REACT: Physically and Chemically Consistent Reconstruction of Marine Active Tracers

Wenbin Dai[1,2,*] **Hao Zheng**[1,*] **Shiyu Liang**[1,2,*,†] **Chaofan Sun**[1] **Xueying Zhang**[1]
Hanbo Huang[1] **Xuan Gong**[1,2] **Yiran Zhang**[1] **Enhui Liao**[1]
[1]Shanghai Jiao Tong University [2]Shanghai AI Laboratory

Abstract

Reconstructing global sea surface pH from sparse observations is critical for monitoring ocean acidification and understanding marine carbon cycling. Traditional assimilation and inverse models are physically grounded but costly for large-scale reconstruction. Recent black-box and physics-guided AI models improve efficiency, but are mainly designed for passive tracers, where the reconstructed variable is also the transported inventory. In contrast, pH is an active carbonate tracer: it is the prediction target, while dissolved inorganic carbon (DIC) is the conserved carbon inventory. This mismatch can produce low pH error while violating carbonate closure and source-free carbon conservation. To address this, we introduce **REACT**, a carbon-first reconstruction framework that decouples transport, active correction, and chemical decoding. REACT transports a latent carbonate state with a conservative advection–diffusion solver, captures non-conservative carbon-cycle variations with a source module, decodes the corrected state into pH, and constrains the output through carbonate equilibrium. This design keeps pH as the target while enforcing consistency on the underlying carbon state. On simulation data, REACT reduces pH NRMSE by 14.7% and chemical consistency error by 24.0% over the best baseline. Cross-temporal-scale evaluations show robustness against error accumulation from coarse to fine temporal scales, and ablation studies validate the effectiveness of each component.

1 Introduction

Reconstructing global sea surface pH from sparse observations is essential for monitoring ocean acidification and marine carbon cycling. However, direct pH measurements remain highly sparse and irregular, with valid records covering only 0.11% of the global spatiotemporal grid according to the World Ocean Database (WOD)[1]. This challenge is further complicated by the fact that pH is not an unconstrained scalar variable, but a chemical coordinate of the ocean carbonate system governed by transport, biological activity, and carbonate equilibrium [2–4]. Consequently, reconstruction models must infer dense pH fields from limited observations while preserving consistency with the physical and chemical structure of the carbonate system.

Recent advances in deep learning have improved the reconstruction of geophysical fields. CNNs and transformers capture spatial or spatiotemporal correlations from gridded data [5–7]. Inpainting models and climate foundation models provide powerful generic completion tools [8–11]. Physics-guided methods further introduce PDE constraints, neural operators, or variational consistency [12–16]. However, most of these methods are designed for passive tracers, where the reconstructed variable is also the transported inventory. This assumption does not hold for pH.

*Equal contribution.
†Corresponding author. Correspondence to lsy18602808513@sjtu.edu.cn.
Technical inquiries: Wenbin Dai (daiwenbin@sjtu.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Reconstructing sea surface pH requires a fundamentally different approach. Unlike conserved tracers, pH is a chemical coordinate of the carbonate system. The conserved carbon inventory is closer to dissolved inorganic carbon (DIC), while pH is determined by carbonate equilibrium jointly with variables such as surface $p\mathrm{CO}_2$. This induces a structural mismatch: direct transport of pH may produce smooth fields with negligible error, while allowing the implied DIC to drift and potentially violating carbonate closure. Therefore, pH error alone is insufficient. A valid reconstruction should simultaneously recover accurate pH fields, maintain consistency with carbonate chemistry and preserve conservation in the source-free carbon state.

The main challenge is to disentangle the three mechanisms jointly represented by the pH observations. Conservative transport redistributes carbon through ocean flow, while non-conservative processes, including air–sea exchange, biological activity, vertical mixing, and unresolved reactions, modify the carbonate state. Carbonate equilibrium then maps the carbon state to pH. Modeling pH as a passive tracer therefore applies transport dynamics to an inappropriate variable, whereas fully black-box reconstruction avoids this mismatch at the cost of providing limited interpretability. A principled formulation should instead transport a carbon-related state, represent non-conservative effects through a source module, and decode pH subject to carbonate-system constraints.

We propose **REACT**, short for **RE**constructing **A**ctive **C**arbonate **T**racers, to address this mismatch via a carbon-first reconstruction principle. At its core is a latent carbonate state that serves as the transported physical coordinate, while pH remains the decoded target. The latent state is propagated through a conservative advection–diffusion solver, corrected by an active source module for non-conservative carbon-cycle effects, and decoded to pH under the carbonate equilibrium constraints of DIC and $p\mathrm{CO}_2$. This design separates transport, active correction, and chemical decoding, preserving source-free carbon mass without imposing conservation on pH. We further show that, under standard conditions, small chemical and physical losses certify carbonate closure and represented DIC conservation, a guarantee that passive pH transport generally cannot provide. Empirically, REACT consistently improves pH reconstruction and chemical consistency on both simulation-based cross-temporal tasks and real-world sparse observations, while ablations validate the role of source correction, mass conservation, and chemical closure. Our contributions are as follows:

- We formulate sea surface pH mapping as *active carbonate-tracer reconstruction*, revealing the mismatch between the reconstructed variable, pH, and the conserved carbon inventory, DIC.
- We propose **REACT**, a carbon-first framework that combines latent carbonate-state transport, active source correction, pH decoding, and carbonate-system closure. We provide a theoretical analysis showing why carbon-state transport can satisfy chemical closure and source-free DIC conservation.
- We empirically demonstrate that REACT achieves state-of-the-art performance on both simulation-based and real-world sparse observation reconstruction. On simulation data, REACT reduces pH NRMSE by 14.7% and chemical consistency error by 24.0% over the best baseline.
- Cross-temporal-scale evaluations and ablation studies further validate the design of REACT, showing robustness against coarse-to-fine error accumulation and confirming the effectiveness of source correction, mass conservation and chemical closure. Additional transport-state comparisons, observational validation, and sensitivity analyses further assess the modeling choices and practical robustness.

2 Related Work

Assimilation and Inversion Models. Traditional data assimilation and inverse modeling methods, including optimal interpolation [17], 3D/4D-Var [18], ensemble Kalman filtering [19], and variational inverse models [20], reconstruct continuous ocean states by combining sparse observations with background dynamics. They provide strong physical consistency, but depend on iterative optimization, prescribed error assumptions, and repeated numerical model runs, making long-term high-resolution reconstruction *computationally costly*.

Black-box Models. Data-driven reconstruction methods replace expensive model integration with learned mappings. CNN-based models such as U-Net [5] and transformer-based models such as ViT [6] and Earthformer [7] capture spatial or spatiotemporal correlations from gridded fields. General-purpose inpainting methods based on GANs [8] and diffusion models [9] improve generic completion quality, while foundation models such as Prithvi [10] and AtmoRep [11] learn representa-

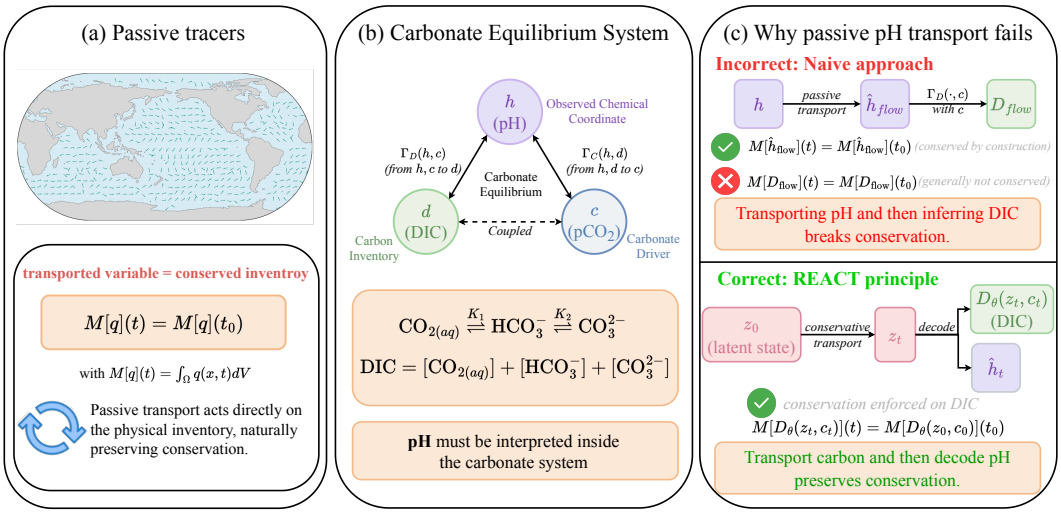

Figure 1: **Conservation-aware transport of pH in the carbonate system.** (a) Passive tracers are transported as conserved physical inventories and therefore obey conservation laws. (b) In contrast, pH is an equilibrium-dependent chemical coordinate coupled to carbonate variables such as DIC and $p\mathrm{CO}_2$. (c) Treating pH as a passive tracer preserves only the transported pH field and does not generally conserve the associated carbon inventory, while introducing a latent conserved state that decodes to both pH and DIC enables transport consistent with DIC conservation.

tions from large-scale weather and climate data. However, these models learn physical structure only implicitly from data, *without enforcing transport, conservation, or carbonate closure*. They can fill missing fields, but they do not guarantee physically faithful ocean reconstruction.

Physics-constrained Models. Physics-constrained models further introduce physical priors into learning. Neural operators such as FNO [12], neural field models such as GridMix [13], and implicit neural representations [14] model PDE-governed fields, while 4DVarNet [15] introduces variational consistency and NUTS [16] explicitly models transport dynamics for ocean reconstruction. However, these methods place dynamics on the reconstructed variable itself. This works for passive tracers, but *fails for active carbonate tracers*: pH is the target, while carbon inventory is the physical state.

3 Problem Formulation: Active Tracer Reconstruction

Notations. Let $\Omega \subset S^2$ denote the ocean surface, with spatial coordinate x and time $t \in [0, T]$. For a field q, write $\dot{q} = \partial q/\partial t$, and use $\nabla\cdot$, ∇, and ∇^2 for spherical divergence, gradient, and Laplacian. Let u be the surface velocity, K the diffusion tensor, h the surface pH, d dissolved inorganic carbon (DIC), and c surface $p\mathrm{CO}_2$. We write $M[r](t) = \int_\Omega r(x,t)\,dx$ for spatial mass, and $\mathcal{I} = \mathcal{Z} \times \mathcal{T}$ for the observed space–time sites.

Active vs. passive tracers: chemical consistency. As illustrated in Figure 1, pH is not a standalone transported scalar. It is a chemical coordinate of the carbonate system. Carbonate equilibrium gives two deterministic maps: $\Gamma_D(h,c)$ returns the DIC implied by pH and $p\mathrm{CO}_2$, while $\Gamma_C(h,d)$ returns the $p\mathrm{CO}_2$ implied by pH and DIC. Therefore, a reconstructed pH field $\hat{h}$ is chemically meaningful only when $\Gamma_D(\hat{h}, c) \approx d$ and $\Gamma_C(\hat{h}, d) \approx c$. We measure this disagreement on the observation set $\mathcal{I}$:

$$\mathcal{E}_{\text{chem}}(\hat{h}) = \|\Gamma_D(\hat{h}, c) - d\|^2_{L^2(\mathcal{I})} + \beta\|\Gamma_C(\hat{h}, d) - c\|^2_{L^2(\mathcal{I})}. \tag{1}$$

Because carbonate equilibrium is nonlinear, low pH RMSE does not guarantee low DIC or $p\mathrm{CO}_2$ error. The reconstructed pH must therefore be checked inside the carbonate system.

Active vs. passive tracers: physical consistency. Figure 1 (c) shows the corresponding conservation mismatch: the source-free conserved inventory is DIC, not pH. Under no-flux boundaries, source-free DIC transport should satisfy $M[d](t) = M[d](t_0)$. Therefore, for a flow-only pH prediction $\hat{h}_{\text{flow}}$, the relevant constraint is $M[\Gamma_D(\hat{h}_{\text{flow}}, c)](t) = M[\Gamma_D(\hat{h}_{\text{flow}}, c)](t_0)$, not conservation of pH or H^+. The full carbonate system may change through biology, air–sea exchange, vertical mixing, and unresolved reactions, but transport alone should not create or destroy carbon.

Reconstruction problem. Given sparse pH observations $\mathcal{O}_h$ on $\mathcal{I}$, velocity u, reference $p\mathrm{CO}_2$ c, and reference DIC d, we reconstruct h by minimizing $\mathcal{L}_{\text{pH}} + \lambda_{\text{chem}}\mathcal{E}_{\text{chem}} + \lambda_{\text{phys}}\mathcal{E}_{\text{phys}}$, where $\mathcal{L}_{\text{pH}} =$

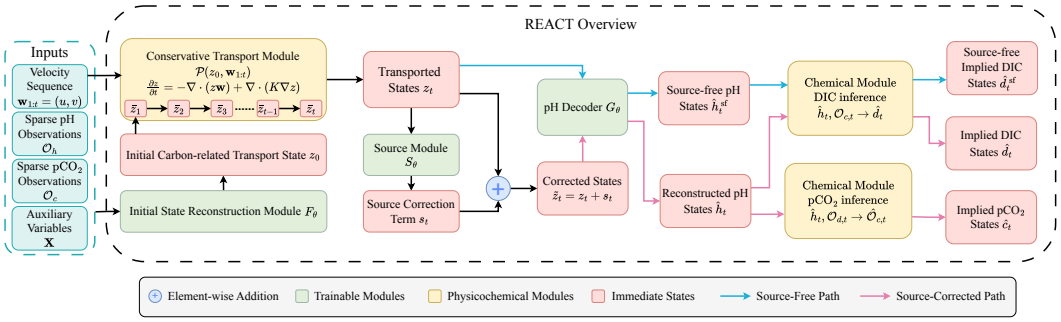

Figure 2: **Overview of REACT**. REACT initializes a carbon-related state from sparse observations and physical drivers, propagates it through a conservative transport module, adds learned active source corrections, and decodes the corrected state into pH. The reconstructed pH is checked by carbonate chemistry, while the source-free branch is constrained to conserve represented DIC mass.

$\|\hat{h} - h\|^2_{L^2(\mathcal{I})}$ and $\mathcal{E}_{\text{phys}} = \int_{t_0}^{T} |M[\Gamma_D(\hat{h}_{\text{flow}}, c)](t) - M[\Gamma_D(\hat{h}_{\text{flow}}, c)](t_0)|^2 dt$. This formulation separates the predicted variable from the conserved inventory: pH is reconstructed, but the carbon state implied by pH and $p\text{CO}_2$ must remain chemically and physically valid.

4 Methodology

4.1 A Naive Passive-Tracer Baseline for pH Reconstruction

Naive passive pH reconstruction. A natural baseline is to treat pH as a transported scalar. Let F_0 be a data-driven initializer and let ξ denote auxiliary fields. Given sparse pH observations $\mathcal{O}_h$, velocity u, reference $p\text{CO}_2$ c, and reference DIC d, the model first estimates the initial field by $\hat{h}(\cdot, t_0) = F_0(\mathcal{O}_h, u, c, d, \xi)$. It then evolves $\hat{h}$ by the passive advection–diffusion equation $\dot{\hat{h}} + \nabla \cdot (u\hat{h}) - \nabla \cdot (K \nabla \hat{h}) = s_h$, where K and s_h are learned. The model is trained only by scalar pH error, $\mathcal{L}_{\text{pH}} = \|\hat{h} - h\|_2^2$. This baseline is stronger than frame-wise inpainting because it uses temporal propagation, velocity fields, and a PDE-based inductive bias.

Why it fails. The failure is structural: the model transports pH, but source-free carbon conservation is about DIC. Let $\hat{h}_{\text{flow}}$ be the flow-only prediction and $D_{\text{flow}} = \Gamma_D(\hat{h}_{\text{flow}}, c)$ be its implied DIC. Even if passive transport conserves the transported scalar, $M[\hat{h}_{\text{flow}}](t) = M[\hat{h}_{\text{flow}}](t_0)$ does not imply $M[D_{\text{flow}}](t) = M[D_{\text{flow}}](t_0)$. Thus the passive baseline can produce smooth, low-RMSE pH fields while still violating carbonate closure and the DIC budget. Appendix B.4 makes this failure formal: even in a source-free, no-flux setting, passive pH transport conserves $M[\hat{h}_{\text{flow}}]$ but generically does not conserve $M[\Gamma_D(\hat{h}_{\text{flow}}, c)]$, because the carbonate map $\Gamma_D(\cdot, c)$ is nonlinear.

4.2 REACT: REconstructing Active Carbonate Tracers

Model overview. We propose **REACT**, short for **RE**constructing **A**ctive **C**arbonate **T**racers. As shown in Figure 2, REACT follows a carbon-first pipeline. It takes velocity sequences, sparse pH observations, sparse $p\text{CO}_2$ observations, and auxiliary variables as inputs. It first reconstructs an initial carbon-related transport state z_0, propagates it by a conservative PDE to obtain transported states z_t, adds an active source correction s_t to form corrected states $\tilde{z}_t = z_t + s_t$, and decodes the corrected states into reconstructed pH $\hat{h}_t$. A chemical module then maps $\hat{h}_t$ back to implied DIC through carbonate equilibrium. The key difference from passive-tracer reconstruction is therefore not the use of a PDE, but what happens around the PDE: the transported state is carbon-related, the residual is an active carbon-cycle source, and the output is checked by carbonate chemistry.

Latent carbonate-state transport and active correction. REACT consists of three coupled components. (1) *Latent carbonate-state initialization*: REACT estimates an initial latent carbonate state $z_0 = F_\theta(\mathcal{O}_h, \mathcal{O}_c, u, \xi)$, where $\mathcal{O}_h$ and $\mathcal{O}_c$ are sparse pH and $p\text{CO}_2$ observations, and ξ denotes auxiliary fields. The state z_0 is not a pH image or raw DIC field; it is a latent coordinate on which conservative carbonate transport is imposed. (2) *Conservative transport*: starting from $z(\cdot, t_0) = z_0$,

REACT evolves z by $\partial_t z + \nabla \cdot (uz) - \nabla \cdot (K_\theta \nabla z) = 0$, where K_θ is learned and $z_t = z(\cdot, t)$. This gives the source-free branch for conservative redistribution. (3) *Active correction and pH decoding*: because the surface carbonate system also changes through air–sea exchange, biological activity, vertical mixing, and unresolved reactions, REACT predicts an active correction $s_t = S_\theta(z_t, u, c, \xi)$, forms $\tilde{z}_t = z_t + s_t$, and decodes pH by $\hat{h}_t = G_\theta(\tilde{z}_t, c, \xi)$. Thus the PDE branch explains conservative transport, while the source module explains non-conservative carbonate-system changes.

Represented DIC and source-free conservation. Since z_t may be latent, we do not require it to equal DIC in raw units. We require only that it has a DIC readout after decoding and carbonate equilibrium, defined by $\mathcal{D}_\theta(z_t, c_t) = \Gamma_D(G_\theta(z_t, c_t), c_t)$. In a DIC-coordinate implementation, $\mathcal{D}_\theta(z_t, c_t)$ reduces to z_t. The corrected state $\tilde{z}_t$ need not conserve mass because it contains active sources, but the source-free branch z_t should conserve represented DIC. We therefore impose $\mathcal{L}_{\text{phys}} = \int_{t_0}^{T} |M[\mathcal{D}_\theta(z_t, c_t)] - M[\mathcal{D}_\theta(z_0, c_{t_0})]|^2 dt$. This placement is crucial: conservation before the source module forces the PDE branch to carry conservative carbon transport, instead of allowing the source module to hide transport errors.

Training objective. REACT is trained with pH supervision, carbonate closure, and source-free DIC conservation. The pH loss is $\mathcal{L}_{\text{pH}} = \|\hat{h} - h\|^2_{L^2(\mathcal{I})}$, and the chemical loss is $\mathcal{L}_{\text{chem}} = \mathcal{E}_{\text{chem}}(\hat{h})$, which checks whether reconstructed pH implies the correct DIC when paired with $p\text{CO}_2$, and the correct $p\text{CO}_2$ when paired with DIC. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pH}} + \lambda_{\text{chem}}\mathcal{L}_{\text{chem}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}$. When the losses are small, the prediction is not only accurate in pH, but also chemically closed and physically conservative in the source-free branch.

Advantage over the passive baseline. REACT follows the PDE-plus-correction principle used in models such as NUTS [16]: both use dynamical propagation and learned residual modules rather than frame-wise inpainting. The difference is the tracer. For passive nutrients, the reconstructed variable is also the transported inventory, so PDE propagation can act directly on the target field. For carbonate reconstruction, the target is pH but the conserved inventory is DIC. A passive pH model therefore has the right dynamical tool but the wrong state variable. REACT keeps the useful part of PDE-based reconstruction while changing the transported state from pH to represented carbon, and adds carbonate closure to ensure that the decoded pH remains chemically meaningful.

Relation to existing reconstruction models. Existing reconstruction methods can be viewed along two axes, with carbonate-aware reconstruction exposing a third missing axis. Frame-wise interpolation or inpainting ignores transport and treats each map independently [17, 5, 6, 8, 9]. PDE-based reconstruction introduces temporal dynamics, but usually assumes the reconstructed variable is itself the transported tracer [18–20, 12–16]. These methods do not directly address pH as an active carbonate-system variable: chemical closure is not part of the reconstruction target, and conservative transport is not explicitly separated from active carbon-cycle sources. REACT takes one step further: it places the PDE on a latent carbon state and evaluates decoded pH through carbonate closure.

4.3 Theoretical Analysis: Why REACT Uses a Carbon State

The theoretical question is whether the hidden state z_t carries a physical certificate. REACT does not require z_t to equal raw DIC, since a latent model may represent the same carbon inventory in another coordinate system. What matters is the DIC readout induced by the decoder and carbonate chemistry: $\mathcal{D}_\theta(z_t, c_t) = \Gamma_D(G_\theta(z_t, c_t), c_t)$. We therefore interpret z_t as a carbon coordinate and $\mathcal{D}_\theta(z_t, c_t)$ as its represented DIC.

Theorem 1 (Informal; carbon-state certificate). Under the standard identifiability and realizability conditions stated in Appendix B, REACT gives the following certificate. If $\mathcal{L}_{\text{chem}} = 0$ and $\mathcal{L}_{\text{phys}} = 0$, then the reconstructed pH closes carbonate chemistry, i.e., $\Gamma_D(\hat{h}, c) = d$ and $\Gamma_C(\hat{h}, d) = c$, and the source-free branch conserves represented DIC mass, i.e., $M[\mathcal{D}_\theta(z, c)](t) = M[\mathcal{D}_\theta(z, c)](t_0)$. Moreover, when the true process decomposes into source-free DIC transport plus an active carbon residual, REACT can realize zero chemical and physical loss. A passive pH-transport model cannot generally satisfy the same certificate: even if it conserves source-free pH mass, it can change the DIC mass implied by carbonate chemistry.

Interpretation. The result explains why REACT transports a carbon state rather than pH itself. For passive tracers, the prediction target is also the conserved inventory. For carbonate tracers, this identity breaks: pH is the target, but DIC is the conserved inventory. Since the carbonate map

$D(h) = \Gamma_D(h, c)$ is nonlinear, conserving pH does not generally conserve DIC. REACT resolves this mismatch by placing the PDE on a carbon state, assigning active carbon-cycle changes to the source module, and decoding pH through carbonate closure. The full proof and the passive-baseline counterexample are given in Appendix B.

5 Experiment

In this section, we evaluate REACT through four questions:

RQ1. Does REACT improve global sea surface pH reconstruction across temporal scales and real observations? Does it also improve carbonate-system consistency?

RQ2. Do the mass and chemical constraints make the reconstruction physically and chemically valid? How much do they reduce DIC drift and carbonate error?

RQ3. How do different design choices, such as auxiliary variable inputs, model depth, and reconstruction window length, affect model performance?

RQ4. Why must REACT transport a carbon state rather than pH?

5.1 Experimental Setup

In this subsection, we describe the datasets, tasks, baselines and metrics. Additional details are provided in Appendix D.

Data. We use two types of datasets in this work: a gridded simulation dataset derived from MOM6–COBALTv2 [21] and a sparse observational dataset in the World Ocean Database (WOD) [1] and the Surface Ocean CO_2 Atlas (SOCAT) [22]. **Simulation Dataset.** The simulation data come from daily MOM6–COBALTv2 outputs from 1959 to 2022. All variables are regridded to a uniform $1^\circ \times 1^\circ$ global grid using bilinear interpolation, resulting in a spatial resolution of 180×360. We aggregate the daily outputs into weekly and monthly products, yielding two cross-scale tasks: month-to-week and week-to-day reconstruction. The dataset provides sea surface pH as the reconstruction target, together with nine auxiliary variables used as model inputs: sea surface CO_2 partial pressure ($p\mathrm{CO}_2^s$), atmospheric CO_2 partial pressure ($p\mathrm{CO}_2^a$), chlorophyll concentration (Chl.), sea surface height (SSH), mixed layer depth (MLD), salinity (Salt), temperature (Temp.), and horizontal velocity components (u and v). **Observational data.** The pH observations are from the WOD, while $p\mathrm{CO}_2^s$ observations are from the SOCAT, both from 1982 to 2022. The records are highly sparse: valid entries cover only 0.11% of the pH grid and 1.67% of the $p\mathrm{CO}_2^s$ grid. All observations are mapped onto the same spatiotemporal grid as MOM6. We additionally evaluate pH reconstruction using GLODAP observations, with auxiliary fields supplied by MOM6.

Tasks. We evaluate the proposed model on two types of tasks: simulation-based and observation-based reconstruction. **Simulation-based reconstruction.** We consider two cross-temporal-scale settings, monthly-to-weekly and weekly-to-daily sea surface pH reconstruction. Given sparse coarse-scale pH and $p\mathrm{CO}_2^s$ observations, together with target-scale auxiliary variables, the model reconstructs the full target-scale pH field. We sample the full MOM6 labels at 0.11%, 1%, and 10%, where 0.11% matches the sparsity of WOD pH observations and 10% follows prior sparse-field reconstruction settings.

Observation-based reconstruction. We use WOD pH and SOCAT $p\mathrm{CO}_2^s$ as sparse constraints, together with MOM6 auxiliary variables, to reconstruct sea surface pH under real observational sparsity. All tasks use chronological splits for training, validation, and testing, as summarized in Table 1.

Table 1: Overview of dataset divisions by year.

Task	Train	Val.	Test
Month→Week	1971–2010	1965–1970; 2011–2016	1959–1964; 2017–2022
Week→Day	1988–2003	1987; 2004	1986; 2005
WOD	1982–2018	2019–2020	2021–2022
GLODAP	1982–2018	2019–2020	2021–2022

Baselines. We compare REACT with seven classes of baselines: (1) spatial interpolation methods, including Kriging with exponential and spherical variogram models [23]; (2) convolutional networks, represented by U-Net [5]; (3) Transformer-based models, represented by ViT [6]; (4) neural operators, represented by FNO [12]; (5) implicit continuous field reconstruction models, represented by Marble [13]; (6) climate foundation models, including Prithvi WxC [10] and AtmoRep [11]; (7) physically guided reconstruction models, including 4DVarNet [15] and NUTS [16]. Kriging is

Table 2: **Main results.** Sea surface pH reconstruction on Month→Week and Week→Day tasks under different sampling ratios. pH NRMSE is reported in units of 10^{-4}, and Chem. is reported in units of 10^{-2}. Best and second-best results are shown in **bold** and underline, respectively. Subscript values denote standard errors over three independent sparse-observation samplings. *Params* denotes model size. *U-Net*^{2} denotes a two-level U-Net variant. *Gain* denotes relative improvement over the strongest baseline. Decomposed $p\mathrm{CO}_2^s$ and DIC errors are reported in Appendix Table 9.

Methods	Params	Month→Week						Week→Day					
		0.11%		1%		10%		0.11%		1%		10%	
		pH	Chem.	pH	Chem.	pH	Chem.	pH	Chem.	pH	Chem.	pH	Chem.
Krig-Exp	–	$50.6_{\pm 0.91}$	$10.1_{\pm 0.34}$	$37.0_{\pm 0.84}$	$7.06_{\pm 0.26}$	$23.3_{\pm 0.72}$	$4.56_{\pm 0.23}$	$51.2_{\pm 0.86}$	$9.95_{\pm 0.29}$	$37.1_{\pm 0.82}$	$6.94_{\pm 0.27}$	$22.2_{\pm 0.68}$	$4.22_{\pm 0.17}$
Krig-Sph	–	$50.9_{\pm 0.89}$	$10.2_{\pm 0.20}$	$41.2_{\pm 0.88}$	$7.83_{\pm 0.27}$	$33.8_{\pm 0.81}$	$6.11_{\pm 0.28}$	$52.2_{\pm 0.83}$	$10.2_{\pm 0.33}$	$39.4_{\pm 0.84}$	$7.35_{\pm 0.30}$	$26.9_{\pm 0.66}$	$5.04_{\pm 0.19}$
4D-Var.	0.27M	$6.92_{\pm 0.03}$	$2.09_{\pm 0.03}$	$7.14_{\pm 0.04}$	$2.00_{\pm 0.02}$	$5.99_{\pm 0.06}$	$1.90_{\pm 0.04}$	$5.93_{\pm 0.07}$	$1.86_{\pm 0.03}$	$6.02_{\pm 0.05}$	$1.88_{\pm 0.03}$	$7.40_{\pm 0.05}$	$2.05_{\pm 0.04}$
Marble	0.83M	$25.7_{\pm 0.03}$	$5.65_{\pm 0.10}$	$25.8_{\pm 0.03}$	$5.71_{\pm 0.07}$	$25.2_{\pm 0.02}$	$5.51_{\pm 0.07}$	$26.9_{\pm 0.04}$	$5.51_{\pm 0.06}$	$26.9_{\pm 0.02}$	$5.55_{\pm 0.03}$	$27.0_{\pm 0.03}$	$5.51_{\pm 0.05}$
FNO	8.42M	$16.7_{\pm 0.07}$	$3.88_{\pm 0.07}$	$16.3_{\pm 0.04}$	$3.81_{\pm 0.02}$	$16.6_{\pm 0.05}$	$3.82_{\pm 0.02}$	$7.34_{\pm 0.05}$	$2.07_{\pm 0.04}$	$8.81_{\pm 0.06}$	$2.29_{\pm 0.02}$	$8.67_{\pm 0.05}$	$2.24_{\pm 0.02}$
U-Net	31.1M	$7.78_{\pm 0.04}$	$2.17_{\pm 0.08}$	$7.56_{\pm 0.05}$	$2.14_{\pm 0.01}$	$7.40_{\pm 0.04}$	$2.12_{\pm 0.03}$	$5.91_{\pm 0.06}$	$1.86_{\pm 0.05}$	$5.81_{\pm 0.05}$	$1.85_{\pm 0.03}$	$5.80_{\pm 0.05}$	$1.84_{\pm 0.03}$
U-Net^2	1.96M	$7.71_{\pm 0.13}$	$2.21_{\pm 0.03}$	$7.78_{\pm 0.09}$	$2.22_{\pm 0.02}$	$7.22_{\pm 0.12}$	$2.11_{\pm 0.03}$	$7.21_{\pm 0.14}$	$2.02_{\pm 0.04}$	$7.29_{\pm 0.12}$	$2.06_{\pm 0.02}$	$6.81_{\pm 0.08}$	$1.93_{\pm 0.02}$
ViT	0.11B	$9.56_{\pm 0.10}$	$2.51_{\pm 0.01}$	$9.47_{\pm 0.04}$	$2.54_{\pm 0.01}$	$9.26_{\pm 0.06}$	$2.50_{\pm 0.01}$	$6.31_{\pm 0.06}$	$1.66_{\pm 0.03}$	$6.26_{\pm 0.04}$	$2.18_{\pm 0.03}$	$5.85_{\pm 0.06}$	$1.88_{\pm 0.02}$
NUTS	0.15B	$6.74_{\pm 0.06}$	$1.95_{\pm 0.10}$	$6.88_{\pm 0.08}$	$2.06_{\pm 0.12}$	$6.87_{\pm 0.05}$	$2.30_{\pm 0.09}$	$5.08_{\pm 0.09}$	$1.69_{\pm 0.07}$	$4.91_{\pm 0.07}$	$1.65_{\pm 0.05}$	$5.05_{\pm 0.05}$	$1.77_{\pm 0.04}$
AtmoR.	0.80B	$14.6_{\pm 0.16}$	$3.52_{\pm 0.04}$	$16.2_{\pm 0.19}$	$3.95_{\pm 0.04}$	$11.0_{\pm 0.17}$	$2.86_{\pm 0.03}$	$9.10_{\pm 0.12}$	$2.35_{\pm 0.05}$	$7.55_{\pm 0.14}$	$2.07_{\pm 0.03}$	$7.60_{\pm 0.11}$	$2.05_{\pm 0.01}$
Prithvi	2.37B	$16.0_{\pm 0.52}$	$3.48_{\pm 0.07}$	$16.1_{\pm 0.47}$	$3.45_{\pm 0.03}$	$14.2_{\pm 0.59}$	$3.16_{\pm 0.05}$	$13.3_{\pm 0.29}$	$2.75_{\pm 0.04}$	$13.4_{\pm 0.36}$	$2.75_{\pm 0.02}$	$12.8_{\pm 0.44}$	$2.73_{\pm 0.07}$
REACT	0.18B	$\mathbf{5.84}_{\pm 0.04}$	$\mathbf{1.59}_{\pm 0.04}$	$\mathbf{5.87}_{\pm 0.03}$	$\mathbf{1.52}_{\pm 0.02}$	$\mathbf{5.81}_{\pm 0.02}$	$\mathbf{1.54}_{\pm 0.03}$	$\mathbf{4.64}_{\pm 0.04}$	$\mathbf{1.35}_{\pm 0.03}$	$\mathbf{4.62}_{\pm 0.03}$	$\mathbf{1.34}_{\pm 0.02}$	$\mathbf{4.54}_{\pm 0.02}$	$\mathbf{1.32}_{\pm 0.02}$
Gain	–	13.4%	18.5%	14.7%	24.0%	3.00%	18.9%	8.66%	18.7%	5.90%	18.8%	10.1%	25.4%

applied independently at each target time slice using only sparse pH observations. All other baselines use the same temporal reconstruction windows, data splits, sampling ratios, and metrics, with sparse observations, auxiliary variables, and flow fields as inputs. Marble uses temporal observations but does not use explicit flow or auxiliary variables.

Metrics. We evaluate model performance using four metrics. **pH** is the latitude-weighted NRMSE of sea surface pH, where the latitude weighting accounts for unequal grid-cell areas. NRMSE is obtained by normalizing RMSE with the mean value of the corresponding target variable. **Chem.** measures carbonate closure, defined as the average of normalized $p\mathrm{CO}_2^s$ error and normalized DIC error. $\mathbf{Drift}_{\mathrm{local}}$ measures short term mass conservation behavior by computing the DIC inventory drift within each reconstruction segment. $\mathbf{Drift}_{\mathrm{global}}$ measures long range mass conservation behavior by computing the accumulated DIC inventory drift over continuous long term reconstruction periods. The detailed computation of all metrics is provided in Appendix D.2.

5.2 Main Results (RQ1)

In this subsection, we present the main results.

Obs 1: REACT improves both pH accuracy and carbonate closure across temporal scales. As shown in Table 2, REACT is the best method on both month-to-week and week-to-day reconstruction, improving both pH NRMSE and Chem. In terms of pH NRMSE, REACT reduces the error by 3.00%–14.7% on monthly-to-weekly reconstruction and by 5.90%–10.1% on weekly-to-daily reconstruction. This suggests that REACT better controls error propagation when refining coarse temporal inputs into finer pH fields. Figure 3 further shows smaller spatial errors than NUTS, especially in the Arctic and Pacific regions. The gains in Chem. are even larger. REACT improves Chem. by 18.5%–24.0% on the monthly-to-weekly task and by 18.7%–25.4% on the weekly-to-daily task. This shows that REACT not only reduces pointwise pH error, but also produces pH fields that are more consistent with the associated $p\mathrm{CO}_2^s$ and DIC fields. The gain comes from modeling pH as an active carbonate tracer rather than a standalone scalar field.

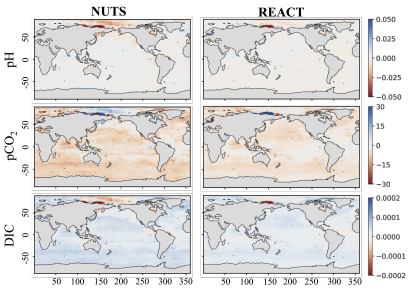

Figure 3: **Qualitative error maps.** Spatial reconstruction errors on MOM6, computed as prediction minus ground truth. Additional qualitative results are provided in Appendix E.2.

Table 3: **Component analysis.** Ablation of the source module, DIC mass constraint, and carbonate chemical constraint. We report results for REACT on both reconstruction tasks and repeat the constraint ablation on U-Net for month-to-week reconstruction at 0.11% sampling. pH NRMSE is in units of 10^{-4}, Chem. is in units of 10^{-2}, and drift is reported in per mille. *Div.* denotes divergent long-range drift. Decomposed pCO_2^s and DIC errors are provided in Appendix Table 10. For REACT, the first row contains only the PDE branch, the second adds the source module, and the third additionally introduces the decoder and mass-conservation loss; the fourth also includes the chemical constraint.

			REACT (Month→Week)				REACT (Week→Day)				U-Net (Month→Week)			
Source	$\mathcal{L}_{\text{mass}}$	$\mathcal{L}_{\text{chem}}$	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
–	–	–	40.8	Div.	–	–	11.8	2.73	–	–	7.74	2.12	–	–
✓	–	–	5.79	1.93	8.84‰	Div.	4.03	1.65	1.11‰	Div.	7.75	2.14	21.2‰	17.3‰
✓	✓	–	5.68	1.94	2.26‰	9.73‰	4.27	1.69	0.57‰	7.41‰	7.78	2.03	4.24‰	3.34‰
✓	✓	✓	5.84	1.59	2.21‰	10.3‰	4.64	1.35	0.63‰	7.68‰	7.92	1.65	5.53‰	4.88‰

Obs 2: REACT transfers to real sparse observations. We evaluate real sparse reconstruction using WOD pH and SOCAT pCO_2^s. After chronological train/validation/test splitting, we use 75% of the available observations within each subset as visible inputs and hold out the remaining 25%. As shown in Figure 4, REACT achieves a pH NRMSE of 5.25×10^{-3}, improving over U-Net, the strongest baseline, which obtains 5.59×10^{-3}. This corresponds to an error reduction of approximately 6.08%. Meanwhile, REACT achieves a Chem. score of 0.102, also outperforming 4DVarNet with 0.113, corresponding to an improvement of approximately 9.73%. Thus, the advantage of REACT is not limited to regular simulation grids; it also holds under sparse and irregular real observations. On GLODAP, REACT achieves a pH NRMSE of 8.25×10^{-3} and a Chem. score of 1.55×10^{-1}, reducing the respective errors by 14.9% relative to NUTS and 18.3% relative to U-Net. Full results and additional observational robustness analyses are provided in Appendix E.8.

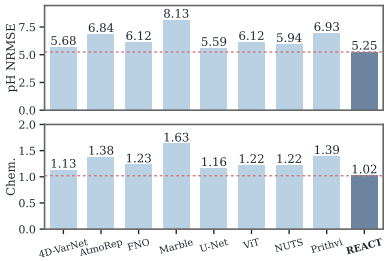

Figure 4: **Real-observation results.** pH NRMSE and Chem. measure are reported in units of 10^{-3} and 10^{-1}, respectively. Lower is better. Decomposed pCO_2^s and DIC results are provided in Appendix E.3.

5.3 Component Analysis (RQ2)

We isolate three active-tracer components: source correction, DIC mass conservation and carbonate chemical closure. They test whether REACT can capture non-conservative pH changes, preserve DIC inventory, and close the carbonate system beyond pointwise pH fitting. Table 3 reports the ablations, Figure 5 shows DIC drift and Figure 6 evaluates uncertainty calibration.

Obs 3: The source module is necessary for accurate pH reconstruction. Table 3 shows that physical propagation alone is insufficient: it captures flow-driven transport but misses non-conservative effects from air–sea exchange, biology, and local model error. With the source term correction module, the model can compensate for residual errors in the propagated intermediate field, leading to a clear reduction in pH error. On month-to-week reconstruction at 0.11% sampling, the pH error drops from 4.08×10^{-3} to 5.79×10^{-4}. This result suggests that sea surface pH reconstruction needs active correction beyond conservative transport.

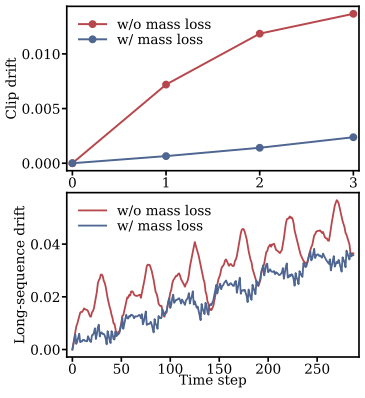

Figure 5: Effect of mass-conservation training on DIC drift.

Obs 4: The mass constraint suppresses DIC drift without hurting pH accuracy. As shown in Table 3 and Figure 5, since pH is not a mass-conservative variable, directly computing inventory drift on the pH field does not have a clear physical interpretation. Therefore, we evaluate mass conservation on the DIC field implied by carbonate chemistry. On month-to-week reconstruction, the mass constraint reduces local drift from 8.84‰ to 2.26‰, and turns global drift from divergence into 9.73‰. Meanwhile, the pH NRMSE slightly decreases from 5.79×10^{-4} to 5.68×10^{-4}. A consistent pattern is observed in the U-Net ablation, where adding the mass constraint

Table 4: **Ablation Study**. (a) Evaluation of reconstruction interval length; (b) analysis of state initializer depth; (c) analysis of pH decoder/source-module depth; (d) comparison of state initializer architectures; (e) quantification of the impact of auxiliary input variables. pH NRMSE is in units of 10^{-4}, Chem. is in units of 10^{-2}, and drift is reported in per mille. All experiments follow the default setting: Conducted on the Monthly-to-Weekly task under sampling ratio of 0.11%, using a 12-layer ViViT [25] as state initializer, pH decoder/source module of depth 4 and interval length set to 4 and 2 for Monthly-to-Weekly and Weekly-to-Daily, respectively. Details are provided in Table 11 and Table 12.

(a) Interval Length.

case	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
1	5.85	1.57	2.10‰	Div.
2	6.52	1.51	2.90‰	Div.
4	5.84	1.59	2.21‰	10.3‰
8	6.09	1.58	2.44‰	10.2‰

(b) Depth of State Initializer.

depth	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
4	6.73	1.59	2.98‰	11.0‰
6	6.16	1.60	2.30‰	8.62‰
12	5.84	1.59	2.21‰	10.3‰
24	5.98	1.59	8.75‰	19.5‰

(c) Depth of pH Decoder/Source Module.

depth	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
2	5.75	1.54	8.41‰	19.1‰
3	5.88	1.54	7.91‰	18.4‰
4	5.84	1.59	2.21‰	10.3‰
5	7.51	1.85	2.75‰	11.4‰

(d) State Initializer Structure.

case	pH	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
ViT	6.58	1.53	3.74‰	Div.
2D CNN	5.96	1.67	8.81‰	537‰
3D CNN	36.6	462	8.44‰	Div.
ViViT	5.84	1.59	2.21‰	10.3‰

(e) Auxiliary Variables.

Task	Metric	pCO_2^s	pCO_2^a	Chl.	SSH	MLD	Salt	Temp.	u	v
Monthly to Weekly	pH	21.7	6.98	5.97	6.18	5.93	5.95	6.05	6.17	6.32
	Chem.	4.97	1.55	1.73	1.58	1.57	1.70	1.61	1.60	1.73
Weekly to Daily	pH	11.9	4.81	4.79	4.76	4.88	4.69	4.77	4.78	4.75
	Chem.	2.57	1.33	1.30	1.26	1.25	1.40	1.29	1.26	1.31

reduces local drift from $21.2‰$ to $4.24‰$ and global drift from $17.3‰$ to $3.34‰$. This confirms that the constraint improves physical validity without sacrificing pH accuracy.

Obs 5: The chemical constraint improves carbonate closure with a small accuracy trade-off. Table 3 shows that optimizing pH alone can give accurate but chemically inconsistent predictions. After introducing the chemical module, the overall chemical consistency metric improves noticeably, indicating improved consistency between the predicted pH field and the carbonate system. For example, on month-to-week reconstruction, the chemical metric improves from 1.94×10^{-2} to 1.59×10^{-2}, while pH error changes only slightly from 5.68×10^{-4} to 5.84×10^{-4}. The U-Net ablation exhibits the same behavior, with the chemical constraint improving Chem. from 2.03×10^{-2} to 1.65×10^{-2}, accompanied by only a small increase in pH NRMSE. This suggests that the chemical constraint trades a small amount of pointwise accuracy for a more physically valid carbonate state.

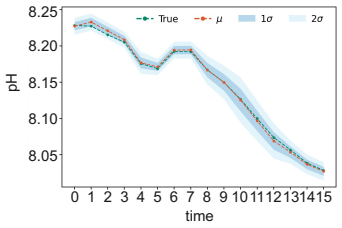

Figure 6: Uncertainty estimation of REACT.

Obs 6: REACT provides calibrated predictive uncertainty. Following ClimODE [24], we add a lightweight uncertainty head on top of the converged deterministic REACT model to estimate the predictive standard deviation, while keeping the predictive mean fixed. As shown in Figure 6, the resulting uncertainty band covers most ground-truth pH trajectories and remains concentrated around the deterministic prediction. The selected checkpoint achieves two-standard-deviation coverage values of 0.963, 0.971, and 0.958 for pH, pCO_2^s, and DIC, respectively, close to the nominal Gaussian coverage of 0.95. This suggests that REACT can provide calibrated predictive uncertainty in addition to accurate deterministic reconstruction. Results for pCO_2^s and DIC, together with training details, are provided in Appendix E.7.

5.4 Ablation Study on Design Choices (RQ3)

This section analyzes how different design choices affect model performance, including the reconstruction horizon, spatiotemporal encoder depth, source and chemical mapping module depth, spatiotemporal encoder architecture, and auxiliary variable inputs. • **Reconstruction Horizon.** Table 4a shows that the reconstruction horizon strongly affects performance. A short horizon can lead to relatively small local mass drift, but it exhibits instability during long-range propagation, whereas an overly long horizon increases error accumulation and optimization difficulty. • **Spatiotemporal Encoder Depth.** Table 4b shows that a moderate encoder depth is preferable. A depth of 4 leads to underfitting, with a pH error of 6.73×10^{-4}, whereas a depth of 12 achieves the best error of 5.84×10^{-4}. Further increasing the depth to 24 does not improve accuracy; instead, it worsens both local and global drift, suggesting that an overly deep encoder may introduce training instability or excessive smoothing. • **Source and Chemistry Module Depth.** Table 4c shows that the U-Net depth

affects both accuracy and conservation. Although a depth of 2 yields a low pH error of 5.75×10^{-4}, its local and global drift increase to $8.41‰$ and $19.1‰$. A depth of 4 provides a better trade-off, reducing the drift to $2.21‰$ and $10.3‰$ while maintaining low pH error. • **Spatiotemporal Encoder Architecture.** Table 4d compares different encoder architectures. The 2D CNN and ViT achieve reasonably low pointwise errors, but exhibit very large global drift, indicating that the lack of explicit temporal modeling leads to severe instability in long-sequence reconstruction. The 3D CNN clearly fails on this task. ViViT achieves the best pH error of 5.84×10^{-4} with lower drift, demonstrating the benefit of explicit spatiotemporal attention. • **Auxiliary Variable Contribution.** Table 4e shows that $p\mathrm{CO}_2^s$ is the most important auxiliary variable. Removing $p\mathrm{CO}_2^s$ increases the pH error from 5.84×10^{-4} to 2.17×10^{-3}, indicating its central role in providing carbonate system constraints. Other variables also contribute to reconstruction performance.

5.5 Transport-State Comparisons (RQ4)

Table 5 compares REACT with pH alignment and direct transport of pH or DIC, including a direct-pH control with matched parameter count. Initialization, masking, and architectural changes are described in Appendix D.3. Soft pH alignment worsens both pH reconstruction and Chem. Direct pH transport causes a much larger increase in Chem. and final-DIC error, which is not resolved by increasing decoder capacity. In contrast, changing the directly transported variable from pH to DIC restores performance comparable to REACT, with no extreme pH or DIC values during training or validation. These comparisons support transporting a carbon-related state rather than pH.

Table 5: **Transport-state comparisons.** pH NRMSE is in units of 10^{-4}; Chem. and the remaining NRMSEs are in 10^{-2}. REACT transports a latent carbon state; pH alignment adds a loss aligning it with source-free decoded pH; Direct pH/DIC transport pH/DIC from exact initial fields. $^{\diamond}$: decoder enlarged to approximately match REACT's parameter count.

Variant	pH	Chem.	$p\mathrm{CO}_2^s$	Final DIC
REACT	5.84	1.59	1.73	1.45
pH alignment	6.58	1.69	–	–
Direct pH	6.46	13.0	1.62	24.4
Direct pH$^{\diamond}$	6.34	65.1	1.56	129
Direct DIC	5.67	1.57	1.66	1.49

Despite stable training, direct pH transport produces localized pH extremes that disproportionately amplify DIC errors through nonlinear carbonate chemistry. These outliers have a smaller effect on $p\mathrm{CO}_2^s$, making its aggregate error less sensitive to this failure. The stable direct-DIC control supports transporting a carbon state to maintain physical consistency.

6 Conclusion

We presented REACT, an active-tracer reconstruction framework for global sea surface pH that models pH through an underlying carbonate-system state. By combining conservative transport, source correction, and carbonate-equilibrium constraints, REACT achieves the best pH NRMSE and chemical consistency on multiple reconstruction tasks. An additional surface $p\mathrm{CO}_2^s$ reconstruction experiment demonstrates applicability to another target within the tested carbonate system (Appendix E.9). Extension to other tracer systems remains future work, together with the limitation-driven directions discussed in Appendix F.

Acknowledgments

This work was supported by the National Key R&D Program of China (2025YFF0516900 & 2025YFF0516904), NSFC U25B2039 and National Natural Science Foundation of China (No. 62306179). The authors declare no competing interests.

References

[1] Alexey V Mishonov, Tim P Boyer, Olga K Baranova, Courtney N Bouchard, Scott L Cross, Hernan E Garcia, Ricardo A Locarnini, Christopher R Paver, James R Reagan, Zhankun Wang, et al. World ocean database 2023. 2024.

[2] Richard E Zeebe and Dieter Wolf-Gladrow. *CO2 in seawater: equilibrium, kinetics, isotopes*, volume 65. Gulf Professional Publishing, 2001.

[3] Andrew G Dickson. The carbon dioxide system in seawater: equilibrium chemistry and measurements. *Guide to best practices for ocean acidification research and data reporting*, 1: 17–40, 2010.

[4] Matthew P Humphreys, Ernie R Lewis, Jonathan D Sharp, and Denis Pierrot. Pyco2sys v1. 8: Marine carbonate system calculations in python. *Geoscientific Model Development*, 15(1): 15–43, 2022.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL `https://arxiv.org/abs/1505.04597`.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[7] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

[8] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

[9] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[10] Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, et al. Prithvi wxc: Foundation model for weather and climate. *arXiv:2409.13598*, 2024.

[11] Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning. *arXiv:2308.13280*, 2023.

[12] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv:2010.08895*, 2020.

[13] Honghui Wang, Shiji Song, and Gao Huang. Gridmix: Exploring spatial modulation for neural fields in pde modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Xihaier Luo, Wei Xu, Yihui Ren, Shinjae Yoo, and Balu Nadiga. Continuous field reconstruction from sparse observations with implicit neural networks. *arXiv:2401.11611*, 2024.

[15] Maxime Beauchamp, Quentin Febvre, Hugo Georgenthum, and Ronan Fablet. 4dvarnet-ssh: end-to-end learning of variational interpolation schemes for nadir and wide-swath satellite altimetry. *Geoscientific Model Development*, 16(8):2119–2147, 2023.

[16] Hao Zheng, Shiyu Liang, Yuting Zheng, Chaofan Sun, Lei Bai, and Enhui Liao. Nuts: Eddy-robust reconstruction of surface ocean nutrients via two-scale modeling. *Advances in Neural Information Processing Systems*, 38:42530–42550, 2026.

[17] Margarita E Conkright, Ricardo A Locarnini, Hernan E Garcia, Todd D O'Brien, Timothy P Boyer, C Stephens, and John I Antonov. World ocean atlas 2001: Objective analyses, data statistics, and figures: Cd-rom documentation. 2002.

[18] PHILIPPE Courtier, J-N Thépaut, and Anthony Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

[19] Lars Nerger and Watson W Gregg. Improving assimilation of seawifs data by the application of bias correction with a local seik filter. *Journal of marine systems*, 73(1-2):87–102, 2008.

[20] Pierre P Brasseur and Jacques A Haus. Application of a 3-d variational inverse model to the analysis of ecohydrodynamic data in the northern bering and southern chukchi seas. *Journal of Marine Systems*, 1(4):383–401, 1991.

[21] Stephen M Griffies et al. Elements of the modular ocean model (mom). *GFDL Ocean Group Tech. Rep*, 7(620):47, 2012.

[22] Dorothee CE Bakker, Benjamin Pfeil, Camilla S Landa, Nicolas Metzl, Kevin M O'brien, Are Olsen, Karl Smith, Cathy Cosca, Sumiko Harasawa, Stephen D Jones, et al. A multi-decade record of high-quality f co 2 data in version 3 of the surface ocean co 2 atlas (socat). *Earth System Science Data*, 8(2):383–413, 2016.

[23] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[24] Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. *arXiv preprint arXiv:2404.10024*, 2024.

[25] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[26] Alistair Adcroft, Whit Anderson, V Balaji, Chris Blanton, Mitchell Bushuk, Carolina O Dufour, John P Dunne, Stephen M Griffies, Robert Hallberg, Matthew J Harrison, et al. The gfdl global ocean and sea ice model om4. 0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, 11(10):3167–3211, 2019.

[27] Charles A Stock, John P Dunne, Songmiao Fan, Paul Ginoux, Jasmin John, John P Krasting, Charlotte Laufkötter, Fabien Paulot, and Niki Zadeh. Ocean biogeochemistry in gfdl's earth system model 4.1 and its response to increasing atmospheric co2. *Journal of Advances in Modeling Earth Systems*, 12(10):e2019MS002043, 2020.

[28] Olivier Aumont, Christian Éthé, Alessandro Tagliabue, Laurent Bopp, and Marion Gehlen. Pisces-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development Discussions*, 8(2):1375–1509, 2015.

[29] John P Dunne, LW Horowitz, AJ Adcroft, P Ginoux, IM Held, JG John, John P Krasting, Sergey Malyshev, V Naik, F Paulot, et al. The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11):e2019MS002015, 2020.

[30] Ariane Verdy and Matthew R Mazloff. A data assimilating model for estimating s outhern o cean biogeochemistry. *Journal of Geophysical Research: Oceans*, 122(9):6968–6988, 2017.

[31] Alexander F Shchepetkin and James C McWilliams. The regional oceanic modeling system (roms): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean modelling*, 9(4):347–404, 2005.

[32] Cecile S Rousseaux and Watson W Gregg. Climate variability and phytoplankton composition in the pacific ocean. *Journal of Geophysical Research: Oceans*, 117(C10), 2012.

[33] Malek Belgacem, Katrin Schroeder, Alexander Barth, Charles Troupin, Bruno Pavoni, Patrick Raimbault, Nicole Garcia, Mireno Borghini, and Jacopo Chiggiato. Climatological distribution of dissolved inorganic nutrients in the western mediterranean sea (1981–2017). *Earth System Science Data*, 13(12):5915–5949, 2021.

[34] Wanqin Zhong, Xin Ma, Tianqi Shi, Ge Han, Haowei Zhang, and Wei Gong. Reconstruction of global ocean surface pco2 and air-sea co2 flux: Based on multigrained cascade forest model. *Journal of Geophysical Research: Oceans*, 130(2):e2024JC021483, 2025.

[35] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv:1901.00212*, 2019.

[36] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[37] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International conference on learning representations*, 2020.

[38] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv:1711.10561*, 2017.

[39] Eugenio Cutolo, Ananda Pascual, Simon Ruiz, Nikolaos D Zarokanellos, and Ronan Fablet. Cloinet: ocean state reconstructions through remote-sensing, in-situ sparse observations and deep learning. *Frontiers in Marine Science*, 11:1151868, 2024.

[40] Bin Lu, Ze Zhao, Luyu Han, Xiaoying Gan, Yuntao Zhou, Lei Zhou, Luoyi Fu, Xinbing Wang, Chenghu Zhou, and Jing Zhang. Oxygenerator: Reconstructing global ocean deoxygenation over a century with deep learning. *arXiv:2405.07233*, 2024.

[41] Jonathan D Sharp, Denis Pierrot, Matthew P Humphreys, Jean-Marie Epitalon, James C Orr, Ernie R Lewis, and Douglas WR Wallace. Co2sysv3 for matlab. *Zenodo*, 2020.

[42] I Loshchilov. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

[43] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[44] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[45] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

A Related Work

A.1 Oceanographic Data

The two main sources of oceanographic data are real-world observational data and model-based simulation data.

Observation data. Observation data are derived from direct in situ measurements of the ocean and therefore provide the most physically grounded source of supervision for ocean biogeochemical variables. Representative observational data sources include the WOD [1] and the SOCAT [22]. We also use GLODAP observations for additional pH reconstruction experiments. WOD is a long-term global archive of high fidelity oceanographic profile observations from diverse observing systems and historical cruises. It contains a wide range of essential ocean variables, including temperature, salinity, and carbonate-system-related measurements such as sea surface pH, making it a valuable resource for characterizing ocean variability over several decades. SOCAT, provides high-quality observations of $p\mathrm{CO}_2^s$ from ships, drifters, and other surface observing platforms. These datasets offer complementary observational constraints on ocean biogeochemistry and air–sea carbon exchange. However, despite their high measurement fidelity, such observations remain highly sparse and irregular in both space and time because their sampling distributions are constrained by ship tracks, regional observing priorities, platform availability, and temporal measurement frequency.

Simulation data. Simulation data are generated by numerical ocean and Earth system models that explicitly couple ocean circulation with marine biogeochemical and carbon-cycle processes. Representative model systems include MOM6-COBALTv2 [26, 27], NEMO-PISCES [28], and GFDL-ESM4 [29, 27]. MOM6 forms the ocean component of GFDL OM4 model, which is used in GFDL climate and Earth system simulations, while COBALTv2 represents marine biogeochemical processes and ecosystem dynamics within the GFDL Earth system modeling framework. NEMO-PISCES couples the NEMO ocean circulation model with PISCES-v2, a biogeochemical model designed for carbon-cycle and ecosystem studies representing lower trophic levels and the cycles of carbon and major nutrients. GFDL-ESM4 further integrates ocean physics, biogeochemistry, ecosystems, atmosphere, sea ice, and land processes within a coupled carbon and chemistry modeling system. These simulation products provide continuous, gridded, and spatially complete fields of pH, $p\mathrm{CO}_2^s$, DIC, and associated physical and biogeochemical variables, making them suitable as dense supervision signals and controlled benchmarks for learning-based models. Nevertheless, simulation outputs are not direct observations and may be affected by structural model assumptions, parameterization choices, numerical resolution, and errors in external forcing.

A.2 Assimilation and Inversion Methods

Since observed sea surface pH records are sparse, irregular, and unevenly distributed, ocean pH reconstruction has traditionally relied on assimilation and inversion methods to infer unobserved states. These methods estimate continuous ocean fields either by incorporating available measurements into numerical ocean models or by solving constrained optimization problems with prescribed physical, statistical, or smoothness assumptions.

Assimilation Methods. Data assimilation methods combine observations with numerical model states to obtain estimates that are more consistent with ocean dynamics and biogeochemical processes. Representative approaches include Optimal Interpolation (OI) [17], which infers missing values through covariance-based statistical weighting of nearby observations; variational methods such as 3D-Var and 4D-Var [18], which optimize model states or controls by reducing the mismatch between observations and simulations; and ensemble-based methods such as the Ensemble Kalman Filter (EnKF) [19], which sequentially updates model states while representing uncertainty through an ensemble. In large-scale ocean studies, these ideas are reflected in systems such as the Biogeochemical Southern Ocean State Estimate (B-SOSE) [30], the Regional Ocean Modeling System (ROMS) [31], MOM6 [21], and the NASA Ocean Biogeochemical Model (NOBM) [32]. By exploiting model dynamics and multivariate coupling, these systems can improve the reconstruction of sparsely observed ocean variables. However, they are still constrained by observation scarcity, model structural errors, uncertain error statistics, and the high computational cost of repeated model integration.

Inversion Methods. Inversion-based methods provide another important route for ocean field reconstruction. Instead of updating a forward model sequentially, they usually cast reconstruction as an optimization problem, seeking ocean states that fit the available observations while satisfying predefined physical, statistical, or regularization constraints. Classical examples include Variational Inverse Models (VIM) [20, 33], which reconstruct oceanographic fields by jointly minimizing data mismatch and regularization terms, such as spatial smoothness or dynamical consistency penalties. Compared with full assimilation systems, inversion methods provide a more direct and flexible formulation for reconstructing sparsely observed ocean biogeochemical variables. Nevertheless, their effectiveness depends heavily on the design of the objective function, the choice of regularization, and the prior assumptions imposed on the solution. In regions with extremely sparse observations, they may still have difficulty recovering fine-scale variability or correcting biased spatial structures.

A.3 Data-Driven Reconstruction Methods

Due to the heavy computational burden and modeling assumptions of assimilation and inversion methods, data-driven reconstruction approaches have received increasing attention for estimating sparse ocean fields.

Supervised Learning from Observable Proxies. A common class of data-driven methods learns a mapping from more accessible proxy variables to target ocean variables. These proxies may include sea surface temperature, SSH, Chl., wind fields, or reanalysis-derived physical fields. Such models are usually trained in a supervised manner using observational products, numerical simulations, or remote sensing datasets. Representative examples include Earthformer [7], a transformer architecture designed for spatiotemporal Earth system data. Tree-based ensemble models, such as gcForest [34], have also been applied to reconstruct ocean fields from satellite and reanalysis inputs. These methods can capture nonlinear relationships between multivariate predictors and target fields, making them useful for large-scale reconstruction. However, because their predictions are mainly driven by learned statistical correlations, they do not explicitly enforce physical or chemical consistency. As a result, their performance may degrade in sparsely sampled regions, unseen dynamical regimes, or distribution shifts beyond the training data.

Foundation Models for Ocean Fields. Recent progress in large-scale pretraining has further introduced foundation models as a general framework for Earth and ocean field reconstruction. Models such as Prithvi [10] and AtmoRep [11] are trained on broad collections of geobiochemical variables across different spatial resolutions and temporal scales, aiming to learn transferable representations for multiple downstream tasks. For ocean reconstruction, such models have the potential to support interpolation, gap filling, forecasting, and cross-variable inference from sparse or heterogeneous inputs. Their scalability and pretraining paradigm make them suitable for general-purpose reconstruction across different ocean variables and regions. However, this line of research remain underexplored because of their heavy demand for large and diverse training datasets, lack of design level guarantees for physical consistency, and poor generalization to sparse, or dynamical complex ocean regimes.

General AI Methods for Reconstruction. Beyond ocean-specific reconstruction models, general-purpose AI methods have also been explored for recovering sparse geobiochemical fields. Image inpainting approaches, including GAN-based models such as CoModGAN [8] and EdgeConnect [35], transformer-based methods such as the Mask-Aware Transformer (MAT) [36], and diffusion-based methods such as RePaint [9], treat gridded ocean variables as images and infer missing regions from learned spatial structures. These methods can handle large or irregular gaps, but they usually lack explicit constraints required by scientific variables. Neural operators learn mappings between function spaces and can approximate solution operators for dynamical systems and PDEs. Typical examples include Fourier Neural Operator (FNO) [12], which models fields in spectral space, and MeshGraphNets [37], which can operate on irregular spatial domains. Although such methods support flexible prediction across resolutions, they often depend on high-quality simulation data and careful training. Implicit Neural Representations (INRs) model fields as continuous coordinate-based functions, enabling arbitrary-resolution reconstruction from sparse observations [14]. Overall, these approaches provide flexible tools for ocean field reconstruction, but their application in scientific settings is still limited by data scarcity, weak physical consistency, and the multi-scale nature of geobiochemical processes.

A.4 Physics-Informed Models

To address the physical inconsistency of purely data-driven methods, recent studies have developed physics-constrained and hybrid learning frameworks for ocean field reconstruction.

Physics-constrained and Hybrid Models. Physics-constrained models aim to improve the reliability and plausibility of learning-based reconstruction by incorporating physical equations, numerical schemes, or domain knowledge into the model design or training objective. Physics-informed neural networks (PINNs) [38] introduce governing equations or partial differential equation constraints directly into the loss function. Hybrid methods such as CLOINet [39] combine OI with deep learning, allowing the model to exploit both statistical interpolation and regime-dependent neural reconstruction. Similarly, 4DVarNet [15] reformulates variational data assimilation as an end-to-end trainable neural framework, preserving an assimilation-inspired structure while learning reconstruction operators from data. Other methods, such as OxyGenerator [40], incorporate graph-based representations and oceanographic domain knowledge for long-term ocean field reconstruction. NUTS [16] further reconstructs ocean nutrients by decoupling ocean flows based on homogenization theory and applying mass conservation regularization to improve physical consistency. These hybrid approaches can enhance generalization and produce more plausible reconstructions than unconstrained black-box models. However, they are often sensitive to the specific design of physical constraints, network architectures, and training strategies. Moreover, they require substantial computational resources when applied to high-resolution or long-term reconstruction. More importantly, many existing physics-constrained reconstruction methods mainly focus on transport-dominated variables. However, for passive tracers, physical transport constraints alone are insufficient. There is still a critical lack of comprehensive physical and biochemical considerations to ensure fully consistent reconstructions in complex marine environments.

B Theory for Carbon-State Reconstruction

This appendix formalizes the claim in Section 4.3. The result has two parts. The positive part shows that REACT provides a certificate for carbonate closure and source-free DIC conservation when the chemical and physical losses vanish. The negative part shows that a passive pH transport model cannot provide the same certificate in general: conserving the mass of pH does not imply conservation of the DIC mass implied by carbonate chemistry.

B.1 Notation and setup

Let Ω be a compact smooth two-dimensional Riemannian domain representing the ocean surface, equipped with surface measure dx. The domain may be closed, as on the sphere, or may have a smooth boundary $\partial\Omega$. When $\partial\Omega \neq \varnothing$, n denotes the outward unit normal. Let

$$\Omega_T = \Omega \times [t_0, T].$$

For any scalar field $r : \Omega_T \to \mathbb{R}$, define its spatial inventory by

$$M[r](t) = \int_\Omega r(x,t)\, dx.$$

We use the convention that all equalities between functions hold almost everywhere unless otherwise stated.

Let $h(x,t)$ denote pH, $d(x,t)$ denote DIC, and $c(x,t)$ denote surface $p\mathrm{CO}_2$. Let Γ_D and Γ_C be the carbonate-equilibrium maps

$$\Gamma_D(h,c) = \text{DIC implied by pH } h \text{ and } p\mathrm{CO}_2\ c,$$

and

$$\Gamma_C(h,d) = p\mathrm{CO}_2 \text{ implied by pH } h \text{ and DIC } d.$$

The chemical residual of a reconstructed pH field $\hat{h}$ is

$$\mathcal{L}_{\mathrm{chem}}(\hat{h}) = \|\Gamma_D(\hat{h},c) - d\|^2_{L^2(\mathcal{I})} + \beta\|\Gamma_C(\hat{h},d) - c\|^2_{L^2(\mathcal{I})}, \qquad \beta > 0,$$

where $\mathcal{I} \subseteq \Omega_T$ is the observed space–time set.

REACT evolves a latent carbonate state z by a source-free conservative transport equation

$$\partial_t z + \nabla \cdot (uz) - \nabla \cdot (K_\theta \nabla z) = 0, \tag{2}$$

with no-flux boundary condition

$$(uz - K_\theta \nabla z) \cdot n = 0 \qquad \text{on } \partial\Omega.$$

If $\partial\Omega = \varnothing$, this boundary condition is omitted. The transported state is then corrected by an active source module,

$$\tilde{z}_t = z_t + s_t, \qquad s_t = S_\theta(z_t, u, c, \xi),$$

and decoded into pH by

$$\hat{h}_t = G_\theta(\tilde{z}_t, c_t, \xi_t).$$

For compactness we suppress ξ when it is not needed. The represented DIC of a latent state w is

$$\mathcal{D}_\theta(w, c) = \Gamma_D(G_\theta(w, c), c). \tag{3}$$

The source-free physical loss is

$$\mathcal{L}_{\text{phys}} = \int_{t_0}^{T} |M[\mathcal{D}_\theta(z_t, c_t)] - M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})]|^2 \, dt. \tag{4}$$

B.2 Assumptions

We use the following assumptions only to make the theorem precise.

A1. Carbonate consistency and invertibility. There exists an admissible interval H of pH values, an admissible interval C of $p\mathrm{CO}_2$ values, and an admissible interval D of DIC values such that

$$\Gamma_D : \mathsf{H} \times \mathsf{C} \to \mathsf{D}, \qquad \Gamma_C : \mathsf{H} \times \mathsf{D} \to \mathsf{C}$$

are well-defined. For each fixed $c \in \mathsf{C}$, the map

$$h \mapsto \Gamma_D(h, c)$$

is one-to-one on H. Its inverse in the first argument is denoted by

$$H(D, c),$$

so that

$$\Gamma_D(H(D, c), c) = D.$$

We also assume carbonate consistency:

$$\Gamma_C(H(D, c), D) = c.$$

A2. Observation identifiability. The observed set $\mathcal{I}$ is sufficient to identify the chemical residuals in the admissible reconstruction class. Concretely, if an admissible $\hat{h}$ satisfies

$$\Gamma_D(\hat{h}, c) = d \quad \text{and} \quad \Gamma_C(\hat{h}, d) = c \qquad \text{on } \mathcal{I},$$

then the same identities hold on the target domain under consideration. In the simplest case, $\mathcal{I} = \Omega_T$, and this assumption is automatic. Without this assumption, the conclusion below holds on $\mathcal{I}$.

A3. Realizability. For the realizability part of the theorem, assume the true DIC field decomposes as

$$d(x, t) = d^0(x, t) + r(x, t),$$

where d^0 is the source-free transported DIC branch and r is the active carbon residual. The source-free branch satisfies

$$\partial_t d^0 + \nabla \cdot (ud^0) - \nabla \cdot (K^\star \nabla d^0) = 0, \tag{5}$$

with no-flux boundary condition

$$(ud^0 - K^\star \nabla d^0) \cdot n = 0.$$

Assume the model class contains the DIC-coordinate realization

$$z_t = d_t^0, \qquad K_\theta = K^\star, \qquad S_\theta(z_t, u, c, \xi) = r_t, \qquad G_\theta(w, c) = H(w, c).$$

This assumption does not require every implementation to use raw DIC as z; it only says that the model class is expressive enough to represent such a carbon-coordinate solution.

B.3 Positive result: carbon-state certificate

Theorem B.1 (Carbon-state certificate and realizability)**.** *Suppose A1–A2 hold and $\beta > 0$. Let z_t, $\tilde{z}_t$, and $\hat{h}_t$ be a REACT trajectory, with represented DIC $\mathcal{D}_\theta$ defined by* (3)*.*

- *If $\mathcal{L}_{\mathrm{chem}} = 0$, then the reconstructed pH closes carbonate chemistry:*

$$\Gamma_D(\hat{h}, c) = d \qquad \textit{and} \qquad \Gamma_C(\hat{h}, d) = c$$

  on the target domain identified by $\mathcal{I}$. In particular, if $\mathcal{I} = \Omega_T$, these equalities hold almost everywhere on Ω_T.

- *If $\mathcal{L}_{\mathrm{phys}} = 0$, then the source-free branch conserves represented DIC mass:*

$$M[\mathcal{D}_\theta(z_t, c_t)] = M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})]$$

  for almost every $t \in [t_0, T]$. If the map

$$t \mapsto M[\mathcal{D}_\theta(z_t, c_t)]$$

  is continuous, then the equality holds for every $t \in [t_0, T]$.

- *If, in addition, A3 holds and the true fields are chemically consistent,*

$$d = \Gamma_D(h, c), \qquad c = \Gamma_C(h, d),$$

  then there exists a REACT realization with

$$\hat{h} = h, \qquad \mathcal{L}_{\mathrm{chem}} = 0, \qquad \mathcal{L}_{\mathrm{phys}} = 0.$$

  If the pH observations are noiseless, this same realization also gives

$$\mathcal{L}_{\mathrm{pH}} = 0.$$

Proof. We prove the three claims separately.

1. Chemical closure. By definition,

$$\mathcal{L}_{\mathrm{chem}} = \|\Gamma_D(\hat{h}, c) - d\|^2_{L^2(\mathcal{I})} + \beta\|\Gamma_C(\hat{h}, d) - c\|^2_{L^2(\mathcal{I})}.$$

Both terms are nonnegative, and $\beta > 0$. Therefore $\mathcal{L}_{\mathrm{chem}} = 0$ implies

$$\|\Gamma_D(\hat{h}, c) - d\|^2_{L^2(\mathcal{I})} = 0$$

and

$$\|\Gamma_C(\hat{h}, d) - c\|^2_{L^2(\mathcal{I})} = 0.$$

Hence

$$\Gamma_D(\hat{h}, c) = d \qquad \text{and} \qquad \Gamma_C(\hat{h}, d) = c$$

almost everywhere on $\mathcal{I}$. By the observation identifiability assumption A2, these identities hold on the target domain. If $\mathcal{I} = \Omega_T$, this conclusion is immediate without any additional identifiability argument.

2. Source-free represented-DIC conservation. Define

$$m(t) = M[\mathcal{D}_\theta(z_t, c_t)] - M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})].$$

Then

$$\mathcal{L}_{\mathrm{phys}} = \int_{t_0}^{T} |m(t)|^2 dt.$$

Since $|m(t)|^2 \geq 0$, the equality $\mathcal{L}_{\mathrm{phys}} = 0$ implies $m(t) = 0$ for almost every t. Thus

$$M[\mathcal{D}_\theta(z_t, c_t)] = M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})]$$

for almost every t. If m is continuous in time, then an almost-everywhere zero continuous function is identically zero, so the equality holds for every $t \in [t_0, T]$.

3. Realizability. Assume A3. Choose the DIC-coordinate realization

$$z_t = d_t^0, \qquad K_\theta = K^\star, \qquad S_\theta(z_t, u, c, \xi) = r_t, \qquad G_\theta(w, c) = H(w, c).$$

Then the corrected latent state is

$$\tilde{z}_t = z_t + s_t = d_t^0 + r_t = d_t.$$

The decoded pH is therefore

$$\hat{h}_t = G_\theta(\tilde{z}_t, c_t) = H(d_t, c_t).$$

Because the true fields are chemically consistent and H is the inverse of $h \mapsto \Gamma_D(h, c)$, we have

$$H(d_t, c_t) = h_t.$$

Hence $\hat{h} = h$. It follows immediately that $\mathcal{L}_{\mathrm{pH}} = 0$ for noiseless pH observations.

Next, since $\hat{h} = h$ and the true fields are chemically consistent,

$$\Gamma_D(\hat{h}, c) = \Gamma_D(h, c) = d$$

and

$$\Gamma_C(\hat{h}, d) = \Gamma_C(h, d) = c.$$

Therefore $\mathcal{L}_{\mathrm{chem}} = 0$.

Finally, for the source-free branch,

$$\mathcal{D}_\theta(z_t, c_t) = \Gamma_D(G_\theta(z_t, c_t), c_t) = \Gamma_D(H(z_t, c_t), c_t) = z_t = d_t^0.$$

Thus the represented DIC of the source-free branch is exactly d_t^0. Integrating the source-free DIC equation (5) over Ω gives

$$\frac{d}{dt} M[d^0](t) = -\int_{\partial\Omega} (u d^0 - K^\star \nabla d^0) \cdot n \, dS.$$

The no-flux boundary condition makes the boundary integral vanish. Therefore

$$M[d^0](t) = M[d^0](t_0) \qquad \text{for all } t \in [t_0, T].$$

Because $\mathcal{D}_\theta(z_t, c_t) = d_t^0$, we obtain

$$M[\mathcal{D}_\theta(z_t, c_t)] = M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})],$$

and hence $\mathcal{L}_{\mathrm{phys}} = 0$. This proves the theorem. □

B.4 Negative result: passive pH transport is not a DIC-conservation certificate

We now show that the passive pH baseline lacks the certificate above. The issue is not that passive pH transport cannot conserve pH mass. It can. The issue is that pH mass is not the conserved carbonate inventory. The conserved source-free inventory is DIC, and the map from pH to DIC is nonlinear.

First, we record a purely algebraic fact.

Lemma B.1 (pH mass does not determine DIC mass)**.** *Let $\gamma : \mathsf{H} \to \mathbb{R}$ be a non-affine continuous function, where $\mathsf{H} \subset \mathbb{R}$ is an interval. Then there exist two bounded pH fields $h_1, h_2 : \Omega \to \mathsf{H}$ such that*

$$M[h_1] = M[h_2],$$

but

$$M[\gamma(h_1)] \neq M[\gamma(h_2)].$$

In particular, when $\gamma(h) = \Gamma_D(h, c_0)$ for fixed c_0, conservation of pH mass alone does not imply conservation of implied DIC mass.

Proof. Because γ is non-affine, there exist $a, b \in \mathsf{H}$ and $\lambda \in (0, 1)$ such that, with

$$m = \lambda a + (1 - \lambda) b,$$

we have

$$\gamma(m) \neq \lambda\gamma(a) + (1 - \lambda)\gamma(b).$$

Choose a measurable subset $A \subset \Omega$ with $|A| = \lambda|\Omega|$. Define

$$h_1(x) = m$$

and

$$h_2(x) = \begin{cases} a, & x \in A, \\ b, & x \in \Omega \setminus A. \end{cases}$$

Then

$$M[h_1] = |\Omega|m = |\Omega|\{\lambda a + (1-\lambda)b\} = M[h_2].$$

However,

$$M[\gamma(h_1)] = |\Omega|\gamma(m),$$

whereas

$$M[\gamma(h_2)] = |\Omega|\{\lambda\gamma(a) + (1-\lambda)\gamma(b)\}.$$

These two quantities are different by construction. Therefore equal pH mass does not determine equal implied DIC mass. □

The next theorem gives a dynamical counterexample for the passive advection–diffusion baseline.

Theorem B.2 (Passive pH transport can conserve pH mass while changing implied DIC mass)**.** *Fix a constant $p\mathrm{CO}_2$ value c_0, and define*

$$\gamma(h) = \Gamma_D(h, c_0).$$

Assume $\gamma \in C^2(\mathsf{H})$ and γ is nonlinear in the sense that there exists $a \in \mathsf{H}$ with

$$\gamma''(a) \neq 0.$$

Let h solve the passive pH transport equation

$$\partial_t h + \nabla \cdot (uh) - \nabla \cdot (K\nabla h) = 0 \tag{6}$$

on Ω, with no-flux boundary condition

$$u \cdot n = 0, \qquad K\nabla h \cdot n = 0 \qquad \text{on } \partial\Omega.$$

Assume u is divergence-free,

$$\nabla \cdot u = 0,$$

and $K(x)$ is symmetric positive definite on some open set $U \subseteq \Omega$.

Then there exists a smooth nonconstant initial pH field $h_0 : \Omega \to \mathsf{H}$ such that the corresponding solution satisfies

$$M[h](t) = M[h](t_0) \qquad \text{for all } t,$$

but

$$\frac{d}{dt} M[\Gamma_D(h, c_0)](t)\Big|_{t=t_0} \neq 0.$$

Consequently,

$$M[\Gamma_D(h, c_0)](t)$$

is not conserved, even though passive pH mass is conserved.

Proof. First, integrate (6) over Ω:

$$\frac{d}{dt} M[h](t) = \int_\Omega \partial_t h \, dx = -\int_\Omega \nabla \cdot (uh) \, dx + \int_\Omega \nabla \cdot (K\nabla h) \, dx.$$

By the divergence theorem,

$$\frac{d}{dt} M[h](t) = -\int_{\partial\Omega} uh \cdot n \, dS + \int_{\partial\Omega} K\nabla h \cdot n \, dS.$$

The no-flux boundary conditions imply both boundary terms vanish. Hence

$$\frac{d}{dt} M[h](t) = 0,$$

so

$$M[h](t) = M[h](t_0)$$

for all t.

Now consider the DIC implied by pH and fixed $p\mathrm{CO}_2$:

$$D_{\mathrm{pass}}(x,t) = \Gamma_D(h(x,t), c_0) = \gamma(h(x,t)).$$

Its mass derivative is

$$\frac{d}{dt} M[D_{\mathrm{pass}}](t) = \frac{d}{dt} \int_\Omega \gamma(h(x,t))\, dx = \int_\Omega \gamma'(h) \partial_t h \, dx.$$

Using (6),

$$\frac{d}{dt} M[D_{\mathrm{pass}}](t) = -\int_\Omega \gamma'(h) \nabla \cdot (uh)\, dx + \int_\Omega \gamma'(h) \nabla \cdot (K \nabla h)\, dx.$$

Because $\nabla \cdot u = 0$, we have $\nabla \cdot (uh) = u \cdot \nabla h$. Therefore

$$-\int_\Omega \gamma'(h) \nabla \cdot (uh)\, dx = -\int_\Omega \gamma'(h) u \cdot \nabla h \, dx = -\int_\Omega u \cdot \nabla \gamma(h)\, dx.$$

By the divergence theorem,

$$-\int_\Omega u \cdot \nabla \gamma(h)\, dx = -\int_{\partial\Omega} \gamma(h) u \cdot n \, dS + \int_\Omega \gamma(h) \nabla \cdot u \, dx = 0.$$

For the diffusion term, integration by parts gives

$$\int_\Omega \gamma'(h) \nabla \cdot (K \nabla h)\, dx = \int_{\partial\Omega} \gamma'(h) K \nabla h \cdot n \, dS - \int_\Omega \gamma''(h) \nabla h^\top K \nabla h \, dx.$$

The boundary term vanishes by $K \nabla h \cdot n = 0$. Hence

$$\frac{d}{dt} M[D_{\mathrm{pass}}](t) = -\int_\Omega \gamma''(h) \nabla h^\top K \nabla h \, dx. \tag{7}$$

It remains to choose an initial field for which the right-hand side is nonzero at $t = t_0$. Since $\gamma''(a) \neq 0$, by continuity there exists a small interval $J \subset \mathsf{H}$ containing a such that γ'' has a fixed nonzero sign on J. Choose a smooth nonconstant function ϕ supported inside U, and choose $\varepsilon \neq 0$ sufficiently small that

$$h_0(x) = a + \varepsilon \phi(x)$$

takes values in J. Then

$$\nabla h_0 = \varepsilon \nabla \phi.$$

Evaluating (7) at $t = t_0$ yields

$$\left.\frac{d}{dt} M[D_{\mathrm{pass}}](t)\right|_{t=t_0} = -\varepsilon^2 \int_\Omega \gamma''(a + \varepsilon\phi) \nabla \phi^\top K \nabla \phi \, dx.$$

On the support of ϕ, K is positive definite, and $\nabla\phi$ is nonzero on a set of positive measure because ϕ is nonconstant. Also, $\gamma''(a + \varepsilon\phi)$ has a fixed nonzero sign. Therefore the integral is nonzero, and so

$$\left.\frac{d}{dt} M[D_{\mathrm{pass}}](t)\right|_{t=t_0} \neq 0.$$

Thus the implied DIC mass changes at t_0, even though $M[h]$ is conserved. This proves the claim. □

Remark. The diffusion term is the essential mixing mechanism in Theorem B.2. In the degenerate case of purely incompressible advection with no diffusion, the full distribution of h is transported, so $\int_\Omega \gamma(h)\, dx$ is also conserved for any fixed nonlinear γ. That special case does not provide a certificate for passive pH reconstruction in the advection–diffusion setting used by the baseline. Once mixing or coarse-grained diffusion is present, nonlinear carbonate chemistry makes conservation of $M[h]$ insufficient for conservation of $M[\Gamma_D(h, c)]$.

B.5 Consequence for REACT versus passive pH transport

Theorem B.1 and Theorem B.2 formalize the modeling distinction. REACT transports a carbon-related state and evaluates the source-free branch through represented DIC,

$$\mathcal{D}_\theta(z_t, c_t) = \Gamma_D(G_\theta(z_t, c_t), c_t).$$

Thus zero physical loss directly certifies conservation of the represented DIC inventory. In contrast, the passive baseline transports pH itself. Even when that baseline exactly conserves

$$M[h](t),$$

Theorem B.2 shows that it can change

$$M[\Gamma_D(h, c)](t),$$

which is the chemically implied DIC inventory. Therefore passive pH transport has the right conservative PDE form but applies it to the wrong variable. REACT avoids this mismatch by placing the source-free transport constraint on a carbon coordinate and decoding pH only after conservative carbon transport and active carbon-cycle correction.

B.6 Exact and finite-loss consistency statements

The exact zero-loss certificate is the limiting case of a finite-loss statement. Define

$$r_D = \Gamma_D(\hat{h}, c) - d, \qquad r_C = \Gamma_C(\hat{h}, d) - c.$$

If $\mathcal{L}_{\text{chem}} \le \varepsilon_{\text{chem}}$, then

$$\|r_D\|_{L^2(\mathcal{I})} \le \sqrt{\varepsilon_{\text{chem}}}, \qquad \|r_C\|_{L^2(\mathcal{I})} \le \sqrt{\varepsilon_{\text{chem}}/\beta}.$$

Similarly, $\mathcal{L}_{\text{phys}} \le \varepsilon_{\text{phys}}$ implies

$$\left[\frac{1}{T - t_0}\int_{t_0}^{T} |M[\mathcal{D}_\theta(z_t, c_t)] - M[\mathcal{D}_\theta(z_{t_0}, c_{t_0})]|^2 \, dt\right]^{1/2} \le \sqrt{\frac{\varepsilon_{\text{phys}}}{T - t_0}}.$$

These bounds follow directly from the nonnegative terms in the loss definitions and do not require zero loss to be attainable. For finitely trained models, the achieved losses therefore quantify the remaining carbonate-closure residuals on $\mathcal{I}$ and source-free represented-DIC inventory drift. The zero-loss theorem recovers exact consistency under its stated assumptions; nonzero losses on real observations give approximate consistency relative to the supplied observations and reference fields.

C Implementation Details of REACT

C.1 Architectural Details

REACT consists of an initial state reconstruction module, a conservative transport module, a source module, a pH decoder and a chemical module.

Initial State Reconstruction Module. The initial state reconstruction module estimates the initial carbon-related state from sparse observations and auxiliary fields. The input contains pH observations, $p\text{CO}_2^s$ observations, and nine MOM6-derived variables. We use a ViViT-style [25] spatiotemporal transformer as the backbone of this module. The input sequence is divided into non-overlapping spatiotemporal tubelets, and each tubelet is projected into a latent token. The tokens are then processed by stacked transformer blocks to capture long-range spatial dependencies and temporal correlations. The output tokens are reshaped back to the global latitude–longitude grid to produce the initial carbon-related transport state:

$$z_0 = F_\theta(\mathcal{O}_h, \mathcal{O}_c, \mathbf{X}_{\text{MOM6}}),$$

where $\mathcal{O}_h$ denotes sparse pH observations, $\mathcal{O}_c$ denotes sparse $p\text{CO}_2^s$ observations, and $\mathbf{X}_{\text{MOM6}}$ denotes the nine MOM6-derived variables. The reconstructed state z_0 serves as the initial carbon-related transport state for subsequent physical propagation.

Conservative Transport Module. The evolution of the reconstructed carbon-related state is governed by the standard advection-diffusion equation:

$$\frac{\partial z}{\partial t} = -\nabla \cdot (z\mathbf{w}) + \nabla \cdot (K \nabla z),$$

where z is the carbon-related transport state, $\mathbf{w} = (u, v)$ denotes the horizontal surface velocity field, and K is the diffusion coefficient. The advection term describes flow-driven transport, while the diffusion term describes unresolved spreading. In our implementation, the diffusion term is removed and only the advective transport term is retained. Given the initial state z_0, the module produces the propagated states

$$z_t = \mathcal{P}(z_0, \mathbf{w}_{1:t}), \quad t = 1, \ldots, T,$$

where $\mathcal{P}$ denotes the physical transport operator.

Source Module. The source module parameterizes the non-conservative residual in the carbon-state evolution. After the conservative transport module produces z_t, the source module predicts a source correction

$$s_t = S_\theta(z_t),$$

where S_θ is implemented as a U-Net [5]. The corrected carbon-related state is then obtained by

$$\tilde{z}_t = z_t + s_t.$$

Here s_t represents unresolved source and sink effects associated with air–sea CO_2 exchange, biological activity, and other biogeochemical processes that cannot be captured by source-free transport alone.

pH Decoder. The pH decoder maps the corrected carbon-related state to the reconstructed sea surface pH field. Given the corrected state $\tilde{z}_t$ and the carbonate equilibrium constants K_1 and K_2, the decoder predicts

$$\hat{h}_t = G_\theta(\tilde{z}_t, K_{1,t}, K_{2,t}),$$

where G_θ is implemented as a U-Net. The constants K_1 and K_2 denote the first and second dissociation constants of the carbonate system, which characterize the nonlinear equilibrium among carbonate species and hydrogen ions. By conditioning on K_1 and K_2, the decoder maps the corrected carbon-related state to pH under carbonate-equilibrium conditions.

Chemical Module. The chemical module implements carbonate-equilibrium formulas that couple reconstructed pH, $p\mathrm{CO}_2^s$, and DIC. Given reconstructed pH, it computes the DIC implied by observed $p\mathrm{CO}_2^s$ and the $p\mathrm{CO}_2^s$ implied by observed DIC. These calculations are used to evaluate and enforce carbonate-system consistency. Standard carbonate-system calculations can be performed by CO2SYS [41] (under MIT license)[3] or PyCO2SYS [4] (under GPL-3.0 license)[4]. However, these solvers are not directly used as differentiable PyTorch layers in our end-to-end training pipeline. We therefore implement the carbonate-equilibrium calculation with explicit analytical formulas, following the steps below:

Given the reconstructed pH $\hat{h}_t$, the hydrogen ion concentration is

$$[\mathrm{H}^+]_t = 10^{-\hat{h}_t}.$$

The dissolved CO_2 concentration is computed from $p\mathrm{CO}_2^s$ as

$$[\mathrm{CO}_2^*]_t = K_{0,t}\mathcal{O}_{c,t},$$

where $K_{0,t}$ is the CO_2 solubility coefficient and $\mathcal{O}_{c,t}$ denotes $p\mathrm{CO}_2^s$ observations. The bicarbonate and carbonate ion concentrations are then given by

$$[\mathrm{HCO}_3^-]_t = \frac{K_{1,t}[\mathrm{CO}_2^*]_t}{[\mathrm{H}^+]_t}, \qquad [\mathrm{CO}_3^{2-}]_t = \frac{K_{1,t}K_{2,t}[\mathrm{CO}_2^*]_t}{[\mathrm{H}^+]_t^2},$$

where $K_{1,t}$ and $K_{2,t}$ are the first and second dissociation constants of the carbonate system. The DIC implied by the reconstructed pH is

$$\hat{d}_t = [\mathrm{CO}_2^*]_t + [\mathrm{HCO}_3^-]_t + [\mathrm{CO}_3^{2-}]_t,$$

[3] https://github.com/jamesorr/CO2SYS-MATLAB

[4] https://github.com/mvdh7/PyCO2SYS

or equivalently,

$$\hat{d}_t = K_{0,t}\mathcal{O}_{c,t}\left(1 + \frac{K_{1,t}}{[\mathrm{H}^+]_t} + \frac{K_{1,t}K_{2,t}}{[\mathrm{H}^+]_t^2}\right).$$

Conversely, when DIC observations are available, the $p\mathrm{CO}_2^s$ implied by the reconstructed pH can be obtained by inverting the same carbonate-equilibrium relation. Specifically, given DIC observation $\mathcal{O}_{d,t}$, we compute

$$\hat{c}_t = \frac{\mathcal{O}_{d,t}}{K_{0,t}\left(1 + \frac{K_{1,t}}{[\mathrm{H}^+]_t} + \frac{K_{1,t}K_{2,t}}{[\mathrm{H}^+]_t^2}\right)}.$$

Here $\hat{c}_t$ denotes the $p\mathrm{CO}_2^s$ implied by reconstructed pH and DIC observation $\mathcal{O}_{d,t}$.

To verify the consistency of the formula-based implementation with PyCO2SYS, we compute DIC using both methods, with pH and sea surface CO_2 partial pressure from MOM6 as inputs. The mean relative difference between the formula-based DIC and the PyCO2SYS DIC is 3.736×10^{-3}. The small discrepancy indicates that the formula-based implementation is an appropriate substitution for PyCO2SYS, while remaining compatible with neural-network training.

C.2 Input and Output Details

REACT accepts an input tensor of shape $B \times C \times T_{\mathrm{in}} \times H \times W$, where B is the batch size, T_{in} is the number of input temporal frames, $H \times W$ is the spatial resolution, and $C = 11$ denotes the input channels. These channels consist of pH observations, $p\mathrm{CO}_2^s$ observations, and nine MOM6-derived variables. The model outputs a tensor of shape $B \times 1 \times T_{\mathrm{out}} \times H \times W$, representing the reconstructed sea surface pH sequence. For monthly-to-weekly reconstruction, $T_{\mathrm{out}} = 4T_{\mathrm{in}}$, while for weekly-to-daily reconstruction, $T_{\mathrm{out}} = 7T_{\mathrm{in}}$.

The initial state reconstruction module processes the full spatiotemporal input sequence with tubelet embedding. The input tensor is partitioned into $N = T_{\mathrm{in}}HW/(tpq)$ non-overlapping tubelets, where t, p, and q denote the temporal and spatial tubelet sizes. Each tubelet is flattened into a vector of size $Ctpq$ and projected into a latent space of dimension D. The ViViT-style encoder maps these tokens back to the latitude–longitude grid and produces the initial carbon-related transport state $z_0 \in \mathbb{R}^{B\times 1\times H\times W}$.

Starting from z_0, the conservative transport module produces a temporally resolved carbon-related state sequence $\{z_t\}_{t=1}^{T_{\mathrm{out}}} \in \mathbb{R}^{B\times 1\times T_{\mathrm{out}}\times H\times W}$. The source module then parameterizes unresolved non-conservative effects as frame-wise residual corrections to the transported state, yielding the corrected carbon-related state sequence $\{\tilde{z}_t\}_{t=1}^{T_{\mathrm{out}}} \in \mathbb{R}^{B\times 1\times T_{\mathrm{out}}\times H\times W}$.

The pH decoder maps the corrected carbon-related state sequence to sea surface pH under carbonate equilibrium conditions. The equilibrium constants K_1 and K_2 provide the nonlinear chemical link between pH and the carbonate system. The decoder outputs the reconstructed pH sequence $\hat{\mathbf{H}} = \{\hat{h}_t\}_{t=1}^{T_{\mathrm{out}}} \in \mathbb{R}^{B\times 1\times T_{\mathrm{out}}\times H\times W}$. The chemical module is used when carbonate-system variables are required for chemical constraints or evaluation, while the primary output of REACT remains the reconstructed sea surface pH sequence.

D Experiment Details

D.1 Datasets

WOD [1][5]. WOD is a publicly available global archive of quality-controlled ocean profile data maintained by NOAA NCEI. The WOD 2023 release contains scientifically quality-controlled ocean profile and plankton data collected over the global ocean from 1772 to 2022, and includes measurements of temperature, salinity, alkalinity, pH, nutrients, and other oceanographic variables. In this work, we use surface pH records from WOD 2023 to build sparse pH observations for sea surface pH reconstruction. The available records span January 1959–December 2022; we use January 1982–December 2022 because earlier observations are too sparse.

[5] https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html

SOCAT [22] (under CC BY 4.0 license)[6]. SOCAT is a community synthesis of quality-controlled surface ocean CO_2 observations, with the released data primarily reported as fugacity of CO_2 (fCO_2). These measurements are collected from ships, moorings, autonomous platforms, and drifting surface platforms over the global ocean and coastal seas. We use surface CO_2 records from SOCAT version 2023 as sparse carbon-system observations. The available records span January 1959–December 2022; we use January 1982–December 2022 because earlier observations are too sparse.

MOM6-COBALTv2 [21] (under LGPLv3 license)[7]. MOM6 is the sixth-generation Modular Ocean Model developed at NOAA GFDL for simulating ocean circulation, and COBALTv2 is the second-generation Carbon, Ocean Biogeochemistry, and Lower Trophics biogeochemical model used in GFDL Earth system modeling. We use MOM6-COBALTv2 outputs from January 1959 to December 2022 as dense gridded reference data. MOM6-COBALTv2 provides dense gridded ocean fields, including sea surface pH, surface CO_2, DIC, temperature, salinity, and velocity fields. All variables are regridded to a global $1^\circ \times 1^\circ$ grid. The daily fields are used to construct datasets for both monthly-to-weekly and weekly-to-daily reconstruction tasks.

D.2 Metrics

Normalized RMSE. We use normalized root mean square error (NRMSE) to measure the discrepancy between the reconstructed field and the ground-truth field. We first compute the latitude-weighted RMSE as

$$\mathrm{RMSE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\alpha_h\,(y_{thw}-u_{thw})^2},$$

where N is the number of time points, H and W denote the numbers of latitude and longitude grid points, respectively, y_{thw} is the ground-truth value at time t and grid location (h, w), and u_{thw} is the reconstructed value at time t and grid location (h, w). The latitude weight α_h accounts for the area variation across latitudes and is defined as

$$\alpha_h = \frac{\cos\phi_h}{\frac{1}{H}\sum_{h'=1}^{H}\cos\phi_{h'}},$$

where ϕ_h is the latitude of the h-th grid row. The NRMSE is then obtained by normalizing RMSE with the mean of the ground-truth values:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\frac{1}{NHW}\sum_{t=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} y_{thw}}.$$

pH. The reported pH NRMSE is computed between the reconstructed and ground-truth sea surface pH fields, given by

$$\mathrm{NRMSE}_{\mathrm{pH}} = \mathrm{NRMSE}(\hat{h}, h).$$

Here, h and $\hat{h}$ denote the ground-truth and reconstructed sea surface pH fields, respectively.

Chem. Let $\hat{h}$ be the reconstructed pH field, and let c and d be the reference pCO_2^s and DIC fields. We denote by $\Gamma_D(\hat{h}, c)$ the DIC implied by reconstructed pH and pCO_2^s, and by $\Gamma_C(\hat{h}, d)$ the pCO_2^s implied by reconstructed pH and DIC. The chemical consistency error is defined as

$$\mathrm{Chem.} = \frac{1}{2}\mathrm{NRMSE}\left(\Gamma_C(\hat{h}, d), c\right) + \frac{1}{2}\mathrm{NRMSE}\left(\Gamma_D(\hat{h}, c), d\right).$$

Drift$_{\mathrm{local}}$. Local drift measures the DIC inventory variation implied by the source-free branch within each prediction clip. Let $\hat{h}_{\mathrm{flow}}(t)$ be the source-free pH prediction at time t, and let c_t be the reference pCO_2^s field at the same time. The corresponding source-free DIC field is

$$\hat{d}_{\mathrm{flow}}(t) = \Gamma_D(\hat{h}_{\mathrm{flow}}(t), c_t).$$

[6] https://socat.info/index.php/version-2023/

[7] https://github.com/NOAA-GFDL/MOM6-examples

Let $M[\hat{d}_{\text{flow}}](t)$ denote the spatial integral of $\hat{d}_{\text{flow}}(t)$ over the ocean domain. For a clip i with time points $\{t_{i,1}, \ldots, t_{i,L}\}$, we define

$$D_i^{\text{loc}} = \frac{1}{L-1} \sum_{\ell=2}^{L} \frac{\left| M[\hat{d}_{\text{flow}}](t_{i,\ell}) - M[\hat{d}_{\text{flow}}](t_{i,1}) \right|}{\left| M[\hat{d}_{\text{flow}}](t_{i,1}) \right|}.$$

The reported local drift is averaged over all clips:

$$\text{Drift}_{\text{local}} = \frac{1}{C} \sum_{i=1}^{C} D_i^{\text{loc}},$$

where C is the number of clips.

Drift$_{\text{global}}$. Global drift measures the DIC inventory variation implied by the source-free branch over longer evaluation segments. For a segment s with time points $\{t_{s,1}, \ldots, t_{s,T_s}\}$, we define

$$D_s^{\text{glob}} = \frac{1}{T_s - 1} \sum_{j=2}^{T_s} \frac{\left| M[\hat{d}_{\text{flow}}](t_{s,j}) - M[\hat{d}_{\text{flow}}](t_{s,1}) \right|}{\left| M[\hat{d}_{\text{flow}}](t_{s,1}) \right|}.$$

Using the head and tail evaluation segments, the reported global drift is

$$\text{Drift}_{\text{global}} = \frac{1}{2} \left(D_{\text{head}}^{\text{glob}} + D_{\text{tail}}^{\text{glob}} \right).$$

D.3 Implementation Details

We train the model for 250 epochs using the AdamW optimizer [42] with momentum parameters $(0.9, 0.999)$. Cosine annealing is used for learning-rate scheduling. We select the learning rate and weight decay by grid search over $\{2, 1, 0.8, 0.5, 0.3, 0.1, 0.08\} \times 10^{-3}$ and $\{8, 5, 2, 1, 0.5, 0.1\} \times 10^{-3}$, respectively. The selected learning rate and weight decay are 1×10^{-3} and 5×10^{-4} for the MOM6 simulation experiments, and 1×10^{-4} and 5×10^{-4} for the WOD real-observation experiments. We tune the mass-conservation loss weight over $\{0.5, 0.1, 0.01, 0.001\}$. For the simulation experiments, we use a mass-conservation loss weight of 0.01 for monthly-to-weekly reconstruction and 0.1 for weekly-to-daily reconstruction. For the WOD real-observation experiments, we use a mass-conservation loss weight of 0.01. GradNorm [43] is applied in all experiments to adaptively balance the active loss terms. We set the GradNorm asymmetry parameter to 1.5, use a learning rate of 0.025 for the loss weights, and update the adaptive weights every 60 iterations. For the relative task-priority coefficients in GradNorm, we assign a larger coefficient to pH reconstruction and smaller coefficients to the carbonate-consistency terms, using 2 for pH and 0.5 for both pCO_2^s and DIC-based chemical constraints. We conduct all experiments on 8 NVIDIA 6000D GPUs with a total batch size of 8 for both datasets.

Transport-state controls. For pH alignment, we add a source-free decoding and alignment loss:

$$\hat{h}_t^{\text{flow}} = G_\theta(z_t, c_t, \xi_t), \qquad \mathcal{L}_{\text{align}} = \|z_t - \hat{h}_t^{\text{flow}}\|_{L^2}^2.$$

This encourages an identity mapping on the source-free branch while leaving the remaining modules and loss terms unchanged. For direct-pH and direct-DIC transport, we remove the Initial State Reconstruction Module and initialize the PDE with the exact pH or DIC field immediately before the target interval. This exact field is used only for PDE initialization and is masked before the source module and decoder, which otherwise retain the REACT inputs and architectures. For the parameter-matched direct-pH control, we enlarge the decoder to approximately match REACT's total parameter count while retaining the same source module and training objectives.

D.4 Sampling Strategies

Simulation Dataset. To simulate realistic sparse observations, we use an observation-guided sampling strategy rather than uniformly sampling over the entire ocean domain, as shown in Figure 7. For each variable, we first collect all grid locations that have valid observations at least once over the observational period. This historical observation footprint defines the candidate sampling region. Let N_{ocean} be the number of ocean grid points and let r be the target sampling ratio. At each time step, we sample approximately rN_{ocean} ocean grid points, rounded to the nearest integer. Sampling is first performed uniformly from the historical observation footprint of the corresponding variable. If the footprint contains fewer points than required by the target ratio, all footprint points are selected and the remaining points are sampled from nearby ocean grid cells around the footprint until the target number is reached. For pH, we use sampling ratios of 0.11%, 1%, and 10%. For $p\text{CO}_2^s$, we use a fixed sampling ratio of 1.67%, matching the sparsity level of SOCAT $p\text{CO}_2^s$ observations. This procedure preserves the spatial bias of real observational networks while allowing controlled comparisons under different sparsity levels.

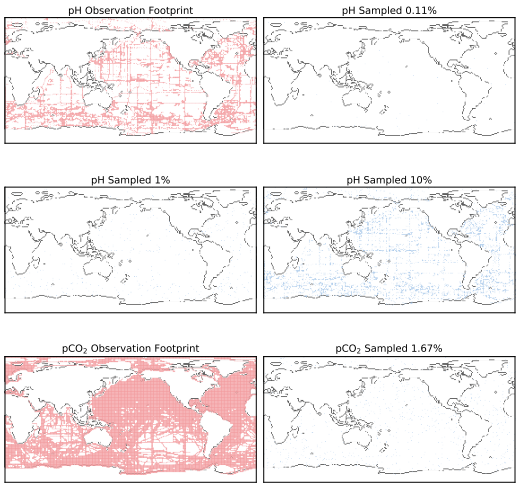

Figure 7: **Observation-guided sampling strategy.** Red points show historical observation footprints, defined as grid locations with at least one valid observation over the observational period. Blue points show sampled locations at one randomly selected time step under the target sampling ratios. pH is sampled at 0.11%, 1%, and 10%, while $p\text{CO}_2^s$ is sampled at 1.67%.

Real Observations. For real-observation experiments, we first identify valid observation locations where pH and $p\text{CO}_2^s$ are both available. Since directly available DIC observations are not provided, we compute observational DIC using PyCO2SYS from WOD pH, SOCAT $p\text{CO}_2^s$, salinity, and temperature. The intersection of valid pH, $p\text{CO}_2^s$, and computed DIC locations is used as the available observation set. For WOD/SOCAT and GLODAP, we first apply the chronological train/validation/test split in Table 1. Within each subset, we uniformly sample 75% of the available observation locations as input observations and hold out the remaining 25% as target observations. The input observations are used to construct the model input, while the held-out observations are used for evaluation.

D.5 Hyperparameter setting.

REACT[8]. The model hyperparameters of REACT are summarized in Table 6. Training hyperparameters are provided in Appendix D.3. The transport-state controls retain REACT's optimizer, learning rate, scheduler, loss weights, training epochs, and data splits, except for the additional alignment loss in the pH-alignment variant. Architectural and initialization changes are described in Appendix D.3.

NUTS [16] (under MIT license)[9]. The hyperparameters used for NUTS are listed in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 5×10^{-4} and the weight decay to 2×10^{-3} for all experiments. Cosine annealing is used for learning-rate scheduling.

4D-VarNet [15] (under CeCILL-C license)[10]. The hyperparameters used for 4D-VarNet are summarized in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 1×10^{-4} and the weight decay to 1×10^{-4} for all experiments. Cosine annealing is used for learning-rate scheduling.

Marble [13] (under MIT license)[11]. The hyperparameters used for Marble are summarized in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters

[8] https://anonymous.4open.science/r/REACT-67E7
[9] https://github.com/Leamonz/NUTS
[10] https://zenodo.org/records/8048541
[11] https://github.com/LeapLabTHU/GridMix

Table 6: Model hyperparameters used for REACT.

Hyper-param.	Meaning	Value
t	Temporal tubelet size in the ViViT initializer	1
p, q	Spatial patch size in the ViViT initializer	12, 12
D	Embedding dimension in the ViViT initializer	768
# Layers	Number of transformer layers in the ViViT initializer	12
# Heads	Number of attention heads in each transformer layer	12
MLP ratio	Expansion ratio of the MLP layers	4.0
# Source levels	Number of U-Net levels in the source module	4
# Decoder levels	Number of U-Net levels in the pH decoder	4
Base channels	Number of channels in the first U-Net level	64
Dropout	Dropout rate in the ViViT initializer	0.0
Attention dropout	Attention dropout rate in the ViViT initializer	0.0

$(0.9, 0.999)$. We set the learning rate to 1×10^{-3} and the weight decay to 5×10^{-3} for all experiments. Cosine annealing is used for learning-rate scheduling.

FNO [12] (under MIT license)[12]. The hyperparameters used for FNO are summarized in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 3×10^{-4} and the weight decay to 1×10^{-4} for all experiments. Cosine annealing is used for learning-rate scheduling.

U-Net [5] (under GPL-3.0 license)[13]. The hyperparameters used for U-Net are specified in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 8×10^{-4} and the weight decay to 5×10^{-3} for all experiments. Cosine annealing is used for learning-rate scheduling.

Vision Transformer (ViT) [6]. The hyperparameters used for ViT are configured as outlined in Table 7. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 5×10^{-4} and the weight decay to 1×10^{-3} for all experiments. Cosine annealing is used for learning-rate scheduling.

AtmoRep [11] (under MIT license)[14]. The hyperparameters used for AtmoRep are described in Table 7. We load pretrained weights from model ID `4nvwbetz`. The model is trained for 250 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 5×10^{-4} and the weight decay to 2×10^{-3} for all experiments. Cosine annealing is used for learning-rate scheduling.

Prithvi [10] (under MIT license)[15]. The hyperparameters used for Prithvi are described in Table 7. We load the official pretrained weights[16] and finetune the input adapter, output head, and last decoder block. The model is trained for 100 epochs using the AdamW optimizer with momentum parameters $(0.9, 0.999)$. We set the learning rate to 3×10^{-4} and the weight decay to 5×10^{-4} for all experiments. Cosine annealing is used for learning-rate scheduling.

D.6 Model efficiency

We further evaluate the computational efficiency of different models under the same hardware setting. All models are trained and evaluated on eight NVIDIA RTX 6000D GPUs with a per-GPU batch size of 1, corresponding to a total batch size of 8. Table 8 reports the number of parameters, training time per epoch, peak training VRAM, inference time on the test set, peak inference VRAM, and the corresponding reconstruction performance. REACT requires more computation than lightweight baselines because it jointly performs transport modeling, source correction, and carbonate-system regulation. Nevertheless, its cost remains substantially lower than large foundation-model baselines such as Prithvi, while achieving the best pH NRMSE and Chem. This indicates that REACT provides a favorable balance between reconstruction quality, chemical consistency, and computational cost.

[12] https://github.com/neuraloperator/neuraloperator
[13] https://github.com/milesial/Pytorch-UNet
[14] https://github.com/clessig/atmorep
[15] https://github.com/NASA-IMPACT/Prithvi-WxC
[16] https://huggingface.co/Prithvi-WxC/prithvi.wxc.2300m.v1

Table 7: Hyper-parameters used for all baseline models.

Baseline	Hyper-param.	Meaning	Value
NUTS	t	Tubelet size in the initializer	2
	p, q	Patch size in the transformer layers	18, 36
	D	Embedding dimension	768
	# Coarse layer	Number of transformer layers in the initializer	12
	# Refine layer	Number of transformer layers in the refinement module	6
	# Heads	Number of attention heads in each transformer layer	16
	MLP dimension	Hidden dimension of the MLP layers	2048
	# Blocks	Number of convolutional blocks in the source/diffusion network	[2, 2]
	Base channels	Number of channels in the first convolutional block of source/diffusion network	16
	Dropout(coarse)	Dropout rate in data-driven initializer	0.0
	Dropout(refine)	Dropout rate in refinement module	0.1
4D-VarNet	Shape data	Input channels of the model	11
	D	Hidden dimension for bilinear units	64
	# Blocks	Number of residual bilinear blocks	5
	Half kernel size	Convolution half-kernel sizes	1
	Subsample stride	Subsampling stride	1
	Dropout	Dropout rate applied in encoder	0.0
Marble	Grid size	Spatial resolution of modulation grid	8
	Grid base	Number of basis functions per grid cell	32
	Hidden dimension	Dimension of hidden layers in Siren [44] network	128
	# Layer	Number of layers in Siren [44] network	4
	Latent dimension	Dimension of latent vector	32
	Modulation dimension	Hidden dimension of modulation network	96
FNO	# Modes	Number of modes to keep in Fourier layer	16
	In Channels	Number of channels in input function	11
	Out Channels	Number of channels in output function	1
	Hidden Channels	Width of FNO	64
	Projection Channels	Number of channels in the projection block	128
	Lifting Channels	Number of channels in the lifting block	64
	# Layers	Number of Fourier layers	4
U-Net	Kernel size	Kernel size of encoder/decoder block	3
	Stride	Stride of encoder/decoder block	1
	Padding size	Padding size of encoder/decoder block	1
	Base channels	Number of channels in the first encoder block	64
	# Blocks	Number of encoder/decoder blocks	4
ViT	p, q	Patch size	12, 12
	D	Embedding dimension	768
	# Blocks	Number of ViT blocks	12
	# Heads	Number of attention heads	12
	MLP dimension	The hidden dimension of the MLP layers	3072
	Dropout	Dropout rate	0.0
AtmoRep	D	Embedding Dimension	2048
	# Encoder layer	Number of transformer layers in the encoder	10
	# Encoder heads	Number of attention heads in the encoder	16
	# Decoder layer	Number of transformer layers in the decoder	10
	# Decoder heads	Number of attention heads in the decoder	16
	# Tail nets	Number of networks in Tail Ensemble	16
	# Tail layer	Number of layers per network in Tail Ensemble (0 represents linear)	0
Prithvi	p, q	Patch size	2, 2
	D	Embedding Dimension	2560
	# Encoder layer	Number of transformer layers in the encoder	25
	# Decoder layer	Number of transformer layers in the decoder	5
	# Heads	Number of attention heads in each transformer layer	16
	MLP dimension	Hidden dimension of the MLP layers	10240
	Window size	Window size of attention layers	30, 32
	Drop path	For stochastic depth [45]	0.0
	Dropout	Dropout rate	0.0

For REACT, the reported inference time of 84.035 s includes both reconstruction and full-domain physicochemical evaluation. Reconstruction alone takes approximately 7.1 s; the remaining cost comes primarily from evaluating DIC conservation over the full spatiotemporal domain.

Table 8: Parameter, efficiency, and performance comparison of different models on MOM6 monthly-to-weekly task under the 0.11% sampling ratio. pH NRMSE is reported in units of 10^{-4}, and Chem. is reported in units of 10^{-2}.

Methods	Params	$\text{Time}_{\text{Train}}$ (s/epoch)	$\text{VRAM}_{\text{Train}}$ (GB)	$\text{Time}_{\text{Infer}}$ (s)	$\text{VRAM}_{\text{Infer}}$ (GB)	pH NRMSE	Chem.
4D-Var.	0.27M	4.654	5.444	0.513	0.821	$6.92_{\pm 0.03}$	$2.09_{\pm 0.03}$
Marble	0.83M	1.237	0.563	0.213	0.274	$25.7_{\pm 0.03}$	$5.65_{\pm 0.10}$
FNO	8.42M	2.711	0.757	0.254	0.380	$16.7_{\pm 0.07}$	$3.88_{\pm 0.07}$
U-Net	31.1M	1.544	1.173	0.190	0.612	$7.78_{\pm 0.04}$	$2.17_{\pm 0.08}$
ViT	0.11B	8.805	2.319	0.298	1.093	$9.56_{\pm 0.10}$	$2.51_{\pm 0.01}$
NUTS	0.15B	17.865	3.640	0.361	1.722	$6.74_{\pm 0.06}$	$1.95_{\pm 0.10}$
AtmoR.	0.80B	73.072	18.429	1.267	6.399	$14.6_{\pm 0.16}$	$3.52_{\pm 0.04}$
Prithvi	2.37B	1127.814	71.069	137.471	9.664	$16.0_{\pm 0.52}$	$3.48_{\pm 0.07}$
REACT	0.18B	100.745	32.087	84.035	6.118	$5.84_{\pm 0.04}$	$1.59_{\pm 0.04}$

Table 9: Decomposed chemical errors for the main reconstruction results in Table 2. The table reports the two components of the chemical consistency metric, i.e., $p\text{CO}_2^s$ and DIC errors, under different sampling ratios for monthly-to-weekly and weekly-to-daily sea surface pH reconstruction. Both errors are reported in units of 10^{-2}. All values are rounded to three significant digits. *Params* denotes model size. *U-Net*^{2} denotes a two-level U-Net variant.

Methods	**Params**	**Month→Week**						**Week→Day**					
		0.11%		1%		10%		0.11%		1%		10%	
		$p\text{CO}_2^s$	DIC	$p\text{CO}_2^s$	DIC	$p\text{CO}_2^s$	DIC	$p\text{CO}_2^s$	DIC	$p\text{CO}_2^s$	DIC	$p\text{CO}_2^s$	DIC
Kriging (Exp.)	–	10.3	9.82	7.33	6.78	4.86	4.26	10.2	9.73	7.16	6.72	4.45	3.99
Kriging (Sph.)	–	10.3	9.99	8.00	7.65	6.41	5.82	10.4	9.91	7.52	7.18	5.30	4.77
4D-VarNet	0.27M	2.28	1.91	2.18	1.82	2.06	1.74	1.93	1.79	1.95	1.81	2.12	1.98
Marble	0.83M	5.95	5.36	5.99	5.42	5.77	5.25	5.66	5.35	5.67	5.44	5.64	5.39
FNO	8.42M	4.07	3.69	4.00	3.63	4.01	3.63	2.14	2.00	2.37	2.21	2.31	2.16
U-Net	31.1M	2.32	2.02	2.27	2.01	2.24	1.99	1.93	1.79	1.91	1.78	1.90	1.78
U-Net^2	1.96M	2.35	2.07	2.37	2.08	2.22	2.00	2.10	1.95	2.14	1.98	2.00	1.86
ViT	0.11B	2.65	2.36	2.68	2.41	2.64	2.35	1.76	1.57	2.23	2.12	1.94	1.82
NUTS	0.15B	2.07	1.82	2.19	1.92	2.42	2.18	1.75	1.63	1.71	1.60	1.83	1.72
AtmoRep	0.80B	3.74	3.30	4.24	3.67	3.08	2.65	2.44	2.26	2.16	1.99	2.14	1.96
Prithvi	2.37B	3.83	3.13	3.76	3.14	3.45	2.88	2.92	2.58	2.91	2.58	2.91	2.54
REACT	0.18B	**1.73**	**1.45**	**1.65**	**1.40**	**1.67**	**1.41**	**1.44**	**1.27**	**1.42**	**1.28**	**1.40**	**1.25**

E Additional Experiments Results

E.1 Decomposed Chemical Errors for Table 2

Table 9 provides the $p\text{CO}_2^s$ and DIC components of the Chem. metric reported in Table 2. Across both monthly-to-weekly and weekly-to-daily reconstruction tasks, REACT achieves the lowest decomposed chemical errors under all sampling ratios. This confirms that the Chem. improvements in the main reconstruction results come from simultaneous reductions in both carbonate-system components, rather than from a single dominant term.

E.2 Additional Qualitative Results

This subsection provides the additional qualitative results referenced in Figure 3. Figure 8 compares the ground-truth and reconstructed sea surface pH, $p\text{CO}_2^s$, and DIC fields from different methods at one randomly selected MOM6 test time slice. Figure 9 shows the corresponding reconstruction errors, computed as prediction minus ground truth. The results show that REACT better preserves the spatial structures of the three variables and produces smaller residual errors than the baselines.

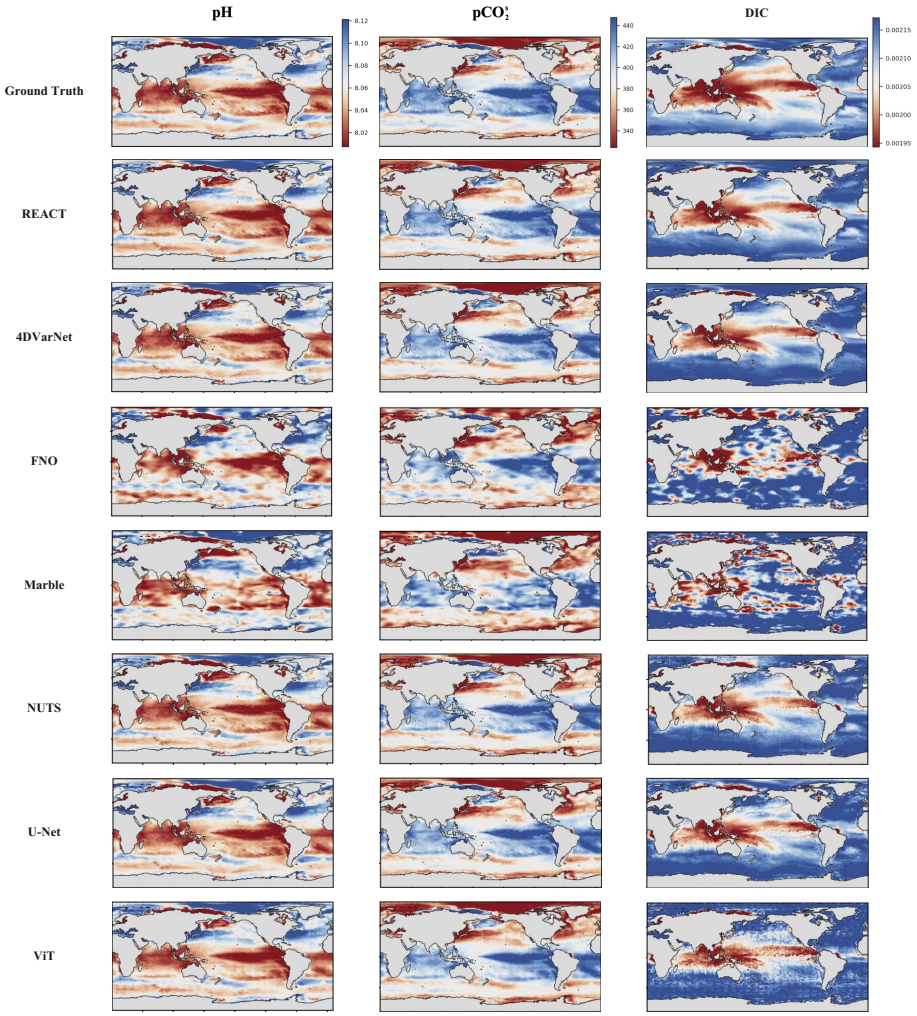

Figure 8: **Qualitative field comparison.** Ground-truth and reconstructed sea surface pH, pCO_2^s, and DIC fields from different methods at one randomly selected MOM6 test time slice.

E.3 Additional Real-Observation Results

This subsection provides the decomposed real-observation results corresponding to Figure 4. Figure 10 reports the NRMSE of pCO_2^s and DIC on the WOD real-observation experiment. Both metrics are reported in units of 10^{-1}. REACT achieves the lowest error on both carbonate variables, showing that its improvement on real observations is consistent not only for pH and Chem., but also for the two decomposed carbonate-system quantities.

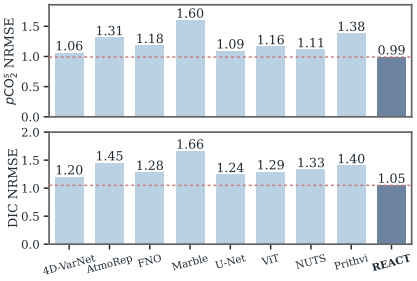

Figure 10: **Decomposed real-observation results.** NRMSE of pCO_2^s and DIC for different methods on real WOD observations. Both metrics are reported in units of 10^{-1}.

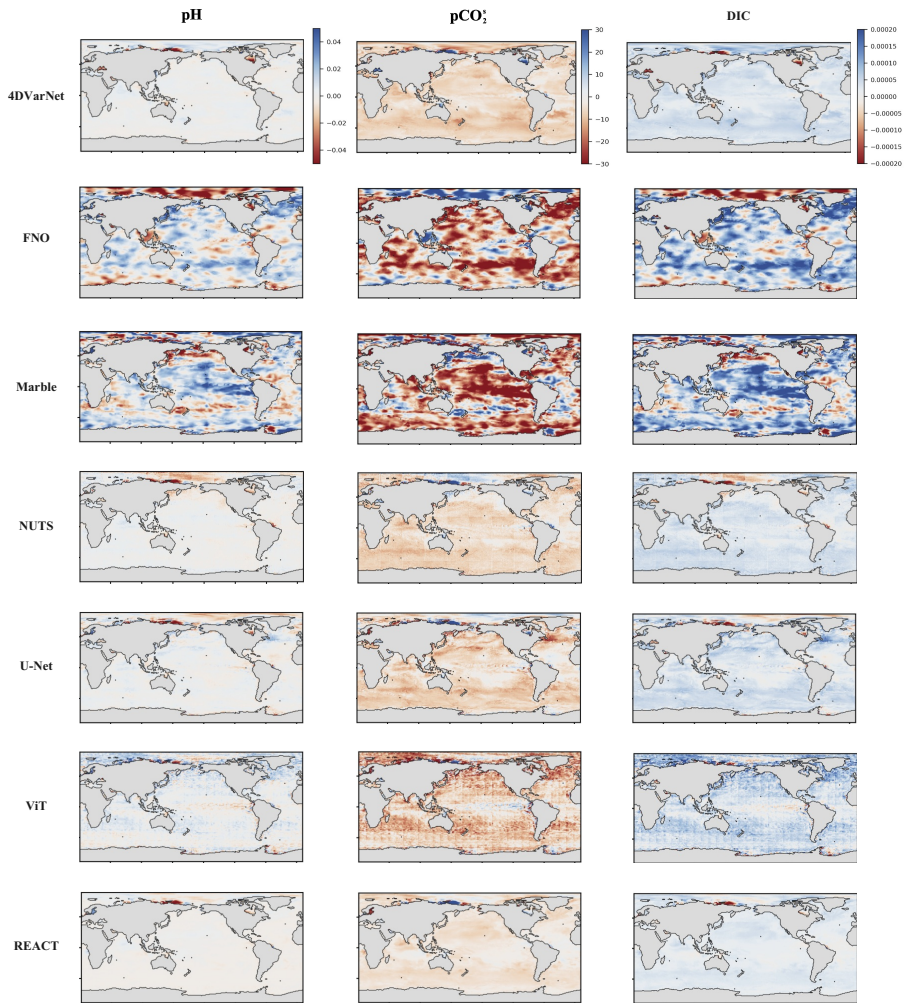

Figure 9: **Qualitative error maps.** Reconstruction errors of different methods for sea surface pH, pCO_2^s, and DIC at the same time slice, computed as prediction minus ground truth.

E.4 Decomposed Chemical Errors for Table 3

Table 10 provides the decomposed pCO_2^s and DIC errors for the source-module and constraint ablations reported in Table 3. The results show that the chemical constraint consistently reduces both carbonate-system components. For REACT on the monthly-to-weekly reconstruction task, adding $\mathcal{L}_{\text{chem}}$ reduces the pCO_2^s error from 2.06×10^{-2} to 1.73×10^{-2} and the DIC error from 1.82×10^{-2} to 1.45×10^{-2}. The same pattern is observed on the weekly-to-daily reconstruction task and in the U-Net ablation, confirming that the chemical constraint improves carbonate closure beyond the REACT architecture alone.

E.5 Decomposed Chemical Errors for Monthly-to-Weekly Design Ablations

Table 11 provides the decomposed pCO_2^s and DIC errors for the monthly-to-weekly design ablations reported in Table 4. The decomposed results show that most architectural choices have similar carbonate-system errors around the default setting, while removing key auxiliary variables causes larger degradation. In particular, removing sea surface pCO_2^s substantially increases both decomposed errors, indicating that sparse pCO_2^s observations are important for maintaining carbonate consistency.

Table 10: Decomposed chemical errors for the source module and constraint ablations. We report pCO_2^s and DIC errors for the ablation variants in Table 3, covering REACT on Month→Week and Week→Day reconstruction under the 0.11% sampling ratio, and the U-Net baseline ablation on Month→Week reconstruction under the 0.11% sampling ratio. Both metrics are reported in units of 10^{-2}. All values are rounded to three significant digits.

			REACT (Month→Week)		REACT (Week→Day)		U-Net (Month→Week)	
Source	$\mathcal{L}_{\text{mass}}$	$\mathcal{L}_{\text{chem}}$	pCO_2^s	DIC	pCO_2^s	DIC	pCO_2^s	DIC
–	–	–	8.94	Div.	2.84	2.62	2.27	1.98
✓	–	–	2.06	1.80	1.71	1.59	2.28	2.00
✓	✓	–	2.06	1.82	1.76	1.62	2.17	1.89
✓	✓	✓	1.73	1.45	1.44	1.27	1.77	1.53

Table 11: Decomposed chemical errors for the monthly-to-weekly ablation study in Table 4. The table reports pCO_2^s and DIC errors in units of 10^{-2} for interval length, state initializer depth, pH decoder/source-module depth, state initializer architecture, and auxiliary-variable ablations. The default setting uses interval length 4, state initializer depth 12, pH decoder/source-module depth 4, a ViViT state initializer, all auxiliary variables, and both mass and chemistry constraints. All values are rounded to three significant digits.

Factor	Setting	pCO_2^s	DIC
Interval length	1	1.71	1.44
	2	1.66	1.36
	4	1.73	1.45
	8	1.70	1.46
Initializer depth	4	1.73	1.46
	6	1.74	1.46
	12	1.73	1.45
	24	1.71	1.47
Decoder/source depth	2	1.66	1.42
	3	1.67	1.42
	4	1.73	1.45
	5	1.99	1.72

Factor	Setting	pCO_2^s	DIC
Initializer	ViT	1.66	1.40
	2D CNN	1.80	1.53
	3D CNN	7.83	916
	ViViT	1.73	1.45
Auxiliary removed	pCO_2^s	5.22	4.73
	pCO_2^a	1.72	1.39
	Chl.	1.84	1.63
	SSH	1.72	1.44
	MLD	1.72	1.42
	Salt	1.82	1.59
	Temp.	1.75	1.47
	u	1.75	1.46
	v	1.90	1.56

The 3D CNN initializer also becomes unstable, consistent with the divergent behavior reported in the main ablation table.

E.6 Weekly-to-Daily Design Ablations

Table 12 provides the weekly-to-daily counterpart of the monthly-to-weekly design ablations in Table 4. The default configuration still provides the best overall trade off. Although the model with an initializer depth of 24 achieves better overall performance than the default configuration, we retain a depth of 12 due to computational cost considerations. The auxiliary-variable ablations again show that removing pCO_2^s causes the largest degradation in both decomposed carbonate errors, confirming its importance for chemically consistent reconstruction across temporal scales.

E.7 Uncertainty Estimation

To evaluate predictive uncertainty, we start from the converged deterministic REACT checkpoint and freeze the reconstruction model. We then train an uncertainty head to estimate the standard deviation of a Gaussian predictive distribution around the deterministic prediction. The predictive mean is fixed by the deterministic model, while the uncertainty head learns the error scale around this mean. For the selected checkpoint, the one-standard-deviation coverage values are 0.821, 0.752, and 0.640 for pH, pCO_2^s, and DIC, respectively. The corresponding two-standard-deviation coverage values are 0.963, 0.971, and 0.958. These two-standard-deviation coverages are close to the nominal Gaussian coverage of 0.95, suggesting that the learned uncertainty intervals are broadly consistent with the observed reconstruction errors.

Table 12: Weekly-to-daily counterpart of the ablation study in Table 4 under the 0.11% sampling ratio. The table reports full metrics for interval length, state initializer depth, pH decoder/source-module depth, and state initializer architecture, with auxiliary-variable ablations shown separately using decomposed chemical errors. The default setting uses interval length 2, state initializer depth 12, pH decoder/source-module depth 4, a ViViT state initializer, all auxiliary variables, and both mass and chemistry constraints. pH NRMSE is reported in units of 10^{-4}, while pCO_2^s, DIC, and Chem. are reported in units of 10^{-2}. Drift values are reported in per mille. All values are rounded to three significant digits. *Div.* denotes divergent results.

Factor	Setting	pH	pCO_2^s	DIC	Chem.	$\text{Drift}_{\text{local}}$	$\text{Drift}_{\text{global}}$
Interval length	1	4.89	1.47	1.29	1.38	0.61‰	Div.
	2	4.64	1.44	1.27	1.35	0.63‰	7.68‰
	3	5.37	1.59	1.50	1.54	0.85‰	8.35‰
	4	4.79	1.35	1.19	1.27	0.77‰	3.59‰
Initializer depth	4	4.74	1.35	1.20	1.27	0.63‰	8.21‰
	6	4.80	1.33	1.18	1.26	0.60‰	7.96‰
	12	4.64	1.44	1.27	1.35	0.63‰	7.68‰
	24	4.56	1.40	1.26	1.33	0.60‰	7.79‰
Decoder/source depth	2	4.58	1.44	1.30	1.37	0.75‰	Div.
	3	4.89	1.33	1.16	1.25	0.64‰	Div.
	4	4.64	1.44	1.27	1.35	0.63‰	7.68‰
	5	4.91	1.25	1.10	1.17	0.53‰	12.1‰
Initializer	ViT	4.91	1.41	1.25	1.33	0.51‰	7.01‰
	2D CNN	4.76	1.33	1.18	1.26	0.54‰	Div.
	3D CNN	4.78	1.42	1.26	1.34	0.57‰	Div.
	ViViT	4.64	1.44	1.27	1.35	0.63‰	7.68‰

Metric	pCO_2^s	pCO_2^a	Chl.	SSH	MLD	Salt	Temp.	u	v
pCO_2^s	2.70	1.40	1.38	1.34	1.33	1.47	1.37	1.33	1.39
DIC	2.44	1.27	1.22	1.19	1.18	1.32	1.21	1.19	1.23

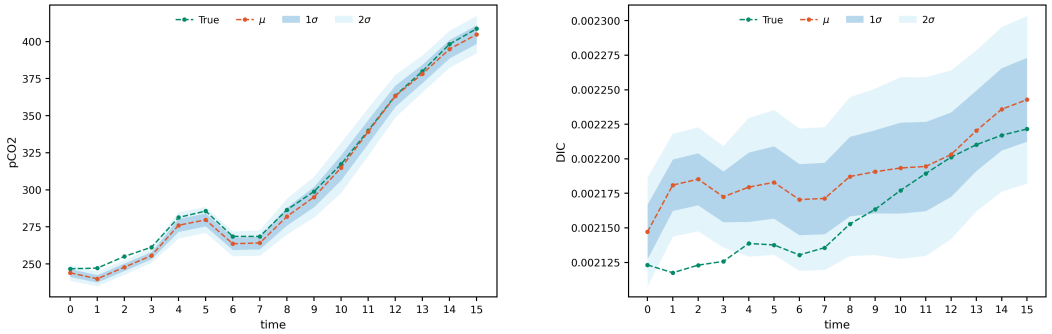

Figure 11: **Uncertainty estimation of REACT for pCO_2^s and DIC.** The predictive means are shown with one- and two-standard-deviation intervals.

E.8 Additional Observational and Sensitivity Experiments

GLODAP observations. We additionally evaluate pH reconstruction using GLODAP observations. We use the same chronological split as WOD: 1982–2018 for training, 2019–2020 for validation, and 2021–2022 for testing. Within each subset, 75% of the available observations are visible and 25% are held out, with auxiliary fields supplied by MOM6. Table 13 reports the results. REACT reduces pH NRMSE by 14.9% relative to NUTS and Chem. by 18.3% relative to U-Net, the strongest baseline for each metric. This extends the observational evaluation beyond WOD/SOCAT.

Table 13: **GLODAP results.** pH NRMSE is reported in units of 10^{-3}, and Chem. in 10^{-1}.

Method	pH	Chem.
NUTS	9.69	2.54
4DVarNet	11.4	1.97
U-Net	10.4	1.90
REACT	**8.25**	**1.55**

Dependence on auxiliary fields. The auxiliary fields in the WOD experiment come from MOM6. To assess this dependence, we mask surface pCO_2^s or mixed-layer depth (MLD) while retaining the rest of the reconstruction setting. Table 14 shows that removing pCO_2^s has a larger effect on both reconstruction and chemical consistency than removing MLD. Thus, the observational results benefit from model-based carbonate information, while being less sensitive to the removal of MLD.

Table 14: **WOD input ablation.** pH NRMSE is reported in units of 10^{-3}, and Chem. in 10^{-1}.

Setting	pH	Chem.
Full REACT	5.25	1.02
w/o pCO_2^s	6.40	1.33
w/o MLD	5.74	1.08

Sensitivity to chemical training labels. For one chemical target at a time, we perturb the normalized training labels as

$$\widetilde{y}_{\mathrm{norm}} = y_{\mathrm{norm}} + \alpha\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1), \qquad \alpha \in \{0.05, 0.20\}.$$

The two noise levels correspond to 5% and 20% of the target's training-set standard deviation. Model architectures, losses, data splits, optimization settings, and evaluation procedures are otherwise unchanged. Table 15 gives the perturbation scales in physical units and the resulting pH errors. Small perturbations increase pH NRMSE by 2.34% for pCO_2^s labels and 0.18% for DIC labels. At the larger noise level, the increases are 22.50% and 2.78%, respectively. The reconstruction is therefore more sensitive to severe corruption of pCO_2^s supervision than to DIC-label perturbations.

Table 15: **Training-label noise.** pH NRMSE: 10^{-4}.

Training label	Noise SD	pH
No noise	0	5.84
pCO_2^s, small	2.27 μatm	5.98
pCO_2^s, large	9.08 μatm	7.15
DIC, small	5.80 μmol kg^{-1}	5.85
DIC, large	23.20 μmol kg^{-1}	6.00

E.9 Surface pCO_2^s Reconstruction

We additionally reconstruct surface pCO_2^s as the target tracer, using pH as an auxiliary carbonate-system variable. For this task, Chem. is the average of the normalized pH and DIC errors. As shown in Table 16, REACT achieves the lowest error for the target and both associated carbonate variables. Relative to the strongest baseline for each metric, it reduces target NRMSE by 17.5% compared with NUTS and Chem. by 40.7% compared with U-Net. These results support applying REACT to another reconstruction target within the tested carbonate system; extension to other tracer systems remains future work.

Table 16: **Surface pCO_2^s reconstruction.** NRMSEs for pCO_2^s and DIC are reported in units of 10^{-2}, pH NRMSE in 10^{-4}, and Chem. in 10^{-3}.

Method	pCO_2^s	pH	DIC	Chem.
4DVarNet	1.38	49.5	2.07	12.8
NUTS	1.25	60.8	1.94	12.7
U-Net	1.48	49.1	2.00	12.4
REACT	**1.03**	**7.89**	**1.40**	**7.38**

F Limitations

A limitation of the REACT model is that the learned source module lacks full interpretability. Although the source term improves reconstruction by capturing non-conservative variations in the carbon cycle, the learned correction is not explicitly decomposed into individual physical or biogeochemical mechanisms, such as air–sea exchange, biological activity, or unresolved mixing. Consequently, while the source module provides a flexible way to account for processes that are not directly resolved by the transport equation, its attribution to underlying mechanisms remains limited.

Furthermore, REACT is currently designed for sea surface pH reconstruction and does not explicitly account for vertical carbon transport or subsurface carbonate dynamics. This limits its ability to characterize the three-dimensional structure of ocean carbonate chemistry, which is important for understanding carbon-cycle variability and marine biogeochemical processes. Extending REACT to three-dimensional ocean fields is therefore a promising direction for future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims and scope are clearly stated in the abstract and Introduction 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of REACT in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The full assumptions, formal statements, and proofs are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed experimental settings in Appendix D and hyperparameter configurations in Appendix D.5; we also provide an anonymous repository containing our code and data through this link.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code and data through an anonymous repository available at this link.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the data split protocol in Table 1 and report all training details in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the standard errors in the main results table, as shown in Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the hardware setting and computational cost details in Appendix D.6, including the GPU type, training time, inference time, and peak VRAM usage.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study follows the ethical guidelines and requirements specified by NeurIPS.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Conclusion 6, including the potential extension of REACT to other marine active tracers.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve data, models, or released assets with high misuse or dual-use risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: All code implementations are cited with license details in Appendix D.5. For the WOD datasets, although we were unable to identify their specific licenses, we cite their official sources and provide the official links.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the newly released code and data, together with documentation for reproducing the experiments, through an anonymized repository available at this link.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this work does not involve LLMs as important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.