget et al., 2015) and Copernicus global ocean products (Lellouche et al., 2018), underpin modern ocean modeling. Their physical foundations support multivariate state estimation and prediction, but high-resolution integrations remain computationally demanding. Data-driven alternatives offer efficient representations of ocean evolution. AI-GOMS (Xiong et al., 2023) develops a reusable ocean modeling backbone, while XiHe (Wang et al., 2024) and WenHai (Cui et al., 2025) demonstrate eddy-resolving global prediction. LangYa (Yang et al., 2024) addresses cross-spatiotemporal prediction, and OceanForecastBench (Jia et al., 2025) provides a basis for systematic model comparison. These advances establish strong deterministic references but do not, by themselves, specify calibrated predictive distributions. In `OceanFlow`, FDO composes factorized dynamics with a learned state-conditioned response to support both the deterministic reference and residual velocity. Coherent flow matching then makes the conditional residual ensemble, rather than only a central estimate, the object of learning.

Probabilistic models in the Earth system. Ensemble prediction (Lorenz, 2017; Leith, 1974; Al Sukhni et al., 2025; Wu et al., 2025) characterizes uncertainty through multiple plausible future states. Operational systems use perturbed initial conditions (Toth & Kalnay, 1993; Molteni et al., 1996), ensemble data assimilation (Evensen, 1994; Houtekamer & Mitchell, 1998), and multi-model diversity (Buizza et al., 2005; Palmer, 2019) to approximate predictive distributions. Data-driven approaches (Larsson et al., 2025; Nai et al., 2025) include precipitation modeling with MetNet (Sønderby et al., 2020), generative radar nowcasting (Ravuri et al., 2021), diffusion-based global weather ensembles with GenCast (Price et al., 2023), and learned ensemble systems such as FuXi-ENS (Zhong et al., 2025). DiffCast (Yu et al., 2024) combines a deterministic backbone with residual diffusion for precipitation nowcasting. Extending probabilistic modeling to the ocean requires accounting for subsurface stratification, variable- and depth-dependent uncertainty, complex land–sea geometry, and long dynamical memory. FuXi-ONS (Huang et al., 2026) learns state-dependent structured perturbations and conditions autoregressive ocean prediction on atmospheric features. Its reported training objective combines member-level prediction errors with atmospheric regularization. `OceanFlow` learns a coherent state-conditioned pathwise law for the residual ensemble around a controlled mean anchor. Member updates depend on the evolving ensemble context, and probabilistic objectives supervise the transported ensemble. The distinction is this joint, center-controlled distributional refinement at each transition, rather than perturbation design or residual generation alone.

C Evaluation Metrics

Evaluation distinguishes the accuracy of the ensemble center from the quality of the distribution around it. RMSE and ACC assess the ensemble mean; CRPS, RankDev, SSR, and CalibErr assess complementary aspects of the ensemble distribution. Ocean-specific diagnostics further examine regional variability, multivariate consistency, and member-level structure. All scalar scores use the ocean mask. Let x and y denote a predicted and target ocean field, let c index the evaluated ocean channels, and let (i, j) index latitude and longitude. The latitude weight is

$$w_i = \frac{H \cos(\phi_i)}{\sum_{r=1}^{H} \cos(\phi_r)}, \tag{23}$$

where H is the number of evaluated latitudes and ϕ_i is the latitude at row i. After applying the ocean mask, land points are treated as missing values and the spatial averages below are computed over valid ocean points. Unless specified otherwise, global probabilistic metrics are evaluated in normalized variable space, and numerical denominators use $\epsilon = 10^{-12}$.

RMSE. RMSE is computed on normalized ocean variables and then averaged over channels:

$$\mathrm{RMSE} = \frac{1}{C} \sum_{c=1}^{C} \sqrt{\mathrm{nanmean}_{i,j} \left[w_i \left(\hat{x}_{c,i,j} - y_{c,i,j} \right)^2 \right]}. \tag{24}$$

ACC. ACC is computed on anomalies relative to the climatological mean. Let $\hat{x}' = \hat{x} - x^{\text{clim}}$ and $y' = y - y^{\text{clim}}$. The channel-wise anomaly correlation is

$$\text{ACC}_c = \frac{\sum_{i,j} w_i \hat{x}'_{c,i,j} y'_{c,i,j}}{\sqrt{\left(\sum_{i,j} w_i (\hat{x}'_{c,i,j})^2\right)\left(\sum_{i,j} w_i (y'_{c,i,j})^2\right)}}, \qquad \text{ACC} = \frac{1}{C}\sum_{c=1}^{C} \text{ACC}_c. \tag{25}$$

CRPS. For an ensemble $\{\hat{x}^{(m)}\}_{m=1}^{M}$, we use the empirical ensemble form (Matheson & Winkler, 1976; Hersbach, 2000; Gneiting & Raftery, 2007),

$$\text{CRPS} = \frac{1}{C}\sum_{c=1}^{C} \text{nanmean}_{i,j}\left[w_i \left(\frac{1}{M}\sum_{m=1}^{M} \left| \hat{x}^{(m)}_{c,i,j} - y_{c,i,j} \right| - \frac{1}{2M^2} \sum_{m,n=1}^{M} \left| \hat{x}^{(m)}_{c,i,j} - \hat{x}^{(n)}_{c,i,j} \right| \right) \right]. \tag{26}$$

Rank deviation (RankDev). At every channel–grid point where the target and all M members are valid, the target rank is

$$r_{c,i,j} = \sum_{m=1}^{M} \mathbf{1}\left\{ \hat{x}^{(m)}_{c,i,j} < y_{c,i,j} \right\} \in \{0, \dots, M\}. \tag{27}$$

Let $\mathcal{V}$ denote the set of valid channel–grid points. The latitude-weighted rank histogram and its deviation from the uniform histogram are

$$h_b = \frac{\sum_{(c,i,j)\in\mathcal{V}} w_i \mathbf{1}\{r_{c,i,j} = b\}}{\sum_{(c,i,j)\in\mathcal{V}} w_i + \epsilon}, \qquad \text{RankDev} = \left[\sum_{b=0}^{M} \left(h_b - \frac{1}{M+1} \right)^2 \right]^{1/2}. \tag{28}$$

We use $M = 20$, giving 21 rank bins. Lower RankDev indicates a rank distribution closer to uniform.

Spread–skill ratio (SSR). Let $\bar{x}_{c,i,j} = M^{-1}\sum_m \hat{x}^{(m)}_{c,i,j}$ and let

$$s_{c,i,j} = \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{x}^{(m)}_{c,i,j} - \bar{x}_{c,i,j} \right)^2 \right]^{1/2} \tag{29}$$

be the sample standard deviation across members. Denoting the latitude-weighted, channel-averaged RMSE in Eq. equation 24 by $\mathcal{E}_w(\cdot,\cdot)$, we compute

$$\text{SSR} = \frac{\mathcal{E}_w(s, 0)}{\mathcal{E}_w(\bar{x}, y) + \epsilon}. \tag{30}$$

The numerator measures ensemble spread and the denominator measures ensemble-mean skill; values closer to one are better.

Calibration error (CalibErr). For nominal central coverage p, let $q_\alpha(c,i,j)$ be the empirical α-quantile of the M members at (c,i,j). Its latitude-weighted empirical coverage is

$$\widehat{c}(p) = \frac{\sum_{(c,i,j)\in\mathcal{V}} w_i \, \mathbf{1}\left\{ q_{(1-p)/2}(c,i,j) \le y_{c,i,j} \le q_{(1+p)/2}(c,i,j) \right\}}{\sum_{(c,i,j)\in\mathcal{V}} w_i + \epsilon}. \tag{31}$$

Using $\mathcal{P} = \{0.1, 0.2, \dots, 0.9\}$, the scalar calibration error is

$$\text{CalibErr} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \left| \widehat{c}(p) - p \right|. \tag{32}$$

Lower CalibErr indicates closer agreement between nominal and empirical coverage. Unless stated otherwise, all six global metrics are first computed separately for each initial condition and then summarized across initial conditions.

Niño-3.4 diagnosis. Let $\mathcal{R}_{3.4}$ denote the region 5°S–5°N, 170°W–120°W. For initial condition q, the member and verifying indices are computed directly from the raw SST anomaly fields as

$$n_q^{(m)} = \frac{1}{|\mathcal{R}_{3.4}|} \sum_{(i,j)\in\mathcal{R}_{3.4}} \hat{s}_{q,i,j}^{(m)}, \qquad n_q^{\star} = \frac{1}{|\mathcal{R}_{3.4}|} \sum_{(i,j)\in\mathcal{R}_{3.4}} s_{q,i,j}^{\star}, \tag{33}$$

without adding a climate mean, normalizing variables, or applying latitude weights. Let $\bar{n}_q = M^{-1}\sum_m n_q^{(m)}$, $\mu_{\bar{n}} = I^{-1}\sum_q \bar{n}_q$, and $\mu_{n^\star} = I^{-1}\sum_q n_q^{\star}$. At each reported lead time, across $I = 10$ initial conditions, we compute

$$\begin{aligned}
\mathrm{RMSE}_{3.4} &= \left[\frac{1}{I}\sum_{q=1}^{I}\left(\bar{n}_q - n_q^{\star}\right)^2\right]^{1/2}, \\
\mathrm{Corr}_{3.4} &= \frac{\sum_q(\bar{n}_q - \mu_{\bar{n}})(n_q^{\star} - \mu_{n^\star})}{\sqrt{\sum_q(\bar{n}_q - \mu_{\bar{n}})^2 \sum_q(n_q^{\star} - \mu_{n^\star})^2}}, \\
\mathrm{CRPS}_{3.4} &= \frac{1}{I}\sum_{q=1}^{I}\left[\frac{1}{M}\sum_{m=1}^{M}|n_q^{(m)} - n_q^{\star}| - \frac{1}{2M^2}\sum_{m,n=1}^{M}|n_q^{(m)} - n_q^{(n)}|\right].
\end{aligned} \tag{34}$$

The three physical-consistency diagnostics below distinguish the evolution of the ensemble center from the spatial variability of individual members. The first two assess central-state consistency, whereas the third computes spatial variability before averaging across members. We use the latitude-weighted mean and standard deviation over valid ocean points $\mathcal{O}$,

$$\langle a\rangle_w = \frac{\sum_{(i,j)\in\mathcal{O}} w_i a_{i,j}}{\sum_{(i,j)\in\mathcal{O}} w_i}, \qquad \sigma_w(a) = \left[\frac{\sum_{(i,j)\in\mathcal{O}} w_i (a_{i,j} - \langle a\rangle_w)^2}{\sum_{(i,j)\in\mathcal{O}} w_i}\right]^{1/2}. \tag{35}$$

SSTa mean tendency error. Let $\bar{s}_t$ be the ensemble-mean SST anomaly field and y_t^s its verifying field at lead day t. Over the $T = 120$ day trajectory, we compute

$$E_{\mathrm{SSTa}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{\left|(\langle\bar{s}_t\rangle_w - \langle\bar{s}_{t-1}\rangle_w) - (\langle y_t^s\rangle_w - \langle y_{t-1}^s\rangle_w)\right|}{\sigma_w(y_t^s) + \epsilon}. \tag{36}$$

The latitude-weighted global mean summarizes the evolution of surface thermal anomalies. Comparing its daily increments tests whether the ensemble center follows the observed aggregate tendency throughout the rollout, rather than merely matching the final field. This provides a conservation-related consistency check on the mean trajectory, not a test of exact energy conservation: SST anomalies alone do not determine the full ocean heat budget.

SSH–current correlation error. For a valid grid set $\mathcal{B}$, let $\widetilde{a}_{i,j} = a_{i,j} - \langle a\rangle_{w,\mathcal{B}}$ and define $\widetilde{b}$ analogously. The latitude-weighted correlation is

$$\mathrm{corr}_w(a, b; \mathcal{B}) = \frac{\sum_{(i,j)\in\mathcal{B}} w_i \widetilde{a}_{i,j}\widetilde{b}_{i,j}}{\left[\sum_{(i,j)\in\mathcal{B}} w_i \widetilde{a}_{i,j}^2 \sum_{(i,j)\in\mathcal{B}} w_i \widetilde{b}_{i,j}^2\right]^{1/2}}. \tag{37}$$

Here $\mathcal{B}$ contains ocean points satisfying $10^\circ \le |\phi_i| \le 80^\circ$ whose immediate grid neighbors are also valid. The superscripts pred and gt denote the ensemble-mean and verifying fields, respectively. For $q \in \{\mathrm{pred}, \mathrm{gt}\}$, using discrete grid gradients of SSH anomaly,

$$r_u^q = \mathrm{corr}_w\left(\mathrm{U0a}^q, -\partial_y \mathrm{SSHa}^q; \mathcal{B}\right), \qquad r_v^q = \mathrm{corr}_w\left(\mathrm{V0a}^q, \partial_x \mathrm{SSHa}^q; \mathcal{B}\right), \tag{38}$$

and the ensemble-mean consistency error is

$$E_{\text{SSH-current}} = \frac{1}{2}\left(|r_u^{\mathrm{pred}} - r_u^{\mathrm{gt}}| + |r_v^{\mathrm{pred}} - r_v^{\mathrm{gt}}|\right). \tag{39}$$

SSH gradients and near-surface currents are dynamically related away from the equator. Comparing their correlations with those of the verifying state tests whether the ensemble center retains this cross-variable relationship, which separate field-wise errors cannot establish. The diagnostic measures agreement in aggregate multivariate structure; it does not impose a geostrophic constraint or measure its pointwise residual.

Member spatial standard-deviation error. For the $K = 5$ surface anomaly variables $\{\mathrm{SSSa}, \mathrm{U0a}, \mathrm{V0a}, \mathrm{SSTa}, \mathrm{SSHa}\}$, we compute

$$E_{\mathrm{std}} = \frac{1}{K} \sum_{k=1}^{K} \left| \log \frac{\frac{1}{M} \sum_{m=1}^{M} \sigma_w(x_k^{(m)}) + \epsilon}{\sigma_w(y_k) + \epsilon} \right|. \tag{40}$$

Unlike the first two diagnostics, this metric computes spatial variability within each member before averaging across members. It therefore complements center-based evaluation by detecting a deficit or excess in the average member variability across thermodynamic and dynamical surface fields. The absolute log-ratio penalizes deficient and excessive variability symmetrically. This amplitude-based diagnostic complements spectral analysis without requiring each member to reproduce the verifying field's spatial phase. All three physical diagnostics are computed independently for 20 initial conditions at the 120-day horizon and then summarized across initial conditions.

Lower RMSE, CRPS, RankDev, CalibErr, and physical-diagnostic errors are better; higher ACC and Niño-3.4 correlation are better; and SSR is optimal at one.

D Data Details

D.1 Dataset

We conduct the experiments on GLORYS12 reanalysis data, which provides daily mean ocean variables covering latitudes between $-80°$ and $90°$ from 1993 to the present. The data can be downloaded from https://data.marine.copernicus.eu. We use 1993–2017 for training, 2018–2019 for validation, and 2020 for the main test set. The additional 2024 evaluation in Table 4 assesses generalization beyond this test year without changing the training split.

The ocean variables used in our experiments include four depth-dependent variables and one surface variable. The four depth-dependent variables are Sea salinity (S), Sea stream zonal velocity (U), Sea stream meridional velocity (V), and Sea temperature (T), each of which contains 23 depth levels. These depth levels correspond to 0.49 m, 2.65 m, 5.08 m, 7.93 m, 11.41 m, 15.81 m, 21.60 m, 29.44 m, 40.34 m, 55.76 m, 77.85 m, 92.32 m, 109.73 m, 130.67 m, 155.85 m, 186.13 m, 222.48 m, 266.04 m, 318.13 m, 380.21 m, 453.94 m, 541.09 m, and 643.57 m. The surface variable is Sea surface height (SSH).

To improve computational efficiency, we downsample all variables to a spatial resolution of $1°$ using bilinear interpolation, resulting in an original grid size of $H = 181$ and $W = 360$. To better adapt the data to the input requirements of different model architectures, we retain the first 180 latitude grids and use data with a spatial size of 180×360 in the experiments.

Four ERA5 variables provide atmospheric forcing: 10 metre u wind component (U10M), 10 metre v wind component (V10M), 2 metre temperature (T2M), and mean sea level pressure (MSLP). Each ocean transition uses these four fields at the current and next day, yielding eight atmospheric input channels. Ocean states are daily means, while forcing fields are sampled at 12:00 UTC. ERA5 data are available from https://cds.climate.copernicus.eu. Table 8 summarizes the variables used in the main experiments.

D.2 Data Preprocessing

Different ocean variables and forcing variables have different physical meanings, units, and numerical ranges. To reduce the influence of these scale differences, we normalize each variable before feeding it into the model. This preprocessing step enables the model to focus more on learning the spatiotemporal evolution of ocean states rather than adapting to the magnitude discrepancies among variables.

For ocean variables, the mean and standard deviation are calculated from the training data spanning the years 1993 to 2017. For forcing variables, the corresponding statistics are computed from an extended ERA5 dataset covering the period 1959 to 2017. Each variable is normalized independently using its own mean and standard deviation. Specifically, for each variable, we first subtract its corresponding mean and then divide the result by its corresponding standard deviation.

Type	Full name	Abbreviation	Layers	Time	Dt	Spatial Res.
Atmosphere	10 metre u wind component	U10M	1	1993–2020	24h	1°
Atmosphere	10 metre v wind component	V10M	1	1993–2020	24h	1°
Atmosphere	2 metre temperature	T2M	1	1993–2020	24h	1°
Atmosphere	Mean sea level pressure	MSLP	1	1993–2020	24h	1°
Ocean	Sea salinity	S	23	1993–2020	24h	1°
Ocean	Sea stream zonal velocity	U	23	1993–2020	24h	1°
Ocean	Sea stream meridional velocity	V	23	1993–2020	24h	1°
Ocean	Sea temperature	T	23	1993–2020	24h	1°
Ocean	Sea surface height	SSH	1	1993–2020	24h	1°
Static	Land-sea mask	LSM	—	—	—	1°

Table 8: The data details used in this work.

For land regions, the original ocean variables contain 'nan' values. Before inputting the data into the model, these missing values are filled with zeros after normalization. The Land-sea mask is used as a static variable to indicate the distribution of land and ocean regions.

E Detailed Method

The construction below expands the coherent state-conditioned pathwise law in the main text. A deterministic reference supplies member conditions and a mean anchor, while ensemble-conditioned flow matching learns the residual distribution around that anchor. FDO parameterizes both transitions through the same operator composition with separate parameters. The distinction is between advancing the reference in physical time and jointly transporting its conditional residual ensemble in flow time.

E.1 Masked Normalization and Member Conditions

All ocean variables are transformed before they are passed to either the deterministic backbone or the flow model. Let q_t denote a raw ocean channel, $q^{\mathrm{clim}}_{d(t)}$ its day-of-year climatology, and $\mathrm{Norm}(\cdot)$ the channel-wise normalization computed over valid ocean points. With ocean mask M_o, the normalized state is

$$\tilde{q}_t = M_o \odot \mathrm{Norm}\Big(q_t - q^{\mathrm{clim}}_{d(t)}\Big). \tag{41}$$

The order matters: climatological anomalies and normalization are computed before land values are filled, because the climatology contains missing land points and the statistics are defined over the ocean. After normalization, land points are set to zero. Member perturbations are then added only to the ocean-state channels in this normalized space and masked again:

$$\tilde{q}^{(m)}_t = M_o \odot \Big(\tilde{q}_t + \sigma_\eta d_q\, \epsilon^{(m)}_q\Big), \tag{42}$$

where d_q is the variable/depth scale and $\epsilon^{(m)}_q$ is a smooth member-specific perturbation. The full member condition concatenates the perturbed ocean state with the unperturbed forcing variables. This ensures that land points are zero in every ocean input while the perturbations are normalized on ocean points only.

E.2 Structured Perturbation Reference

The perturbation reference provides structured variation in the input conditions and a common baseline for learning the conditional ensemble. It does not replace distributional training: the reference supplies member-specific centers and a spread scale, while coherent residual transport learns how the resulting ensemble should be adjusted. Smooth perturbations are generated by variable

group and normalized over valid ocean cells:

$$\epsilon_q^{(m)} = \frac{M_o \odot \xi_q^{(m)}}{\sqrt{\frac{\sum_{i,j} M_o(i,j)\left(\xi_q^{(m)}(i,j)\right)^2}{\sum_{i,j} M_o(i,j)+\varepsilon} + \varepsilon}}. \tag{43}$$

The scale d_q depends on the physical variable and depth. Salinity, temperature, velocity, and sea-surface height use separate base scales, and subsurface variables decay with depth. The global amplitude σ_η is a positive learnable scalar optimized jointly with the probabilistic model, whereas d_q fixes the relative scaling across variable groups and depths. Neither the shared forcing channels nor the land–sea mask is perturbed. Learning this amplitude calibrates the reference input; the ensemble-conditioned flow subsequently learns the residual adjustment rather than relying on perturbation amplitude alone.

E.3 Mean-Anchored Residual Decomposition

Mean anchoring gives central-state accuracy and ensemble diversity distinct roles within one conditional distribution. The mean branch learns a bounded response relative to the reference center, while the residual branch refines member departures without determining the final ensemble mean. For each channel, define the ocean-domain projection

$$\mathcal{P}_o r = M_o \odot \left(r - \frac{\sum_{i,j} M_o(i,j) r(i,j)}{\sum_{i,j} M_o(i,j)} \right), \tag{44}$$

where the mask contains at least one valid cell. Equation equation 3 applies this projection both before and after the bounded nonlinearity: the second projection restores the zero-spatial-mean constraint that a nonlinear transformation need not preserve. This spatial constraint is distinct from ensemble recentering, which fixes the member average at μ for every generated ensemble. Distributional training can update the mean branch, but residual transport cannot displace a given anchor after the final projection.

E.4 Coherent Residual Transport

The coherent law acts on the conditional residual ensemble around the mean anchor. Member-specific reference anomalies preserve the variation supplied by the deterministic reference, while learned transport reshapes the distribution through velocities informed by the evolving ensemble. For the M members used by the residual branch, write $\bar{x}_{\text{base}} = M^{-1} \sum_m x_{\text{base}}^{(m)}$ and

$$b^{(m)} = x_{\text{base}}^{(m)} - \bar{x}_{\text{base}}. \tag{45}$$

The member-specific centers are $\mu + b^{(m)}$. During training, the reference mean used to construct μ is estimated with $M_b = 20$ members, while the residual loss uses $M = 10$ particles; at evaluation, $M_b = M = 20$. The reference anomalies are centered over the residual member set and also supply its spread reference.

Residual particles are initialized independently of the reference anomalies as

$$z_0^{(m)} = \sigma_0 \epsilon_0^{(m)}, \tag{46}$$

where $\epsilon_0^{(m)}$ is smooth, spatially centered noise with per-member RMS normalization. Member centering precedes normalization, so an exactly zero empirical ensemble mean is imposed at the output rather than assumed for every initial particle set. Let $e = y - \text{sg}(\mu)$, where sg denotes stop-gradient. The spread target combines a spatially smoothed absolute error and a channel-wise ocean RMS, followed by scaling and bounding. With $r_{\text{pat}}^{(m)}$ denoting centered, normalized member coefficients multiplying the spatial error pattern, the target is

$$z_1^{(m)} = (1-\gamma_p) s_{\text{target}} \odot \epsilon_0^{(m)} + \gamma_p \, r_{\text{pat}}^{(m)}(e), \qquad \frac{1}{M} \sum_m r_{\text{pat}}^{(m)} = 0, \tag{47}$$

followed by ocean masking. We use $\gamma_p = 0.05$. This target defines a state-dependent residual transport task, while mean and probability objectives supervise the resulting ocean ensemble directly.

A single τ is sampled for all members in a training ensemble, so their velocities are learned against the context of one common interpolation state. The member paths therefore remain distinct while participating in the same ensemble-conditioned transport.

Training and inference both use two transport updates with $\Delta\tau = 1/2$. At each update, the ensemble context is recomputed from the current residual particles, the shared velocity is evaluated at $\tau_k = (k + 1/2)\Delta\tau$, and $\mathcal{P}_r$ applies the residual bound and ocean mask as in Eq. equation 7. Here pathwise refers to residual evolution along flow time within a single ocean transition. Let $\mathcal{H}_\theta$ denote residual centering, smoothing, and learned spread calibration, and let $\mathcal{B}_{\mu,B}$ stabilize the composed members, recenter them at μ, and apply the reference-spread floor. The composition entering the final projection is

$$\widetilde{Z}_{N_\tau} = \mathcal{K}^M_{\theta,c}(Z_{N_\tau}, \Xi) = \mathcal{B}_{\mu,B}(\mathbf{1}_M \otimes \mu + B + \mathcal{H}_\theta(Z_{N_\tau}; c)) - \mathbf{1}_M \otimes \mu, \tag{48}$$

where $B = (b^{(1)}, \ldots, b^{(M)})$. Spread calibration blends the centered residual with a version normalized to the learned spread map; stabilization bounds member departures, and the reference floor rescales their channel-wise ensemble RMS when it falls below the reference fraction. Equation equation 8 is applied last. Thus the reported distribution includes reference anomalies and learned residual structure, with an empirical center controlled independently of these operations.

All members share an embedding of the current ocean–forcing condition. In the reported one-step training and inference configuration, this context is initialized anew at each ocean transition; recurrent updates are not carried across days. The embedding supplies common state information, whereas $S(Z_\tau)$ changes during integration and supplies the evolving ensemble context. Coherence therefore arises from joint residual conditioning within each transition, not from persistent cross-day memory.

E.5 Ensemble-Conditioned Transport

Ensemble conditioning connects member-specific paths to a collective residual state. For particles $Z = \{z^{(m)}\}_{m=1}^M$, the velocity receives symmetric summaries

$$S(Z) = \left[\frac{1}{M} \sum_m z^{(m)},\ \sqrt{\frac{1}{M} \sum_m (z^{(m)} - \bar{z})^2},\ \frac{1}{M} \sum_m |z^{(m)} - \bar{z}| \right]. \tag{49}$$

The mean locates the current residual center, while the standard deviation and mean absolute deviation describe complementary aspects of its dispersion. These are conditioning features, not fixed moment constraints. The mean and standard deviation enter each member's input directly, and a learned projection of all three summaries supplies an additional ensemble feature. Recomputing them at each integration step allows the velocity to respond to changes in the joint residual state. Together with shared velocity parameters, this gives

$$\mathcal{T}^M_\theta(\Pi Z \mid \Pi C) = \Pi \mathcal{T}^M_\theta(Z \mid C), \tag{50}$$

where particles and member conditions are permuted together. Masking, clipping, reference composition, and recentering respect the same symmetry. Member ordering therefore does not determine the dynamics. The substantive coupling comes from the dependence of each velocity on the current ensemble summaries, as formalized in Eq. equation 19; the symmetry ensures that this coupling is expressed without assigning significance to arbitrary member indices.

E.6 Factorized Dynamics Operator

FDO expresses a conditional transition through factorized dynamics and a learned response to the state on which those dynamics act. This operator composition supports both an accurate deterministic reference and an ensemble-conditioned residual velocity without identifying either task with the other. For the residual instance, the spatial input is $\mathbf{X}^{(m)} = [z^{(m)}, c_t^{(m)}, \bar{z}, \sigma(Z), \mathcal{J}_\theta(S(Z))]$, where $\mathcal{J}_\theta$ projects the ensemble summaries. Its context $\mathbf{C}$ combines flow time with an embedding of the current ocean–forcing condition. The deterministic instance instead takes the ocean–forcing tensor as input and uses a learned static context embedding. The conditional patch encoder gives

$$H_0 = \mathcal{E}^{\text{patch}}_\theta\Big(\mathbf{X}^{(m)}; \mathbf{C}\Big) \in \mathbb{R}^{D \times H_p \times W_p}. \tag{51}$$

Within each block, contextual modulation determines the representation on which factorized dynamics operate,

$$U_\ell = a_\ell(\mathbf{C}) \odot \mathcal{N}_\ell(H_\ell) + b_\ell(\mathbf{C}), \tag{52}$$

where $\mathcal{N}_\ell$ is normalization. A spatial transformation $\mathcal{M}_\ell$ and learned projection $\mathcal{P}_\ell$ induce K factors of width $d = D/K$. Factor-specific evolution and cross-factor interaction then take the form

$$\begin{aligned} F_{\ell,k} &= [\mathcal{P}_\ell(\mathcal{M}_\ell(U_\ell))]_k, \qquad G_{\ell,k} = \mathcal{T}_{\ell,k}(F_{\ell,k}), \\ \widehat{G}_{\ell,k} &= \sum_{j=1}^{K} B_{\ell,kj} G_{\ell,j}, \qquad B_\ell = I + \beta \tanh(A_\ell). \end{aligned} \tag{53}$$

Separate factor operators allow distinct transformations within the learned representation, while the near-identity matrix B_ℓ introduces cross-factor interactions. This matrix is learned but state independent; state-dependent selection determines the contribution of the interacting factors at each location. The complete update is

$$\begin{aligned} \Gamma_\ell &= \mathrm{softmax}_k(\mathcal{G}_\ell(U_\ell)), \\ H'_\ell &= H_\ell + \Psi_\ell\Big(\mathrm{Concat}_k(\Gamma_{\ell,k} \odot \widehat{G}_{\ell,k})\Big), \\ H_{\ell+1} &= H'_\ell + \mathcal{C}_\ell(H'_\ell; \mathbf{C}). \end{aligned} \tag{54}$$

Here Ψ_ℓ integrates the weighted factors into the state representation, and $\mathcal{C}_\ell$ supplies a normalized, context-modulated channel update. Together they instantiate Φ_ℓ in Eq. equation 10. Factor indices denote learned subspaces rather than predefined physical variables: factorization organizes the transition without requiring a prescribed correspondence between factors and ocean processes. These updates act in both encoding and decoding streams, connected by a learned width projection. After normalization and patch restoration through $\mathcal{R}_\theta$, the state-conditioned response completes the operator composition,

$$\widetilde{v}_\theta = \mathcal{R}_\theta(H_L), \qquad v_\theta = \left[\mathcal{I} + \gamma \mathcal{S}_\theta(\cdot \mid \mathbf{X}^{(m)})\right](\widetilde{v}_\theta) = \widetilde{v}_\theta + \gamma \mathcal{S}_\theta\left(\widetilde{v}_\theta, \mathbf{X}^{(m)}\right). \tag{55}$$

The response jointly conditions on the decoded factorized evolution and the complete operator input. It learns how that evolution contributes to the output in the current state, rather than prescribing a fixed persistence addition. For the velocity instance, this input includes member residuals and ensemble context; for the deterministic instance, it is the ocean–forcing state. The two instances share the operator form, not network weights. The residual output projection and response coefficient are initialized to zero, while the deterministic instance is trained before ensemble learning. The ablations in Table 7 assess the complete factorized block and the contribution of a learned, rather than absent or fixed, state response.

E.7 Training Objectives

Training establishes central-state skill before learning the residual ensemble distribution. The reference FDO first learns the normalized one-step transition

$$\widehat{\mathbf{O}}_{t+1} = f_{\mathrm{det},\vartheta}(c_t) = \mathcal{M}_\vartheta(\mathbf{O}_t, \mathbf{F}_t, \mathbf{F}_{t+1}), \tag{56}$$

using the relative L_2 risk in Eq. equation 12. Relative errors are computed separately for each ocean channel and then averaged across channels and training samples; the spatial norm is neither squared nor latitude weighted. The selected pretrained parameters ϑ remain frozen throughout the probabilistic stage, while θ denotes its trainable parameters.

Let $\ell_{\mathrm{cw}}(a, y)$ denote the channel-averaged relative L_2 error appearing inside that expectation. State-field errors use the ocean mask, with land cells retained as zeros in the fixed-grid average,

$$\|u\|_{p,o}^p = \frac{1}{C_o HW} \sum_{q,i,j} |M_{o,q}(i,j) u_q(i,j)|^p. \tag{57}$$

This convention matches the velocity error and empirical CRPS used in training. The expectation in Eq. equation 13 averages these sample-level objectives over training transitions and stochastic

particle construction. Its mean component combines central-state supervision with improvement relative to the frozen reference,

$$\begin{aligned}\ell_{\text{mean}} = a_\mu \ell_{\text{cw}}(\mu, y) + a_{\text{ens}} \ell_{\text{cw}}(\bar{x}, y) \\ + a_{\text{imp}} \left[\|\bar{x} - y\|_{2,o}^2 - \text{sg}\big(\|\mu^{\text{base}} - y\|_{2,o}^2\big) \right]_+ ,\end{aligned} \tag{58}$$

where $\bar{x} = M^{-1} \sum_m \hat{x}^{(m)}$, $[r]_+ = \max(0, r)$, and the coefficients absorb the relative weights of the mean objectives. The ensemble coefficient includes supervision of the corrected reference through $\mu^{\text{base}} + (\bar{x} - \mu^{\text{base}})$, which has the same relative-field loss value as $\bar{x}$. The reference comparison is treated as a fixed target under differentiation.

Probability supervision acts on the ensemble produced by the learned transport, not solely on the prescribed interpolation targets. It combines empirical CRPS in Eq. equation 14 with the reference-relative penalty $\ell_{\text{gain}} = [\mathsf{S}_M(\widehat{X}, y) - \text{sg}(\mathsf{S}_M(X_{\text{base}}, y))]_+$. Gradients propagate through the integrated residual paths, linking velocity learning to the quality of the terminal ensemble. The M^2 denominator corresponds to the empirical ensemble score rather than the finite-member fair-score correction. Member-specific reference centers are detached in the probability and member losses, separating their construction from learned residual adjustment.

The remaining terms supervise member error and the amplitude of the learned distribution. Let $\sigma_M(X)$ be the pointwise ensemble standard deviation with denominator M, $\mathcal{H}$ the spatial smoothing map, and $\hat{s}_\theta$ the learned spread field. For derived spread fields, write $\|u\|_{2,g}^2 = (C_o HW)^{-1} \sum_{q,i,j} u_q(i,j)^2$ for the full-grid mean square; smoothing acts on masked fields without an additional mask before reduction. The spread objectives are

$$\begin{aligned}\ell_{\text{spread}} &= \|\mathcal{H}(\sigma_M(\widehat{X})) - s_{\text{target}}\|_{2,g}^2 + a_s \|\hat{s}_\theta - s_{\text{target}}\|_{2,g}^2, \\ \ell_{\text{anchor}} &= \|[\rho\, \sigma_M(X_{\text{base}}) - \sigma_M(\widehat{X})]_+\|_{2,o}^2.\end{aligned} \tag{59}$$

The regularization group ℓ_{reg} is a weighted sum of these terms and $M^{-1} \sum_m \ell_{\text{cw}}(\hat{x}^{(m)}, y)$. Spread targets are derived from the detached center error. The training spread penalty acts pointwise, whereas the sampling floor in $\mathcal{B}_{\mu,B}$ acts on channel-wise ensemble RMS. The objectives thus connect local velocity supervision to terminal ensemble quality and spread control. Exact center preservation is imposed by the final projection rather than learned as a soft penalty.

E.8 Reference-Guided Autoregressive Inference

The rollout begins from an observed ocean state and its forcing, with structured perturbations establishing the initial member conditions. At each day, the deterministic reference advances these conditions; mean correction and coherent residual transport then produce the evaluated ensemble through Eq. equation 48 and the final projection in Eq. equation 8. The next member conditions use the reference outputs, not the transported members, and the shared context is reinitialized from the new condition. The conditional distribution is therefore refined at every transition while the reference determines the autoregressive progression. All reported `OceanFlow` ensemble metrics evaluate the transported members, not the raw reference ensemble.

For simulation, the forcing pair $(\mathbf{F}_t, \mathbf{F}_{t+1})$ is read from the test reanalysis sequence. For prediction with external forcing, the atmospheric state is advanced by OneForecast at 6-hour resolution. If the ocean initial condition corresponds to daily index t, the atmospheric rollout is initialized from the aligned 12:00 UTC atmospheric state and advanced four 6-hour steps for each ocean day; the ocean model then consumes the current-day and next-day atmospheric fields as its forcing pair. This keeps the ocean rollout protocol identical between simulation and prediction except for the source of future forcing.

The preliminary coupled-forcing experiment in Table 5 instead uses atmospheric states from an independently running two-way PnP-Corrector system. The label "two-way" identifies the system supplying the forcing; `OceanFlow` does not feed its generated ensemble back into that atmospheric model. This extension tests the benefit of coupled forcing while leaving end-to-end probabilistic ocean–atmosphere coupling to future work.

E.9 Ablation Design

The comparisons in Table 7 distinguish the conditional operator from the probabilistic transport built upon it. The deterministic variants are evaluated against FDO. *w/o Factorized Dynamics* replaces the complete factorized block with a plain convolutional block while retaining patch encoding, decoding, and the final state response. It therefore tests the proposed block as a whole. *w/o State Response* retains factorized dynamics but returns the decoded field directly. *w/o Adaptive State Response* instead adds the input ocean field to the decoded output with a fixed unit coefficient. The latter two comparisons distinguish direct decoding, fixed persistence, and the learned state-conditioned response within the same factorized architecture.

The probabilistic comparisons retain the deterministic FDO reference. *FDO+Perlin* evaluates the perturbation ensemble without the learned mean correction or residual transport. *w/o mean correction/recentering* retains learned transport but removes the mean correction and center-preserving projections. *w/o learned pathwise flow* retains the reference, mean correction, and final recentering but disables the learned residual-flow contribution. Together, these comparisons assess whether the observed gains arise from deterministic representation alone or from learning the conditional residual distribution under explicit center control.

F Additional Results

The additional results examine whether the gains in aggregate scores extend across depth, lead time, and individual ensemble members. They complement the main comparisons by separating the quality of the predictive center from the spatial structure and diversity of the conditional ensemble.

Figure 9 reports the 1–120-day simulation RMSE, ACC, and CRPS at representative subsurface levels (109.73 m, 222.48 m, and 453.94 m), with curves and shaded bands denoting the mean and standard deviation over the first 20 initial conditions starting from January 1, 2020 at one-day intervals. Due to ocean states typically vary smoothly in the vertical direction, adjacent depth levels often exhibit similar error patterns. We therefore report representative depths to provide a compact and interpretable assessment of the model's ability to capture three-dimensional ocean structure.

Figure 10 compares ensemble-mean and individual-member simulation fields at the 120-day horizon for the first IC. Each model is represented by its ensemble mean and the same three fixed member indices (members 1, 10, and 20), rather than members selected by their agreement with the ground truth. The examples show how `OceanFlow` combines a spatially organized center with structured variation among individual members. These fields complement the member-averaged spectra and physical diagnostics in the main text: they illustrate the generated distribution beyond its mean, while the aggregate diagnostics assess its behavior across ICs and members.

Additional visual results are shown in Figures 11, 12, 13, 14, covering 17 key ocean variables at lead times from 30 to 120 days, initialized from Jan. 1, 2020. For salinity, temperature, and SSH, whose variations are relatively small compared with their background values, we visualize anomalies to better highlight these subtle yet meaningful changes. In contrast, the original fields are directly shown for the zonal and meridional velocity components, U and V. For clarity, we provide several examples of the variable naming convention: SSSa denotes the sea surface salinity anomaly, while Sa109, Sa222, and Sa453 denote salinity anomalies at depths of 109.73 m, 222.48 m, and 453.94 m, respectively. Similarly, U0 denotes the sea surface zonal velocity, and U109, U222, and U453 denote the zonal ocean velocity at the corresponding depths. Collectively, these additional visualizations further demonstrate the effectiveness of the proposed `OceanFlow` in capturing multivariate ocean-state evolution over extended simulation horizons.

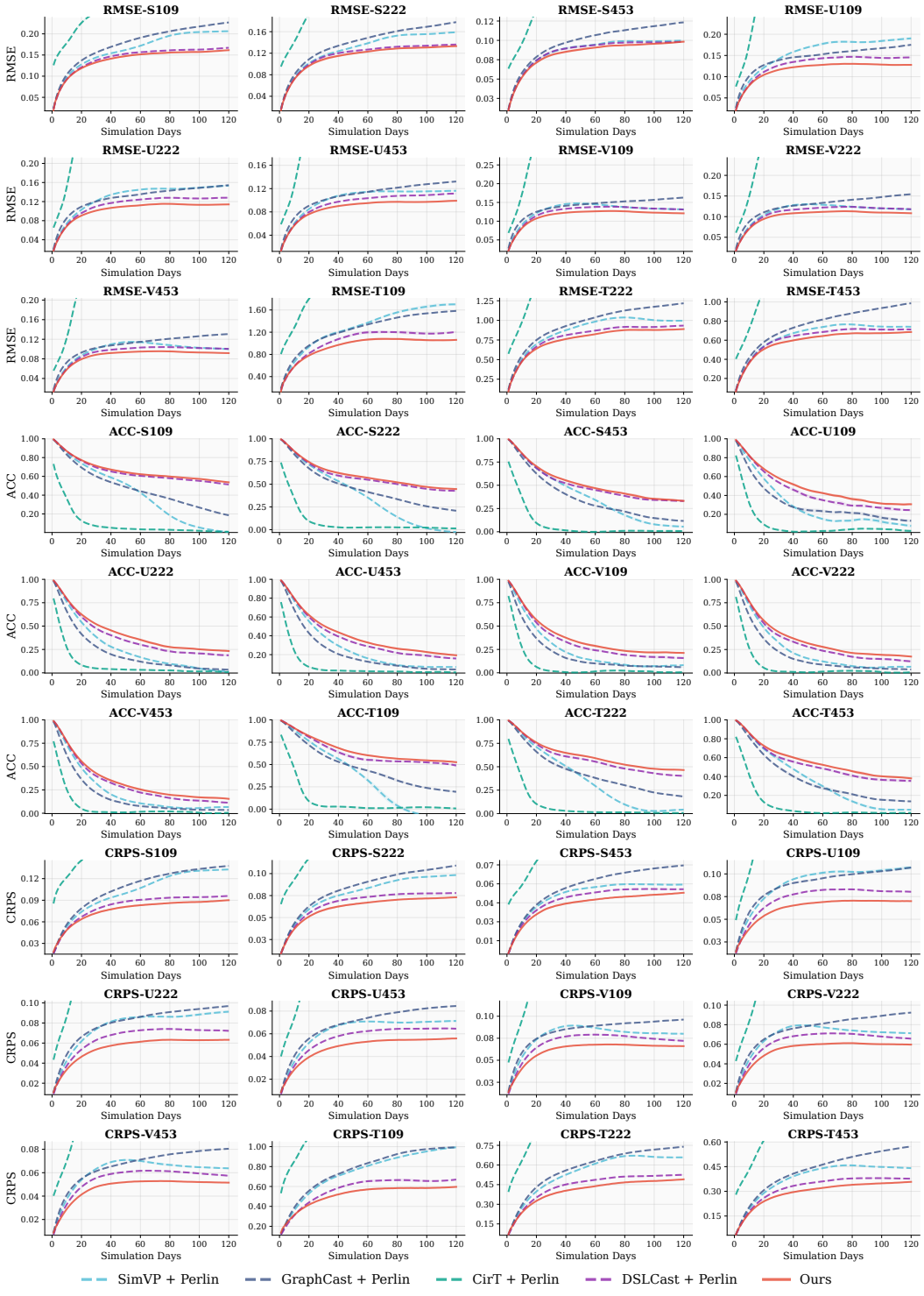

Figure 9: Simulation RMSE, ACC, and CRPS over 1–120 days at depths of 109.73 m, 222.48 m, and 453.94 m. Lines and shaded bands show the mean and standard deviation across 20 ICs. Lower RMSE and CRPS and higher ACC are better.

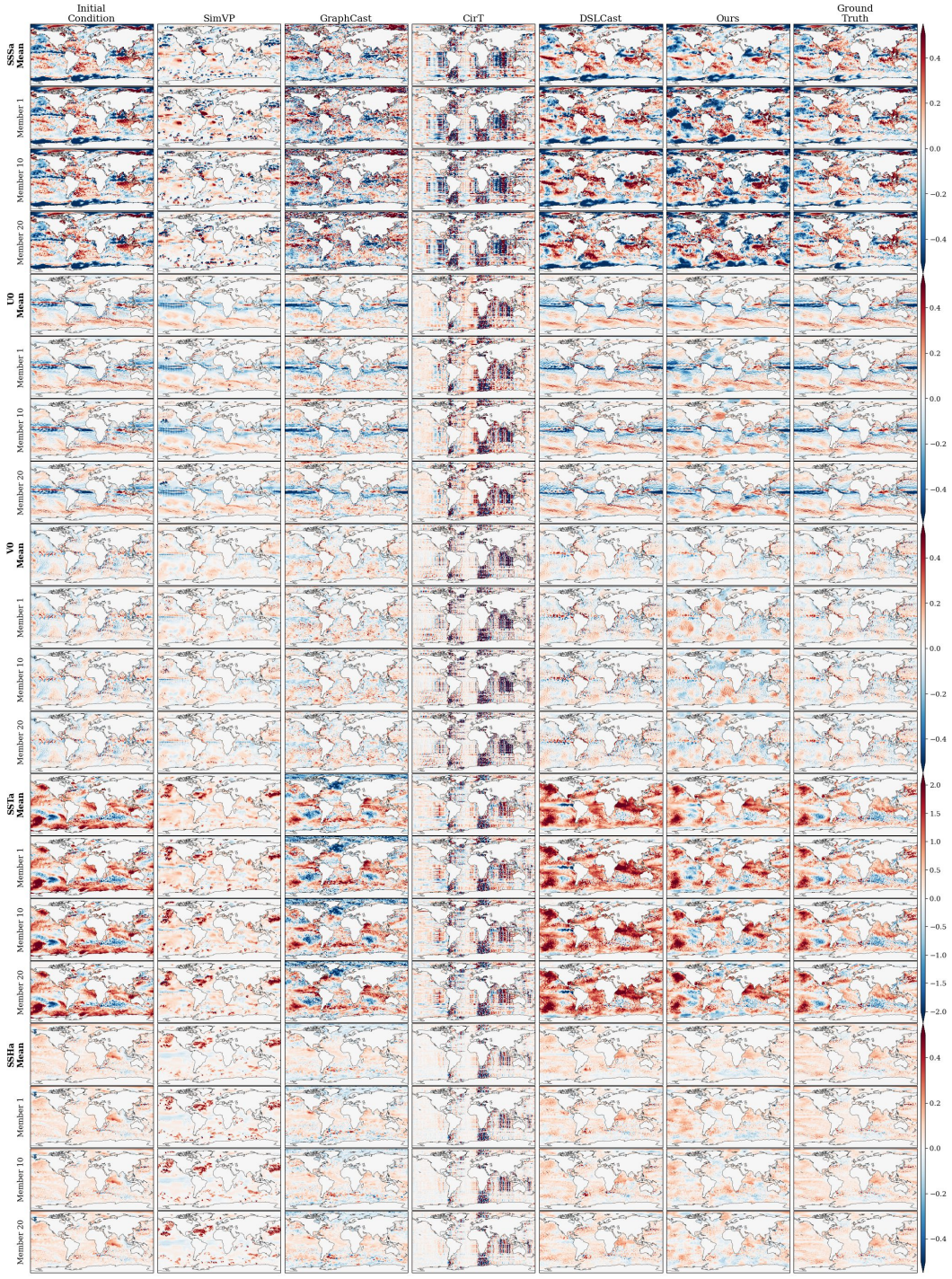

Figure 10: 120-day ensemble-mean and member-level simulation results at representative depth levels for the first IC. Individual members use fixed indices 1, 10, and 20 for every model.

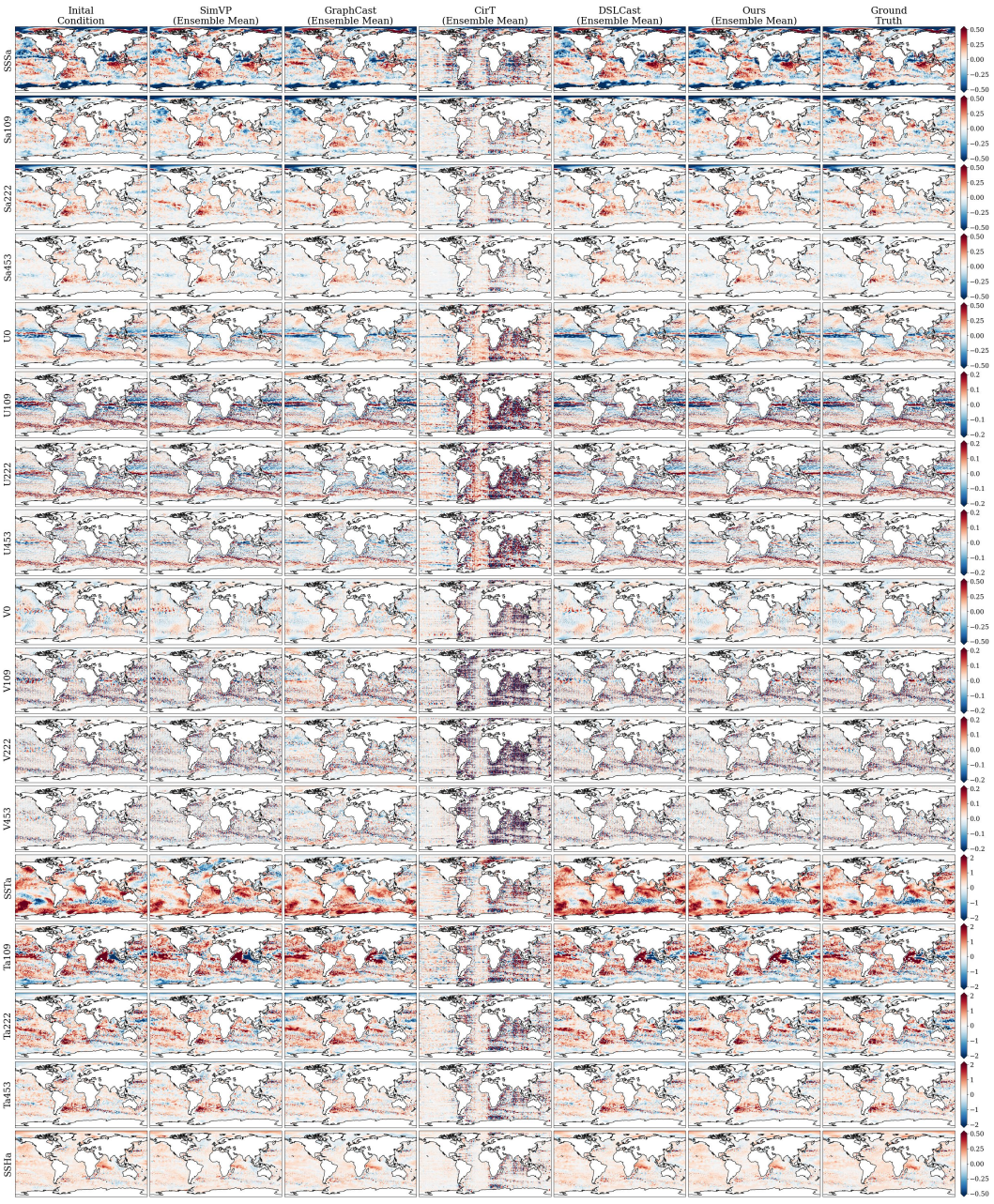

Figure 11: 30-day simulation results of different models at several representative depth levels.

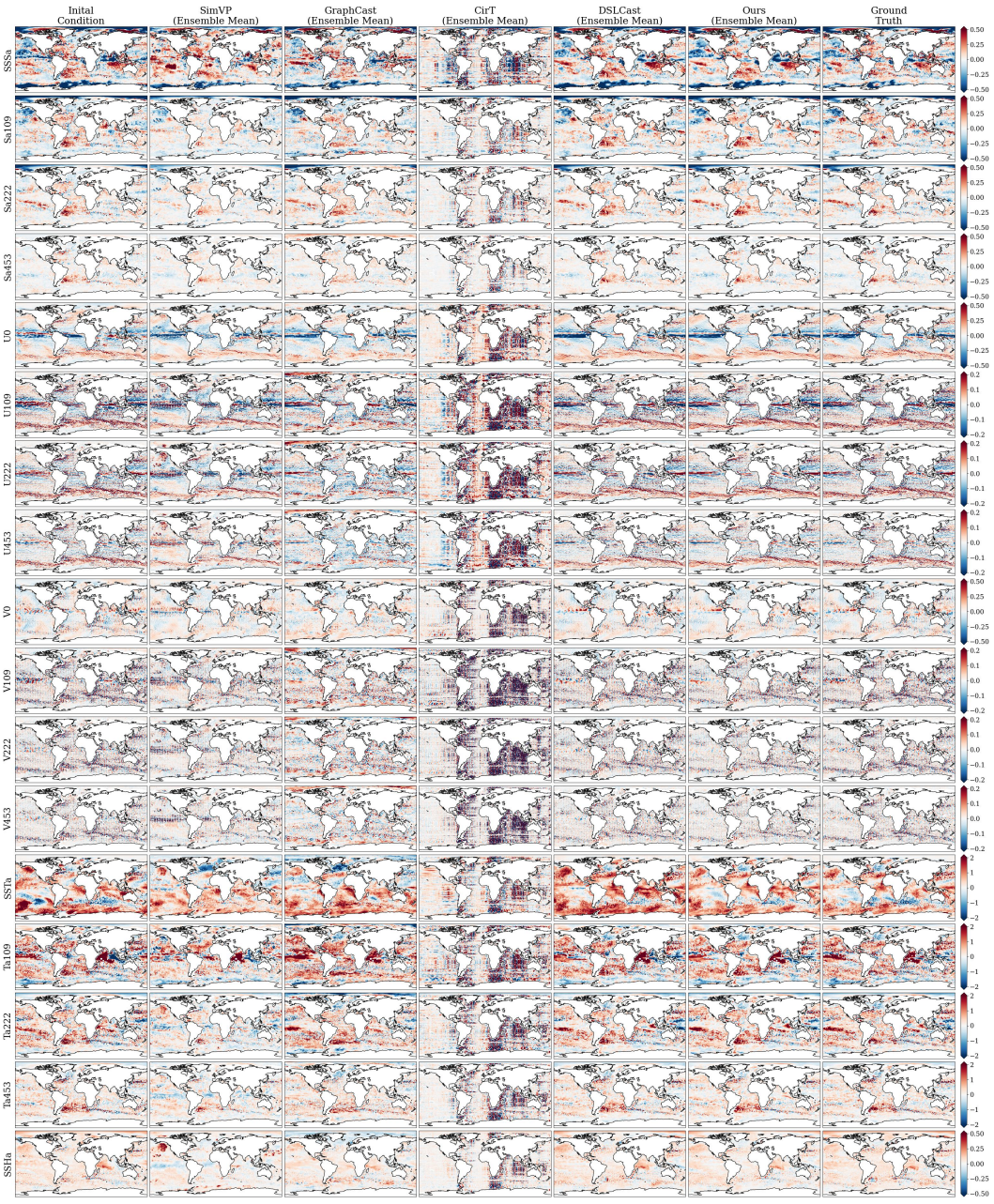

Figure 12: 60-day simulation results of different models at several representative depth levels.

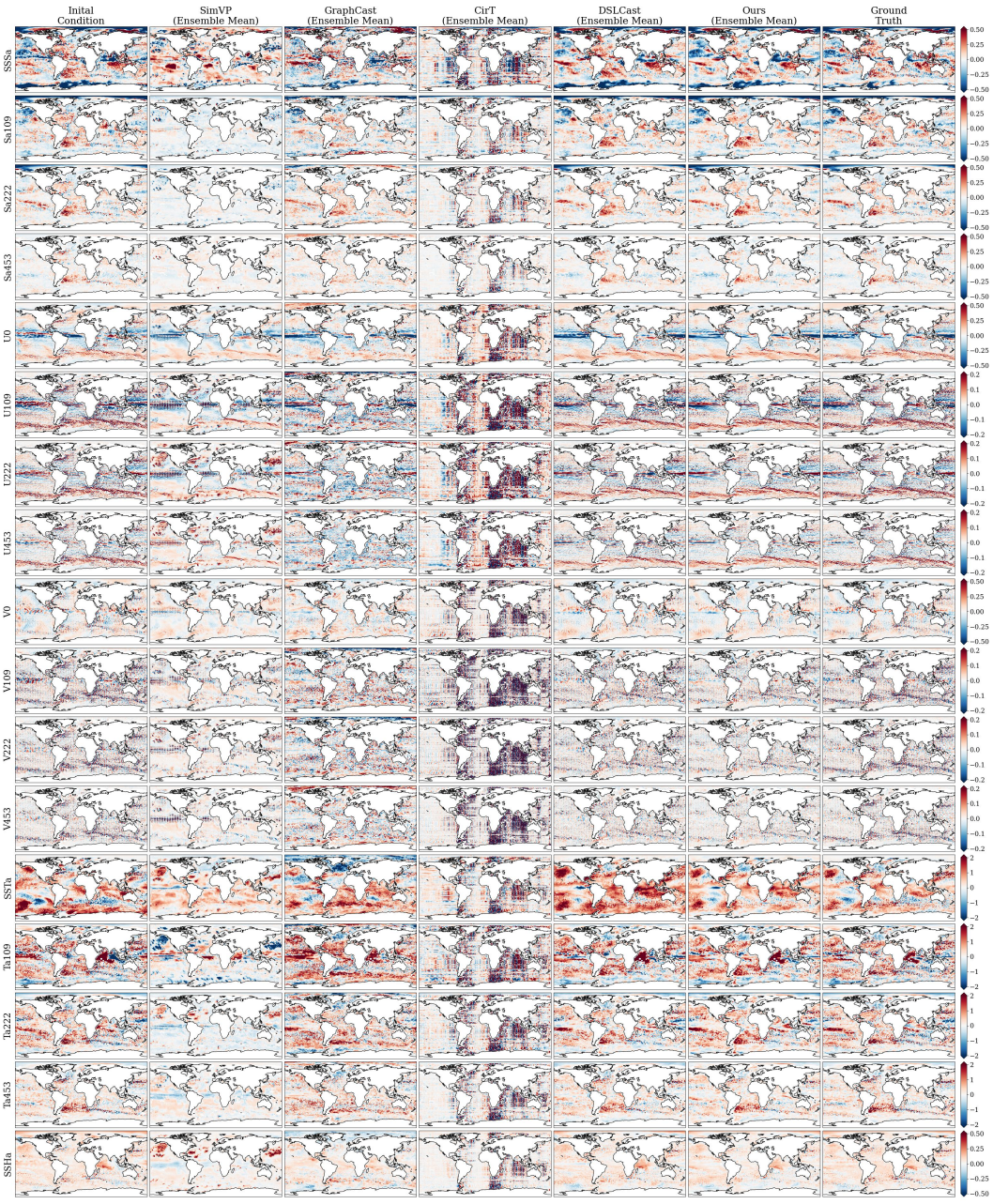

Figure 13: 90-day simulation results of different models at several representative depth levels.

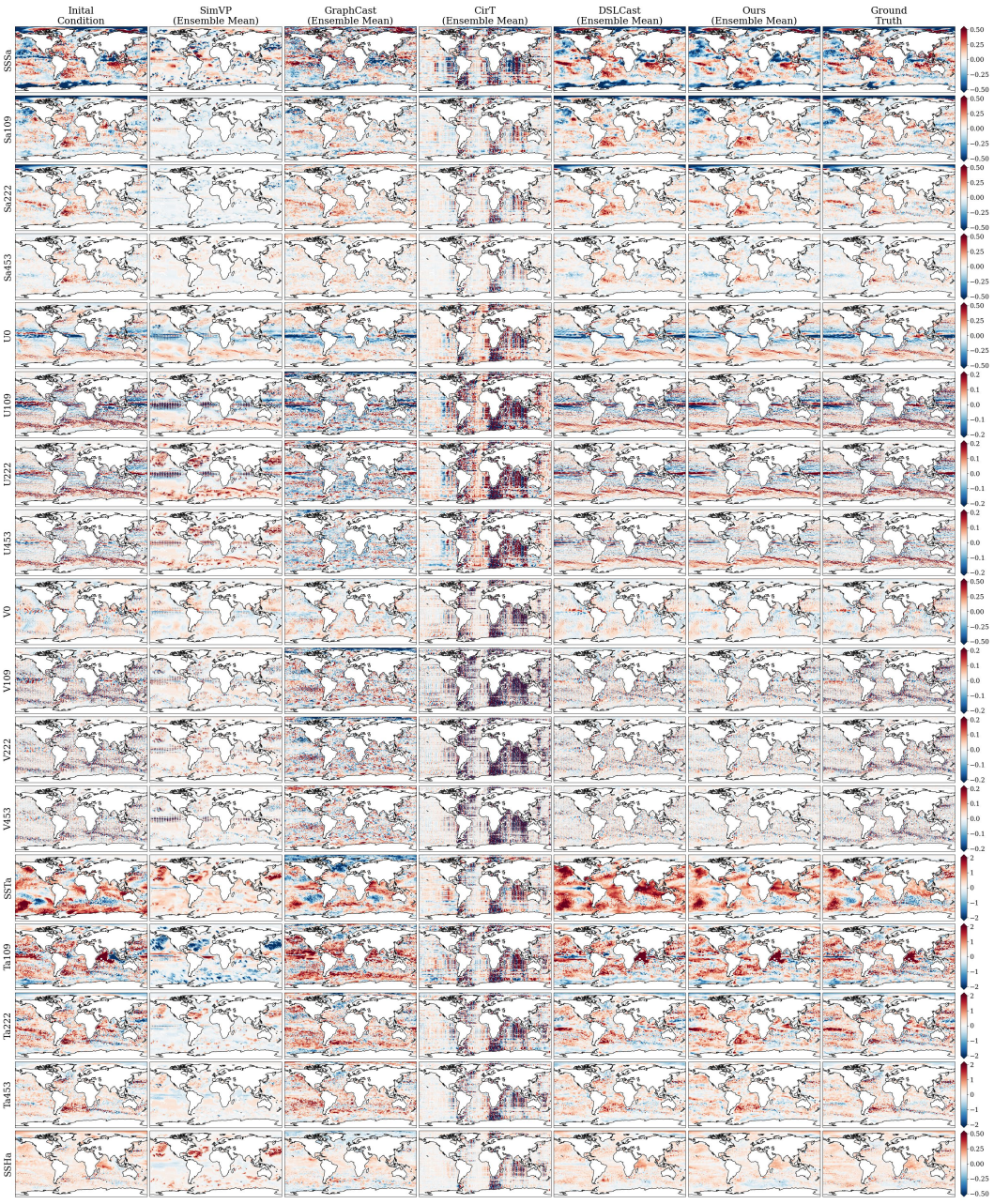

Figure 14: 120-day simulation results of different models at several representative depth levels.