CALM: CARBON-AWARE LEARNING FOR MARINE CHLOROPHYLL RECONSTRUCTION IN THREE DIMENSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing three-dimensional ocean chlorophyll from sparse observations is essential for understanding marine ecosystems and carbon cycling. Direct tracer reconstruction overlooks that chlorophyll is a regulated pigment whose abundance depends on both phytoplankton carbon biomass and pigment allocation. We introduce CALM, a deep learning framework that reconstructs chlorophyll through coupled organic and inorganic carbon states. Conditioned on sparse chlorophyll, dense inorganic carbon, and environmental drivers, separate effective transport operators preserve each carbon pool's global inventory, learned interconversion preserves their local sum, and surface and lower-boundary corrections account for external exchange. A constrained neural decoder uses a variable chlorophyll-to-carbon ratio to ensure that pigment carbon does not exceed available organic carbon. These properties follow from the architecture without conservation penalties. Experiments across three observation densities show improved reconstruction accuracy over the evaluated baselines, supported by carbon-budget audits and component ablations. At 10% sampling, CALM reduces NRMSE and RMSE by 11.6% and 19.5% over the strongest baseline.

1 INTRODUCTION

Chlorophyll is a central indicator of marine primary producers and their role in the carbon cycle. Reconstructing its three-dimensional distribution would improve our understanding of subsurface productivity and ecosystem variability. Yet in-situ profiles are sparse, and satellite ocean color mainly constrains the near-surface ocean. Recovering the full water column therefore requires combining sparse biological observations with physical and biogeochemical context.

Deep learning offers flexible priors for this inverse problem. Convolutional networks, spatiotemporal transformers, and neural operators can exploit the relationships between chlorophyll, hydrography, circulation, and nutrient supply (Ronneberger et al., 2015; Arnab et al., 2021; Gao et al., 2022; Li et al., 2021). Transport-based methods such as NUTS (Zheng et al., 2025) add dynamical structure by propagating nutrient fields and learning corrections. Applying that structure directly to chlorophyll leaves a biological mismatch: the observed pigment is carried by phytoplankton whose carbon biomass and pigment allocation evolve differently (Geider et al., 1997). Chlorophyll can change through photoacclimation without a corresponding change in carbon biomass. A direct chlorophyll transport model with an unrestricted source correction therefore leaves carbon fixation, remineralization, and pigment regulation entangled, even when its reconstruction error is low.

We introduce CALM, a framework that reconstructs chlorophyll through coupled organic and dissolved inorganic carbon states. Three mechanisms impose complementary structure. Separate OC and DIC transport operators learn pool-specific effective diffusivities and preserve each pool's global inventory in a closed domain. A learned interconversion transfers carbon between the pools while preserving their sum at every location. Explicit boundary corrections account for changes in the represented domain's total carbon. These identities follow from the architecture, without conservation penalties. A constrained neural decoder then maps organic carbon to chlorophyll using a variable molar ratio, with an upper bound derived from the carbon content of chlorophyll a. Thus the model distinguishes redistribution, interconversion, external exchange, and pigment allocation.

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

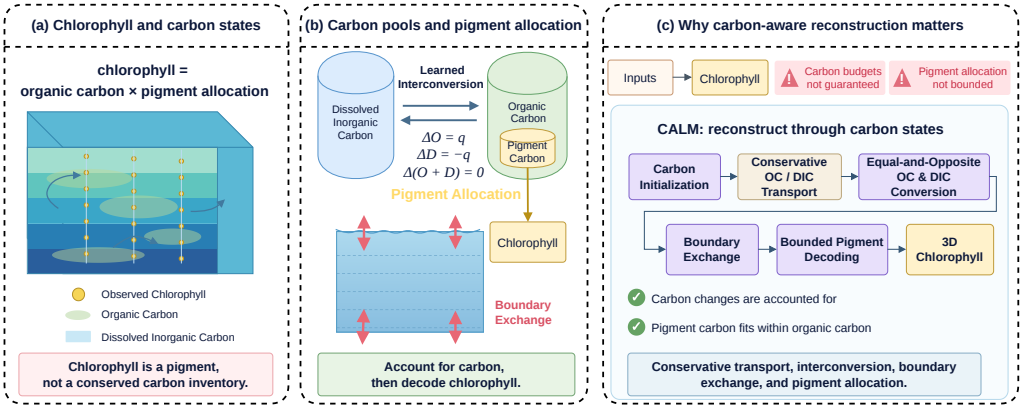


Figure 1: **Carbon-aware chlorophyll reconstruction.** Sparse pigment observations and environmental context constrain coupled carbon states and pigment allocation.

This structure addresses a limitation of simple factorized decoders. We call a UNet3D that predicts organic carbon, DIC, and a positive greenness coefficient **UNet3D (Factorized)**. Such a model can still violate carbon budgets or assign more carbon to pigment than the organic pool contains. The identity "chlorophyll equals organic carbon times greenness" guarantees neither property. CALM adds explicit constraints on the state evolution and on the ratio, while allowing all observed inputs to condition its spatiotemporal representation.

Figure 1 summarizes the reconstruction setting and the carbon-accounting motivation. Our contributions are:

- A carbon-aware formulation of sparse 3D chlorophyll reconstruction that separates globally conservative transport, locally conservative interconversion, and boundary exchange.
- A biologically constrained neural decoder with an explicit pigment-carbon budget, together with continuous integral proofs of the resulting conservation and feasibility properties.
- A monthly benchmark at three observation densities with three matched seeds, separating reconstruction accuracy from carbon-budget and biological-consistency audits. Component ablations distinguish predictive utility from structural guarantees and carbon-state identification.

2 Related Work

Assimilation methods and inversion models. Objective mapping, variational assimilation, and ensemble filtering combine sparse ocean observations with statistical or dynamical priors (Conkright et al., 2002; Courtier et al., 1994; Nerger & Gregg, 2008). Variational inverse methods recover three-dimensional structure through regularized fitting (Brasseur & Haus, 1991; Belgacem et al., 2021). Their dependence on background estimates and iterative optimization motivates learned conditional reconstruction with explicit scientific structure.

Neural Network Black-box models. UNet3D, ViViT, and Earthformer learn spatial and temporal dependencies (Çiçek et al., 2016; Arnab et al., 2021; Gao et al., 2022); CoModGAN and RePaint address generative completion (Zhao et al., 2021; Lugmayr et al., 2022). Atmospheric foundation models offer pretrained representations for adaptation (Lessig et al., 2023; Schmude et al., 2024). These priors do not themselves constrain the carbon budget underlying chlorophyll. UNet3D (Factorized) tests whether predicting carbon and pigment factors alone suffices.

Physics-informed models. Physics-informed objectives penalize equation residuals (Raissi et al., 2017), and 4DVarNet learns variational interpolation (Beauchamp et al., 2023). Neural operators and implicit fields provide useful representations without automatic conservation (Li et al., 2021; Wang et al., 2025; Luo et al., 2024). NUTS combines two-scale transport and corrections for surface

nutrients (Zheng et al., 2025). Direct pigment transport leaves biomass turnover and chlorophyll acclimation entangled (Geider et al., 1997). CALM instead constrains carbon transport and interconversion structurally, then decodes pigment with a bounded ratio. Appendix A expands these connections.

3 Problem Formulation

Notations. Let Ω be the wet three-dimensional ocean domain, $\mathbf{x} = (\lambda, \varphi, z) \in \Omega$ a location, and $t \in [t_1, t_T]$ time. We write ∂_t, ∇, and $\nabla\cdot$ for the time derivative, spatial gradient, and divergence in physical coordinates. Chlorophyll is $y(\mathbf{x}, t)$, dense dissolved inorganic carbon (DIC) is d, and environmental drivers $\mathbf{e}$ include the three-dimensional velocity $\mathbf{u} = (u, v, w)$. At sampled times, $m_t \in \{0, 1\}$ marks sparse chlorophyll observations $o_t = m_t y_t$; the mask distinguishes missing values from observed zeros. For a carbon concentration C in mol C kg^{-1}, its represented inventory is

$$\mathcal{I}[C] = \rho_0 \int_\Omega C(\mathbf{x})\,\mathrm{d}V, \tag{1}$$

where ρ_0 is the constant reference density. We distinguish organic carbon O, inorganic carbon D, and their sum $O + D$ from the pigment y. Under the chlorophyll-a-equivalent convention, y is in $\mu\mathrm{g\,kg}^{-1}$ and $A = 893.49 \times 10^6\,\mu\mathrm{g\,mol}^{-1}$ converts chlorophyll moles to mass. The molar chlorophyll-to-organic-carbon ratio is g, giving $y = AOg$.

Reconstruction problem. Given $(\mathbf{e}_{1:T}, d_{1:T}, o_{1:T}, m_{1:T})$, reconstruct the complete chlorophyll fields $\hat{y}_{1:T}$. Each sample contains $T = 4$ consecutive monthly frames. The benchmark provides full simulated CHL and DIC labels for supervision; only the CHL inputs are sparsified. For standardized fields z^y, z^D and the joint valid-label mask M_t, define the volume-weighted reconstruction error

$$\mathcal{E}(\hat{z}, z) = \frac{\sum_{t=1}^{T} \int_\Omega M_t(\mathbf{x})[\hat{z}_t(\mathbf{x}) - z_t(\mathbf{x})]^2\,\mathrm{d}V}{\sum_{t=1}^{T} \int_\Omega M_t(\mathbf{x})\,\mathrm{d}V}. \tag{2}$$

We seek a reconstruction with low $\mathcal{E}(\hat{z}^y, z^y)$ whose underlying carbon states also admit consistent accounting: source-free transport preserves global inventories, internal conversion preserves $O + D$ locally, and external exchange accounts for net carbon changes. Pigment allocation must satisfy $55\hat{y}/A \le O$, since a chlorophyll a molecule contains 55 carbon atoms. These are constraints on the model's represented states, not extra observations of organic carbon. Appendix D.1 specifies the grid, masks, and units.

4 Methodology

4.1 A Tracer Baseline

Transport followed by a source correction. A natural starting point is the transport-plus-correction principle of NUTS (Zheng et al., 2025), whose surface-nutrient model combines a learned initializer, homogenized transport, mesoscale refinement, and a source module. A three-dimensional adaptation initializes chlorophyll from the observed window, propagates it with advection–diffusion, and adds a learned residual:

$$\begin{aligned} \hat{y}_1^{\mathrm{tr}} &= F_0(\mathbf{e}_{1:T}, d_{1:T}, o_{1:T}, m_{1:T}), \\ \hat{y}^{\mathrm{tr}}(\mathbf{x}, t) &= \hat{y}^{\mathrm{tr}}(\mathbf{x}, t_1) + \int_{t_1}^{t} \left[-\nabla\cdot(\mathbf{u}\hat{y}^{\mathrm{tr}}) + \nabla\cdot(K_y \nabla \hat{y}^{\mathrm{tr}})\right](\mathbf{x}, \tau)\,\mathrm{d}\tau, \\ \hat{y}_t &= \hat{y}_t^{\mathrm{tr}} + s_t^y. \end{aligned} \tag{3}$$

Here K_y and s_t^y are learned; the residual can absorb unresolved sources, sinks, and fine-scale corrections. This is a conceptual adaptation of the tracer modeling strategy, not a reported NUTS experiment on 3D chlorophyll.

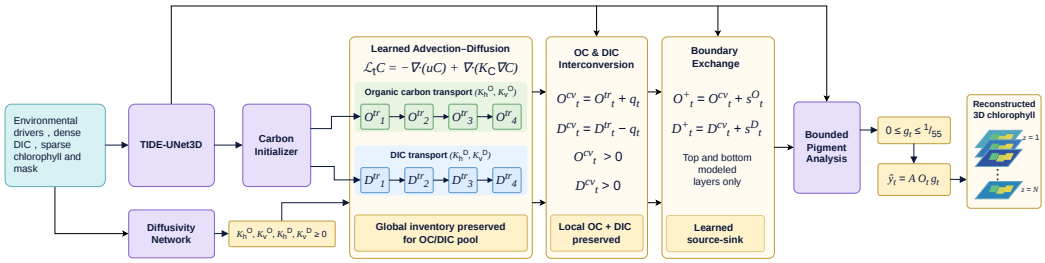


Figure 2: **Overview of CALM.** Separate carbon transport, local interconversion, boundary exchange, and bounded pigment decoding condition chlorophyll reconstruction.

Why it fails. The mismatch is the transported state. Chlorophyll is a regulated pigment carried by phytoplankton, not a standalone conserved inventory. Its concentration reflects both carbon biomass and a variable pigment-to-carbon ratio, which responds to light, nutrients, and temperature (Geider et al., 1997). Even without transport, $\partial_t y = A(g\,\partial_t O + O\,\partial_t g)$: pigment can change through acclimation while carbon remains unchanged. Conversely, conserving $\int_\Omega y\,\mathrm{d}V$ does not conserve $\mathcal{I}[O+D]$. An unrestricted pigment residual cannot distinguish carbon fixation or remineralization from a change in pigment allocation, nor require opposing OC and DIC transfers. Advection–diffusion can be part of a chlorophyll balance, but directly applying a tracer model to chlorophyll provides no carbon-accounting mechanism. Appendix B.5 gives elementary counterexamples. This structural limitation motivates changing the state on which transport and sources act.

4.2 CALM: Carbon-Aware Learning for Marine Chlorophyll

From pigment to coupled carbon states. Phytoplankton biomass is linked to organic carbon; fixation and remineralization connect organic carbon to DIC. CALM therefore reconstructs two non-negative carbon pools, transports them with separate effective diffusivities, exchanges carbon locally, and represents external exchange at the surface and lower boundary. A variable pigment ratio then decodes chlorophyll. This is a reduced model of represented organic and inorganic carbon, rather than an explicit food-web model. The constraints are built into the neural computation, so their validity does not depend on minimizing a physical penalty.

Spatiotemporal representation and initialization. TIDE-UNet3D (Figure 5) combines depth-preserving spatial encoding with temporal attention to produce context $h_{1:T} = F_\theta(\mathbf{e}_{1:T}, d_{1:T}, o_{1:T}, m_{1:T})$. A positive initializer gives $(O_1^{\mathrm{tr}}, D_1^{\mathrm{tr}}) = I_\theta^+(h_{1:T})$. Figure 2 shows the framework; Appendix C details the backbone.

Two transport operators with global conservation. For $C \in \{O, D\}$, source-free transport satisfies $\partial_t C^{\mathrm{tr}} + \nabla\cdot\mathbf{J}_C = 0$, with $\mathbf{J}_C = \mathbf{u}C^{\mathrm{tr}} - K_C\nabla C^{\mathrm{tr}}$ and $K_C = \mathrm{diag}(K_{C,h}, K_{C,h}, K_{C,v})$. A shared network predicts separate non-negative OC and DIC diffusivities, allowing pool-specific unresolved redistribution under the same resolved velocity. These are effective closures, not molecular constants, and need not differ everywhere. For $C_\Sigma = O^{\mathrm{tr}} + D^{\mathrm{tr}}$, the total flux is $\mathbf{u}C_\Sigma - K_D\nabla C_\Sigma - (K_O - K_D)\nabla O^{\mathrm{tr}}$; tying the coefficients removes its partition-dependent term. Periodic longitude and zero outward flux on closed boundaries give

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}[C^{\mathrm{tr}}] = -\rho_0\int_{\partial\Omega}\mathbf{J}_C\cdot\mathbf{n}\,\mathrm{d}S = 0, \qquad C\in\{O,D\}. \tag{4}$$

The numerical implementation uses a positivity-preserving update followed by a global inventory correction. Thus each pool preserves its inventory up to floating-point error; this does not certify locally conservative numerical advection. The complete four-frame source-free trajectories are computed before applying the following framewise corrections.

Organic–inorganic conversion with local conservation. A head conditioned on context, transported carbon, and dense DIC predicts $a_t \in [-1, 1]$. With $[a]_+ = \max(a, 0)$, define the donor-limited transfer

$$q_t = [a_t]_+ D_t^{\mathrm{tr}} - [-a_t]_+ O_t^{\mathrm{tr}}, \qquad O_t^{\mathrm{cv}} = O_t^{\mathrm{tr}} + q_t, \quad D_t^{\mathrm{cv}} = D_t^{\mathrm{tr}} - q_t. \tag{5}$$

Positive q_t transfers DIC to OC; negative q_t reverses the direction. Both pools remain non-negative, with $O_t^{\mathrm{cv}} + D_t^{\mathrm{cv}} = O_t^{\mathrm{tr}} + D_t^{\mathrm{tr}}$ at every location. This represents net carbon conversion without identifying a specific biological reaction rate.

Surface and lower-boundary sources and sinks. External exchange changes the modeled carbon inventory. Let χ_∂ mark wet cells in the top and bottom depth layers. A boundary head predicts $\beta_t^C \in [-\eta, \eta]$, with $\eta = 0.25$, and applies $s_t^C = \chi_\partial \beta_t^C C_t^{\mathrm{cv}}$ and $C_t^+ = C_t^{\mathrm{cv}} + s_t^C$ for $C \in \{O, D\}$. These signed, boundary-supported concentration corrections provide a reduced representation of exchange with the atmosphere, surface inputs, and the deeper ocean. They preserve non-negativity and explicitly account for net carbon change:

$$\mathcal{I}[O_t^+ + D_t^+] - \mathcal{I}[O_t^{\mathrm{tr}} + D_t^{\mathrm{tr}}] = \mathcal{I}[s_t^O + s_t^D]. \tag{6}$$

Corrections act independently on each transported frame, rather than accumulating as time-integrated fluxes. They are supported on finite-thickness boundary layers; no external correction is applied in the interior.

Biologically constrained chlorophyll decoding. Greenness g is the molar chlorophyll-to-organic-carbon ratio. At observed cells with $O_t^+ > 0$, $g_t^{\mathrm{obs}} = o_t/(AO_t^+)$ supplies ratio information. A mask-normalized analysis network combines sparse ratios, context, and carbon features to predict $r_t = R_\theta(h_t, O_t^+, D_t^+, g_t^{\mathrm{obs}}, m_t) \in [0, 1]$, giving $g_t = r_t/55$ and $\hat{y}_t = AO_t^+ g_t$. Consequently, the pigment-carbon concentration satisfies

$$C_t^{\mathrm{pigment}} = 55\frac{\hat{y}_t}{A} = 55O_t^+ g_t \le O_t^+. \tag{7}$$

This necessary elemental constraint permits variable pigment allocation without assigning more carbon to chlorophyll than the organic pool contains. It is not a complete physiological model. Appendix B proves the conservation and feasibility properties and explains why unconstrained factorization cannot guarantee them.

Learning objective and relation to NUTS. Training minimizes $\mathcal{L} = \mathcal{E}(\hat{z}^y, z^y) + \lambda_D \mathcal{E}(\hat{z}^D, z^D)$, with $\lambda_D = 0.05$, using cell-volume quadrature of Equation (2) over training samples. NUTS motivates transport followed by learned corrections; CALM changes the state and the correction structure for a biological target. Separate carbon transport, equal-and-opposite conversion, boundary-supported exchange, and bounded pigment allocation give the network explicit ocean-carbon meaning. Global accounting, local cancellation, and pigment feasibility hold by construction, without adding conservation or feasibility losses. Supervision still determines reconstruction quality and helps anchor the inorganic state.

5 Experiment

We assess reconstruction at varying observation densities, audit the represented carbon budgets and pigment allocation, and use component ablations to determine what each constraint contributes.

5.1 Experimental Setup

Monthly benchmark. We use monthly MOM6–COBALTv2 output (Adcroft et al., 2019; Stock et al., 2020; Dunne et al., 2020) from January 1980 through December 2022. The grid has 180×360 horizontal cells and 12 depth levels from 2.5 to 312.5 m. Four-month windows are split chronologically into 1980–2010 for training (369 windows), 2011–2016 for validation (69), and 2017–2022 for testing (69). Depth-stratified sampling retains 0.1%, 1%, or 10% of jointly valid chlorophyll voxels. Seeds $\{0, 1, 42\}$ determine fixed observation masks and neural training randomness. Repeated calendar months in overlapping windows share their masks. Appendix D.1 details the protocol.

Baselines and optimization. The suite includes exponential and spherical kriging, 4DVarNet, Marble, FNO, UNet3D, ViViT, AtmoRep, and Prithvi WxC. The additional UNet3D (Factorized) predicts the same carbon/pigment factors without hard conservation or a greenness upper bound. Neural methods receive dense environmental inputs, dense DIC, sparse chlorophyll, and its mask;

Table 1: **Monthly 3D chlorophyll reconstruction.** Mean ± sample SD over seeds 0, 1, and 42; lower is better. Physical RMSE is in $10^{-2}\,\mu\mathrm{g\,kg^{-1}}$.

Method	Params	0.1% observed		1% observed		10% observed	
		NRMSE	RMSE	NRMSE	RMSE	NRMSE	RMSE
Kriging (Exp.)	–	$0.755_{\pm 0.00329}$	$11.8_{\pm 0.0651}$	$0.586_{\pm 0.00222}$	$9.05_{\pm 0.00870}$	$0.383_{\pm 0.000178}$	$5.60_{\pm 0.00614}$
Kriging (Sph.)	–	$0.756_{\pm 0.00312}$	$11.9_{\pm 0.0620}$	$0.587_{\pm 0.00318}$	$9.08_{\pm 0.0162}$	$0.382_{\pm 0.000198}$	$5.55_{\pm 0.00571}$
4DVarNet	2.06M	$0.275_{\pm 0.00142}$	$5.05_{\pm 0.0348}$	$0.265_{\pm 0.00146}$	$4.84_{\pm 0.0369}$	$0.187_{\pm 0.00110}$	$3.32_{\pm 0.0608}$
Marble	3.02M	$0.630_{\pm 0.00404}$	$10.3_{\pm 0.121}$	$0.627_{\pm 0.00374}$	$10.2_{\pm 0.131}$	$0.621_{\pm 0.0109}$	$10.0_{\pm 0.393}$
FNO	22.6M	$0.268_{\pm 0.000861}$	$4.42_{\pm 0.0245}$	$0.267_{\pm 0.000541}$	$4.41_{\pm 0.0418}$	$0.245_{\pm 0.000737}$	$4.03_{\pm 0.0216}$
UNet3D	71.7M	$\underline{0.199_{\pm 0.00408}}$	$\underline{3.08_{\pm 0.0410}}$	$\underline{0.195_{\pm 0.00229}}$	$\underline{3.04_{\pm 0.0209}}$	$\underline{0.163_{\pm 0.00208}}$	$\underline{2.53_{\pm 0.0132}}$
UNet3D (Factorized)	71.7M	$0.204_{\pm 0.00147}$	$3.21_{\pm 0.0354}$	$0.201_{\pm 0.00303}$	$3.18_{\pm 0.0273}$	$0.171_{\pm 0.00179}$	$2.71_{\pm 0.135}$
ViViT	144M	$0.428_{\pm 0.000485}$	$5.64_{\pm 0.0154}$	$0.428_{\pm 0.00129}$	$5.66_{\pm 0.0458}$	$0.428_{\pm 0.000406}$	$5.66_{\pm 0.0148}$
AtmoRep	525M	$0.356_{\pm 0.000982}$	$4.82_{\pm 0.0395}$	$0.353_{\pm 0.000407}$	$4.84_{\pm 0.00400}$	$0.334_{\pm 0.000308}$	$4.47_{\pm 0.00995}$
Prithvi WxC	2.37B	$0.233_{\pm 0.00189}$	$3.62_{\pm 0.00818}$	$0.233_{\pm 0.00190}$	$3.61_{\pm 0.00676}$	$0.219_{\pm 0.00227}$	$3.38_{\pm 0.0153}$
CALM	**68.0M**	$0.186_{\pm 0.00100}$	$2.99_{\pm 0.0483}$	$0.179_{\pm 0.00327}$	$2.76_{\pm 0.0862}$	$0.144_{\pm 0.000909}$	$2.04_{\pm 0.0440}$
Promotion	–	**6.34%**	**2.85%**	**8.01%**	**8.97%**	**11.6%**	**19.5%**

kriging uses chlorophyll observations only. CALM has 68.0M parameters and uses AdamW for up to 300 epochs. Each checkpoint is selected by validation CHL NRMSE. All methods are evaluated at all three sampling rates, and results are reported as mean ± sample standard deviation over seeds 0, 1, and 42. Parameter counts include both trainable and frozen weights. Appendix D.3 records the baseline protocols; Appendix D.5 gives their hyperparameters.

Reconstruction metrics. We report volume-weighted RMSE in physical space and after training-set channel/depth standardization:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_j V_j M_j (\hat{y}_j - y_j)^2}{\sum_j V_j M_j}}, \qquad \mathrm{NRMSE} = \sqrt{\frac{\sum_j V_j M_j (\hat{z}_j^y - z_j^y)^2}{\sum_j V_j M_j}}. \tag{8}$$

Here j indexes test samples, frames, and spatial cells; V_j is cell volume and M_j the joint label-validity mask. NRMSE uses layerwise training-set standardization, so it has no fixed proportional relationship to physical RMSE. Observation-derived and simulated chlorophyll use separate training statistics because their means and variances differ.

5.2 Monthly Reconstruction at Three Observation Densities

Table 1 shows that CALM improves both metrics over all baselines at all three densities. The direct-output UNet3D is the strongest baseline in each of the six metric columns. Relative to its mean errors, CALM reduces NRMSE by 6.34%, 8.01%, and 11.6% at 0.1%, 1%, and 10% sampling, respectively, and physical RMSE by 2.85%, 8.97%, and 19.5%. These columnwise relative reductions are reported in the Promotion row.

Increasing observation density improves CALM's mean NRMSE from approximately 0.186 to 0.144. Dense DIC is supplied to all neural models, so these experiments evaluate chlorophyll reconstruction with carbon context rather than joint sparse reconstruction of both tracers.

Figure 3 compares spatial errors at three depths. In this selected slice, CALM reduces volume-weighted RMSE relative to UNet3D by 28.0%, 21.3%, and 17.5% at 10, 100, and 200 m, respectively. These maps illustrate a favorable case; the aggregate comparisons remain those in Table 1.

5.3 Component Analysis

Transport and interconversion serve different roles. Removing learned diffusion increases CHL NRMSE from 0.186 to 0.192, corresponding to a 3.0% reduction for the full model relative to this ablation. DIC error does not improve in the same direction. Removing interconversion gives slightly lower CHL NRMSE (0.185), but increases physical DIC RMSE from 1.06 to 8.43 μmol C kg^{-1}. Interconversion therefore chiefly supports a faithful inorganic-carbon state in this comparison. Disabling exchange trivially preserves the pool sum, so its zero conversion residual alone is not evidence of a useful carbon model.

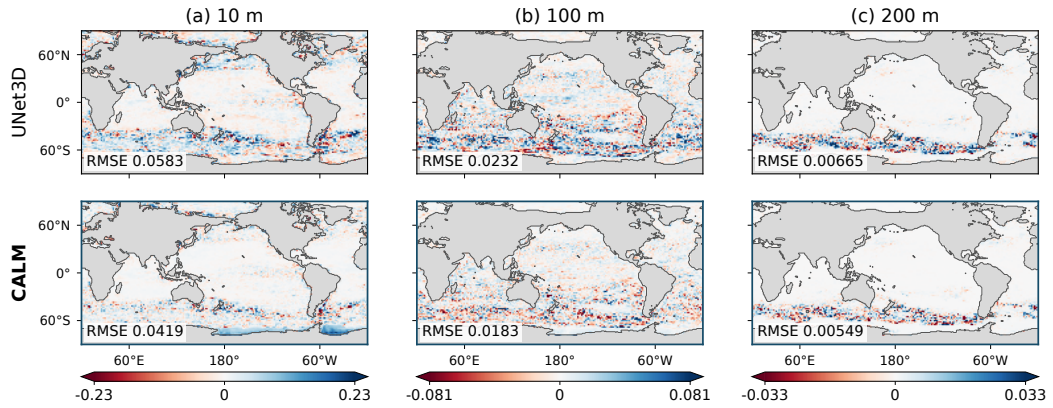


Figure 3: **Qualitative error maps.** UNet3D and CALM at 10, 100, and 200 m in October 2019 (10% sampling, seed 42); errors are in μg kg^{-1}.

Table 2: **Component analysis.** At 0.1% sampling, seed 42; CHL RMSE is in 10^{-2} μg kg^{-1} and DIC RMSE in μmol C kg^{-1}.

	Components / objective				Reconstruction error		
Variant	K	q	s	λ_D	CHL NRMSE	CHL RMSE	DIC RMSE
Full CALM	✓	✓	✓	0.05	0.186	2.98	1.06
No diffusion	–	✓	✓	0.05	0.192	3.05	0.980
No conversion	✓	–	✓	0.05	0.185	2.99	8.43
No boundary source	✓	✓	–	0.05	0.187	2.96	1.10
No DIC loss	✓	✓	✓	0.00	0.183	2.87	8.05×10^4
$\lambda_D = 0.01$	✓	✓	✓	0.01	0.188	3.05	2.39
$\lambda_D = 0.10$	✓	✓	✓	0.10	0.188	3.02	0.870

Boundary terms allow accounted external exchange. Removing boundary sources changes CHL NRMSE only slightly (0.186 to 0.187) and gives a slightly lower physical CHL RMSE. Their demonstrated role is explicit treatment of boundary corrections, rather than a uniform gain across accuracy metrics. Conservation should subtract the declared sources; requiring the final carbon sum to remain constant in their presence would impose a different physical assumption.

DIC supervision anchors the inorganic state. Without DIC supervision, chlorophyll error decreases, but physical DIC RMSE rises to approximately 8.05×10^4 μmol C kg^{-1}. The loss-weight sweep illustrates the tradeoff: $\lambda_D = 0.10$ improves DIC RMSE to 0.870 μmol C kg^{-1}, with a small increase in CHL NRMSE. DIC supervision therefore anchors the inorganic state.

Greenness-bound sensitivity. The unbounded decoder also attains 100% BCR on the evaluated data. Indeed, all bounded and unbounded runs have $f_{\max} < 4.2 \times 10^{-5}$, far below the feasibility limit of 1 . The bound thus supplies a structural certificate, while these experiments do not demonstrate reduced empirical violation rates or a uniformly better accuracy tradeoff. The 0.1% ablation favors the bounded decoder, but the 1% single-seed comparison favors the unbounded one.

5.4 Ablation Study on Design Choices

TIDE-UNet3D achieves the lowest CHL NRMSE (0.186) among matched-size backbones, compared with 0.192 for UNet3D, 0.301 for ViViT, and 0.260 for a 3D CNN (Table 3a). This comparison evaluates the complete backbone rather than temporal attention alone.

Removing shortwave radiation, mixed-layer depth, or temperature increases NRMSE by 2.33%, 1.86%, and 1.78%, respectively (Table 3b), consistent with the roles of light and water-column structure. Silicate and wind-v removal slightly reduce error. Appendix D.6 gives the single-seed protocol.

Table 3: **Design ablations.** CHL NRMSE at 0.1% sampling, seed 42: (a) matched-size backbones; (b) auxiliary inputs removed from both neural pathways.

(a) Backbone architecture

Backbone	TIDE-UNet3D	UNet3D	ViViT	3D CNN
NRMSE	0.186	0.192	0.301	0.260

(b) Auxiliary input removed

Removed	None	NO_3	PO_4	SiO_4	Temp.	Salt	Wind u	Wind v	SW rad.	MLD	SSH
NRMSE	0.186	0.187	0.187	0.186	0.190	0.187	0.187	0.185	0.191	0.190	0.187

Table 4: **Paired carbon and consistency audit.** CALM and UNet3D (Factorized) at 0.1% sampling, seed 42.

Diagnostic	CALM	UNet3D (Factorized)
Within each four-month window		
BCR (%)	100	100
$E_{\mathrm{unacc}} \downarrow$	4.72×10^{-8}	1.08×10^{-3}
Initial-frame unaccounted variation $\downarrow$	6.80×10^{-8}	1.94×10^{-3}
DIC inventory-change error $\downarrow$	2.00×10^{-5}	1.18×10^{-4}
Across overlapping windows / stitched months		
CHL relative spread (%) $\downarrow$	1.59	5.62
DIC relative spread (%) $\downarrow$	0.0137	0.0711

5.5 Carbon Accounting and Cross-Window Consistency

Across nine CALM runs, mean relative transport drift is of order 10^{-8}, the maximum local conversion residual is 2.38×10^{-7} mol C kg^{-1}, and the source-accounted relative inventory error is below 2.1×10^{-7}. All runs attain 100% biochemical consistency rate (BCR). Definitions and full internal audits appear in Appendix D.8.

Table 4 converts UNet3D's mass greenness to the same molar convention. Both models achieve 100% BCR. CALM has much smaller within-window unaccounted inventory variation after subtracting its declared source corrections.

Across overlapping windows, CALM reduces CHL and DIC relative spread, with each window initialized independently.

5.6 Predictive Uncertainty

A learned scale head estimates Gaussian uncertainty around each frozen reconstruction (Figure 4). At 10% sampling, CALM reduces CRPS by 21.9% relative to UNet3D, with 94.1% coverage for nominal 95% intervals. Appendix E.1 defines the metrics, protocol, and calibration limitations.

5.7 Transfer to Argo-Derived Chlorophyll

After simulation pretraining, we fine-tune on an interpolated Argo-derived product (Appendix E.2). At 10% sampling, CALM achieves NRMSE 0.0663 and RMSE 0.0264 μg kg^{-1} for interpolated-product reconstruction (Table 5).

6 Conclusion

CALM reconstructs three-dimensional chlorophyll through coupled carbon states and a constrained pigment decoder. Conservative transport, local interconversion, and explicit boundary corrections make its carbon accounting auditable, while the bounded molar ratio guarantees that pigment carbon fits within the organic pool. Monthly experiments at three observation densities show improved reconstruction over the evaluated learning-based baselines. Component and backbone ablations separate chlorophyll accuracy from DIC fidelity and structural feasibility. A paired audit confirms

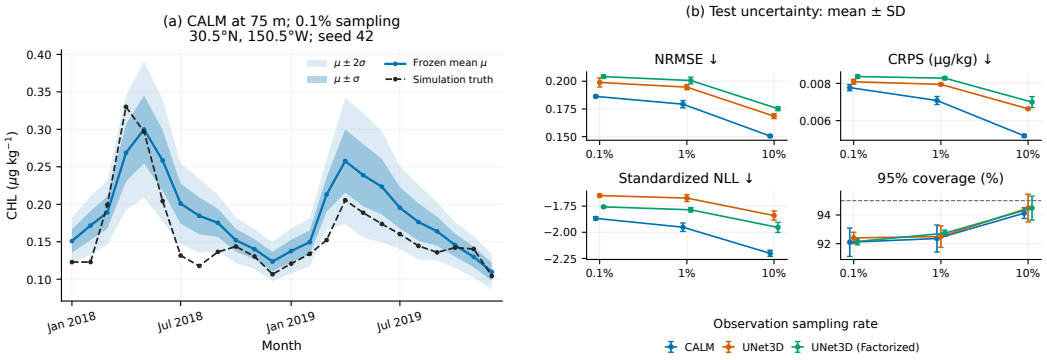


Figure 4: **Predictive uncertainty.** Left: CALM's local reconstruction with uncertainty bands. Right: test metrics across sampling rates; error bars denote three-seed SD.

Table 5: **Argo-derived product reconstruction.** Volume-weighted held-out metrics at 10% sampling, seed 42. RMSE is in μg kg^{-1}; lower is better.

Method	NRMSE	RMSE	Method	NRMSE	RMSE
CALM	**0.0663**	**0.0264**	AtmoRep	0.1869	0.0580
UNet3D	0.0790	0.0297	FNO	0.1942	0.0620
UNet3D (Factorized)	0.0982	0.0341	ViViT	0.3022	0.0973
4DVarNet	0.1012	0.0352	Marble	0.3679	0.1177
Prithvi WxC	0.1901	0.0569			

small source-accounted variation within windows; uncertainty estimation and Argo-derived product reconstruction extend the evaluation.

AI Assistance

Generative AI assisted writing and language polishing, literature retrieval and discovery, and research ideation and execution, including code development, evaluation, and visualization. AI-assisted outputs were checked against source literature, experiment records, and executable analyses. The authors take responsibility for the final text, claims, and artifacts.

REFERENCES

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. doi: 10.1029/2019MS001726.

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*, pp. 6836–6846, 2021.

Olivier Aumont, Christian Ethé, Alessandro Tagliabue, Laurent Bopp, and Marion Gehlen. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development*, 8:2465–2513, 2015. doi: 10.5194/gmd-8-2465-2015. URL https://gmd.copernicus.org/articles/8/2465/2015/.

Dorothee C. E. 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 fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT). *Earth System Science Data*, 8:383–413, 2016. doi: 10.5194/essd-8-383-2016. URL https://essd.copernicus.org/articles/8/383/2016/.

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. doi: 10.5194/gmd-16-2119-2023.

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. doi: 10.5194/essd-13-5915-2021. URL https://essd.copernicus.org/articles/13/5915/2021/.

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. doi: 10.1016/0924-7963(91)90006-G. URL https://www.sciencedirect.com/science/article/abs/pii/092479639190006G.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016. doi: 10.1007/978-3-319-46723-8_49.

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. Internal Report 17, National Oceanographic Data Center, 2002. URL https://repository.library.noaa.gov/view/noaa/1174.

Philippe Courtier, Jean-Noël 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.

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. doi: 10.3389/fmars.2024.1151868. URL https://doi.org/10.3389/fmars.2024.1151868.

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. In *International Conference on Learning Representations*, 2021.

John P. Dunne, L. W. Horowitz, A. J. Adcroft, P. Ginoux, I. M. Held, J. G. John, John P. Krasting, Sergey Malyshev, V. Naik, F. Paulot, et al. The GFDL earth system model version 4.1 (GFDL-ESM4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11):e2019MS002015, 2020. doi: 10.1029/2019MS002015.

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.

Richard J. Geider, Hugh L. MacIntyre, and Todd M. Kana. Dynamic model of phytoplankton growth and acclimation: responses of the balanced growth rate and the chlorophyll a:carbon ratio to light, nutrient-limitation and temperature. *Marine Ecology Progress Series*, 148:187–200, 1997. doi: 10.3354/meps148187.

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 preprint arXiv:2308.13280*, 2023.

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*, pp. 10758–10768, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_MAT_Mask-Aware_Transformer_for_Large_Hole_Image_Inpainting_CVPR_2022_paper.html.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33219–33242. PMLR, 2024. URL https://proceedings.mlr.press/v235/lu24n.html.

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*, pp. 11461–11471, 2022.

Xihaier Luo, Wei Xu, Yihui Ren, Shinjae Yoo, and Balu Nadiga. Continuous field reconstruction from sparse observations with implicit neural networks. *arXiv preprint arXiv:2401.11611*, 2024. URL https://arxiv.org/abs/2401.11611.

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, Dan Seidov, Alexandra I. Grodsky, and James G. Beauchamp. World ocean database 2023. NOAA Atlas NESDIS 97, NOAA National Environmental Satellite, Data, and Information Service, 2024. URL https://repository.library.noaa.gov/view/noaa/66204.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. EdgeConnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. URL https://arxiv.org/abs/1901.00212.

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. doi: 10.1016/j.jmarsys.2007.09.007.

Margaret A. Oliver and Richard Webster. Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4(3):313–332, 1990. doi: 10.1080/02693799008941549.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

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 preprint arXiv:2409.13598*, 2024.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.nips.cc/paper_files/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html.

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. doi: 10.1029/2019MS002043.

Ariane Verdy and Matthew R. Mazloff. A data assimilating model for estimating Southern Ocean biogeochemistry. *Journal of Geophysical Research: Oceans*, 122(9):6968–6988, 2017. doi: 10.1002/2016JC012650. URL https://doi.org/10.1002/2016JC012650.

Honghui Wang, Shiji Song, and Gao Huang. GridMix: Exploring spatial modulation for neural fields in PDE modeling. In *International Conference on Learning Representations*, 2025.

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.

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. In *Advances in Neural Information Processing Systems*, 2025.

A Related Work

A.1 Oceanographic Data

Observations. Ocean reconstruction must reconcile heterogeneous measurements with uneven spatial and temporal coverage. The World Ocean Database assembles quality-controlled profiles of physical, chemical, and biological variables, including chlorophyll (Mishonov et al., 2024). SOCAT instead curates measurements of surface-ocean CO_2 fugacity (fCO_2), supporting studies of air–sea carbon exchange (Bakker et al., 2016). These products constrain different parts of the marine system: surface carbon measurements do not provide a volumetric chlorophyll field, and individual profiles leave extensive gaps between samples. This motivates using environmental context alongside sparse pigment observations. Neither archive supplies the dense chlorophyll labels used in our experiments.

Simulations. Coupled circulation and biogeochemical models produce internally related fields of transport, nutrients, carbon, and ecosystem state. Examples include MOM6-based ocean configurations (Adcroft et al., 2019), the COBALT ecosystem model used within GFDL-ESM4.1 (Stock et al., 2020; Dunne et al., 2020), and PISCES-v2 (Aumont et al., 2015). Their dense outputs permit controlled observation masking and evaluation throughout the water column. Our benchmark uses monthly MOM6–COBALTv2 fields for this purpose. The resulting experiments test reconstruction within a simulated system; transfer to measured chlorophyll must additionally address measurement errors, observation selection, and simulation bias. CALM adopts a reduced carbon-accounting structure rather than reproducing every plankton compartment or reaction in the data-generating simulator.

A.2 Assimilation and Inversion Methods

Assimilation. Objective analyses such as the World Ocean Atlas provide gridded summaries of sparse ocean measurements (Conkright et al., 2002). Variational assimilation optimizes agreement with observations around a model trajectory (Courtier et al., 1994); ensemble methods estimate state corrections from a sampled error distribution. Assimilation of SeaWiFS ocean-color data with a local SEIK filter illustrates the latter approach and the importance of bias correction (Nerger & Gregg, 2008). At the coupled-system level, B-SOSE estimates Southern Ocean biogeochemistry by fitting a dynamical model to observations (Verdy & Mazloff, 2017). These methods can propagate information through physical and ecological relationships, but their estimates inherit the chosen dynamics, error statistics, and optimization cost. CALM uses learned, window-conditioned inference and is not evaluated as an operational sequential assimilation system.

Inversion. Variational inverse models estimate spatial fields by balancing observational fit against regularity and other prescribed constraints. Early three-dimensional applications analyzed coupled physical and biological observations in the Bering and Chukchi seas (Brasseur & Haus, 1991). More recently, variational mapping produced a multidepth climatology of nitrate, phosphate, and silicate in the western Mediterranean (Belgacem et al., 2021). Such methods are valuable where sampling is irregular, although a smooth reconstruction alone cannot determine the biological processes behind an observed pigment concentration. Our formulation addresses that ambiguity by distinguishing carbon redistribution, organic–inorganic transfer, boundary exchange, and pigment allocation, while retaining an explicit limitation on latent-state identification.

A.3 Data-Driven Reconstruction Methods

Spatiotemporal prediction and pretraining. The U-Net family encodes multiscale spatial features with convolutional skip connections (Ronneberger et al., 2015). ViT and ViViT introduce image and video attention architectures (Dosovitskiy et al., 2021; Arnab et al., 2021), while Earthformer uses cuboid attention for Earth-system forecasting (Gao et al., 2022). These architectures offer reusable priors for conditional field reconstruction. AtmoRep and Prithvi WxC extend this approach with large-scale atmospheric pretraining (Lessig et al., 2023; Schmude et al., 2024); their application to subsurface chlorophyll requires explicit channel, geometry, and output adaptations. We therefore report the actual adapted architectures, trainable subsets, and schedules in Appendix D.5. Their results here characterize these adaptations rather than their original weather tasks.

Completion and continuous representations. Image completion provides several ways to condition predictions on partially observed data: CoModGAN uses conditional co-modulation, EdgeConnect predicts edge structure, MAT uses mask-aware attention, and RePaint conditions a pretrained diffusion model during sampling (Zhao et al., 2021; Nazeri et al., 2019; Li et al., 2022; Lugmayr et al., 2022). These studies concern images; their visual completion criteria do not establish carbon consistency for ocean fields. Fourier neural operators learn mappings between functions (Li et al., 2021); SIREN parameterizes continuous signals using periodic activations (Sitzmann et al., 2020). GridMix adds spatial modulation to neural fields (Wang et al., 2025), and implicit networks have also been studied for field reconstruction from sparse observations (Luo et al., 2024). Our Marble baseline adapts GridMix to a three-dimensional grid-modulated representation. Such representational choices can improve expressivity without, on their own, enforcing the integral identities required of the represented carbon pools.

A.4 Physics-Informed Models

Scientific knowledge can enter learned reconstruction through an objective, an inference algorithm, or the structure of the state update. Physics-informed neural networks use differential-equation residuals during training (Raissi et al., 2017). CLOINet combines optimal interpolation with learned clustering to integrate remote-sensing and sparse in-situ information (Cutolo et al., 2024), and 4DVarNet learns components of a variational interpolation scheme (Beauchamp et al., 2023). Graph-based approaches offer another route: MeshGraphNets learns mesh dynamics (Pfaff et al., 2021), while OxyGenerator uses a spatiotemporal graph model tailored to ocean dissolved-oxygen reconstruction (Lu et al., 2024). These designs encode different forms of prior information and should not all be interpreted as exact conservation mechanisms.

NUTS is particularly relevant because its two-scale formulation combines transport with learned corrections for surface nutrients (Zheng et al., 2025). A direct extension to chlorophyll can redistribute the observed scalar and fit its residual changes, but it does not separate phytoplankton carbon from pigment regulation. Physiological models explicitly allow chlorophyll-to-carbon ratios to respond to light, nutrients, and temperature (Geider et al., 1997). CALM consequently applies transport to separate OC and DIC states, constrains their local transfer to cancel, and accounts for external exchange at the boundaries. Distinct learned effective diffusivities allow pool-specific unresolved redistribution; they are not interpreted as measurements of molecular diffusion. A bounded pigment decoder connects the carbon state to chlorophyll. These mechanisms enforce the stated accounting identities without a conservation-loss weight, while leaving unresolved the unique identification of latent biomass and a complete physiological description.

B THEORY

We state the guarantees using continuous fields and volume integrals. The transport theorem describes the ideal continuous operator; the implemented global rescaling enforces its inventory identity numerically. The conversion and decoder results are algebraic and apply directly at each grid cell as well. We make no claim that a one-step monthly numerical trajectory exactly solves the continuous PDE.

B.1 SETUP AND ASSUMPTIONS

Let Ω be a fixed wet ocean domain with volume measure $\mathrm{d}V$ and piecewise smooth boundary. Longitude is periodic. Let $\rho_0 > 0$ be constant, and define

$$\mathcal{I}_{\mathcal{U}}[C] = \rho_0 \int_{\mathcal{U}} C \,\mathrm{d}V, \qquad \mathcal{I}[C] = \mathcal{I}_{\Omega}[C] \tag{9}$$

for any measurable subdomain $\mathcal{U} \subseteq \Omega$. Suppose the source-free fields $O^{\mathrm{tr}}, D^{\mathrm{tr}}$ are non-negative and sufficiently regular for the divergence theorem, with fluxes $\mathbf{J}_O, \mathbf{J}_D$ satisfying the source-free continuity equation in Section 4.2. The net outward flux is zero for each pool: periodic contributions cancel and $\mathbf{J}_C \cdot \mathbf{n} = 0$ on other closed boundaries. Non-negative continuous solutions are assumed; the discrete implementation enforces non-negativity separately.

At each reconstruction time t, a measurable transfer q obeys $-O^{\mathrm{tr}} \le q \le D^{\mathrm{tr}}$, and the converted fields are $O^{\mathrm{cv}} = O^{\mathrm{tr}} + q$, $D^{\mathrm{cv}} = D^{\mathrm{tr}} - q$. Boundary corrections are supported in a designated boundary layer $\Omega_\partial \subset \Omega$, representing the top and bottom finite-volume layers, and satisfy $|s^C| \le \eta C^{\mathrm{cv}}$ with $0 \le \eta < 1$. This is a volume correction on a layer of positive thickness, not a nonzero volume integral over a zero-measure boundary surface. Finally, $O^+ = O^{\mathrm{cv}} + s^O$, $D^+ = D^{\mathrm{cv}} + s^D$, $0 \le g \le 1/55$, and $\hat{y} = AO^+g$.

B.2 GLOBAL TRANSPORT AND LOCAL INTERCONVERSION

Proposition B.1 (Source-free carbon inventory)**.** *Under the stated boundary conditions, each carbon pool and their sum preserve their global inventories:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}[C^{\mathrm{tr}}(t)] = 0, \qquad \mathcal{I}[O^{\mathrm{tr}}(t) + D^{\mathrm{tr}}(t)] = \mathcal{I}[O^{\mathrm{tr}}(0) + D^{\mathrm{tr}}(0)]. \tag{10}$$

Proof. For either $C = O$ or $C = D$, integrate the transport equation and apply the divergence theorem:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}[C^{\mathrm{tr}}] = \rho_0 \int_{\Omega} \partial_t C^{\mathrm{tr}} \,\mathrm{d}V \tag{11}$$

$$= -\rho_0 \int_{\Omega} \nabla \cdot \mathbf{J}_C \,\mathrm{d}V = -\rho_0 \int_{\partial\Omega} \mathbf{J}_C \cdot \mathbf{n} \,\mathrm{d}S = 0. \tag{12}$$

Summing the two identities and integrating in time proves the result. Separate diffusivities for the two pools do not alter the argument. □

Proposition B.2 (Local total-carbon conservation and positivity)**.** *Donor-limited conversion preserves non-negativity and the total carbon in every subdomain $\mathcal{U} \subseteq \Omega$:*

$$O^{\mathrm{cv}}, D^{\mathrm{cv}} \ge 0, \qquad \mathcal{I}_{\mathcal{U}}[O^{\mathrm{cv}} + D^{\mathrm{cv}}] = \mathcal{I}_{\mathcal{U}}[O^{\mathrm{tr}} + D^{\mathrm{tr}}]. \tag{13}$$

Proof. The bounds on q immediately give $O^{\mathrm{tr}} + q \ge 0$ and $D^{\mathrm{tr}} - q \ge 0$. The signed-fraction construction in Equation (5) satisfies those bounds: for $a \ge 0$, $q = aD^{\mathrm{tr}}$; for $a < 0$, $q = -|a|O^{\mathrm{tr}}$. For any $\mathcal{U}$,

$$\rho_0 \int_{\mathcal{U}} \big[(O^{\mathrm{tr}} + q) + (D^{\mathrm{tr}} - q)\big] \,\mathrm{d}V = \rho_0 \int_{\mathcal{U}} (O^{\mathrm{tr}} + D^{\mathrm{tr}}) \,\mathrm{d}V. \tag{14}$$

The cancellation is pointwise, so the identity holds at arbitrary spatial resolution. This local statement concerns interconversion; transport can move carbon across the boundary of $\mathcal{U}$. □

B.3 Boundary Accounting and Independent Frame Corrections

Proposition B.3 (Source-accounted total-carbon identity)**.** *The corrected pools remain non-negative. With $S = s^O + s^D$,*

$$\mathcal{I}[O^+(t) + D^+(t)] = \mathcal{I}[O^{\mathrm{tr}}(0) + D^{\mathrm{tr}}(0)] + \mathcal{I}[S(t)]. \tag{15}$$

For $\mathcal{U} \subseteq \Omega \setminus \Omega_\partial$, the source correction vanishes identically and local total carbon equals the transported total at that time.

Proof. Since $s^C \geq -\eta C^{\mathrm{cv}}$, $C^+ \geq (1-\eta)C^{\mathrm{cv}} \geq 0$. Adding the two corrected fields, integrating, and applying Propositions B.1 and B.2 gives

$$\rho_0 \int_\Omega (O^+ + D^+)\,\mathrm{d}V = \rho_0 \int_\Omega (O^{\mathrm{tr}} + D^{\mathrm{tr}} + S)\,\mathrm{d}V \tag{16}$$

$$= \mathcal{I}[O^{\mathrm{tr}}(0) + D^{\mathrm{tr}}(0)] + \rho_0 \int_{\Omega_\partial} S\,\mathrm{d}V. \tag{17}$$

On an interior subdomain $\mathcal{U}$, the final integral has no contribution. □

For independent frame corrections, subtraction of Equation (15) at two times gives

$$\mathcal{I}[O^+(t_2) + D^+(t_2)] - \mathcal{I}[O^+(t_1) + D^+(t_1)] = \mathcal{I}[S(t_2) - S(t_1)]. \tag{18}$$

If transport is only approximately conservative, its inventory difference must also be retained on the right. In particular, the right side is not $\int_{t_1}^{t_2} \mathcal{I}[S(t)]\,\mathrm{d}t$: $S(t)$ is a frame correction in concentration units, not a source rate. This distinction is essential for auditing the implemented reconstruction.

For comparison, a physical continuous evolution with source rates σ_O, σ_D and interconversion rate f would satisfy

$$\partial_t O + \nabla\cdot \mathbf{J}_O = f + \sigma_O, \tag{19}$$

$$\partial_t D + \nabla\cdot \mathbf{J}_D = -f + \sigma_D. \tag{20}$$

Its general space–time balance is

$$\mathcal{I}[O(t_2) + D(t_2)] - \mathcal{I}[O(t_1) + D(t_1)] = -\rho_0 \int_{t_1}^{t_2} \int_{\partial\Omega} (\mathbf{J}_O + \mathbf{J}_D)\cdot \mathbf{n}\,\mathrm{d}S\,\mathrm{d}t + \rho_0 \int_{t_1}^{t_2} \int_\Omega (\sigma_O + \sigma_D)\,\mathrm{d}V\,\mathrm{d}t. \tag{21}$$

Internal conversion cancels in both formulations. Equation (21) is the physical rate-based balance, whereas Equation (18) describes the actual framewise corrections used here.

B.4 Pigment Feasibility and the Combined Certificate

Theorem B.1 (Carbon and pigment accounting certificate)**.** *Under the assumptions above, CALM has non-negative represented pools; preserves each source-free global inventory; preserves local total carbon under interconversion; satisfies Equation* (15)*; and, for every measurable $\mathcal{U} \subseteq \Omega$,*

$$0 \leq \rho_0 \int_{\mathcal{U}} 55\frac{\hat{y}}{A}\,\mathrm{d}V \leq \rho_0 \int_{\mathcal{U}} O^+\,\mathrm{d}V. \tag{22}$$

Here A is the fixed mass-conversion constant. Pigment carbon fits within the organic pool pointwise and hence in every region. BCR equals 100% wherever the fields are finite.

Proof. The carbon statements follow from Propositions B.1–B.3. Since $0 \leq 55g \leq 1$ and $O^+ \geq 0$,

$$0 \leq 55\hat{y}/A = 55O^+ g \leq O^+ \tag{23}$$

at every location. Integration over any region preserves the inequalities. The BCR indicator is therefore one at every evaluated wet voxel. □

B.5 Why Pigment Transport Does Not Certify Carbon Conservation

The distinction in Section 4.1 already appears in a motionless closed domain. Take spatially constant $O_0, D_0 > 0$, no internal transfer or external source, and any differentiable $g(t) \in (0, 1/55]$. Carbon is conserved, but

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega y \, \mathrm{d}V = A|\Omega| O_0 g'(t) \tag{24}$$

can be nonzero. Thus source-free carbon does not imply source-free pigment: a pigment source would be needed to describe acclimation even when there is no carbon source. In the reverse direction, the following construction conserves chlorophyll while changing represented total carbon. Together these examples explain why an unconstrained pigment source cannot stand in for carbon conversion and external carbon exchange.

B.6 Why Unconstrained Factorization Is Insufficient

Proposition B.4 (A product decoder supplies neither guarantee)**.** *The constraints $O > 0$, $g > 0$, and $\hat{y} = AOg$, even with non-negative D, do not imply pigment feasibility or source-free conservation of $\mathcal{I}[O + D]$.*

Proof. For feasibility, take any $O > 0$ and set $g = 2/55$. The product is a positive chlorophyll field, but $55\hat{y}/A = 2O > O$.

For conservation, consider spatially constant fields in a closed domain with no sources. Let $O(t) = c(t)O_0$, $g(t) = g_0/c(t)$, and $D(t) = D_0$, where $O_0, D_0, g_0 > 0$ and $c(t) > 0$ varies in time. Chlorophyll remains exactly constant:

$$\hat{y}(t) = A\, c(t) O_0 \frac{g_0}{c(t)} = AO_0 g_0. \tag{25}$$

Nevertheless,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{I}[O + D] = \rho_0 |\Omega| O_0 c'(t), \tag{26}$$

which need not vanish. Choosing $c(t) \geq 1$ and $g_0 \leq 1/55$ even makes this example pigment-feasible throughout. Thus accurate chlorophyll and a feasible ratio alone cannot certify carbon conservation. □

This is a statement about what the architecture guarantees, not a claim that every trained UNet3D violates either constraint on every dataset. An unconstrained model can satisfy the inequalities empirically, as the unbounded CALM ablations do.

C Implementation Details of CALM

C.1 TIDE-UNet3D and Positive Carbon Initialization

TIDE-UNet3D (Temporal Integration and Depth-preserving Encoding; Figure 5) implements the shared context network. A bias-free $1 \times 1 \times 1$ 3D convolution maps the 16 channels of each frame to 80 features. The encoder feature widths are $[80, 160, 368, 720]$. Only latitude and longitude are pooled, with stride $(1, 2, 2)$; all 12 depth levels and four frames are retained. At the bottleneck, the 720-channel representation receives learned frame embeddings and a coordinate embedding of depth and spherical latitude/longitude. Factorized attention then uses 16 heads per axis, with residual feed-forward blocks. Residual 3D convolutional blocks use GroupNorm and SiLU, with periodic longitude and replicated latitude/depth padding. Skip connections and trilinear interpolation restore the original horizontal resolution. The decoder yields 40 context channels per frame.

Each month is encoded with the same convolutional weights, giving all four frames a common spatial representation. Keeping every depth level avoids merging distinct layers during pooling. At the bottleneck, attention passes successively along depth, time, latitude, and longitude, combining vertical structure, information across months, and broader spatial context. The skip connections return finer spatial features to the shared decoder, complementing the context assembled at the bottleneck. These features condition the carbon and pigment modules; the subsequent constrained operations enforce carbon accounting.

The carbon initializer reads all four contexts and produces positive OC and DIC using exponentiated, clipped log-state coordinates based on training DIC statistics. Both pools are learned; the initializer does not copy dense DIC directly.

Table 6: CALM configuration used by the bounded three-seed experiments.

Setting	Value
Trainable parameters	68,010,226
Frames / channels per frame	4 / 16
Encoder widths	$[80, 160, 368, 720]$
Context / mechanism width	40 / 80
Context attention layers / heads	1 / 16
Context / analysis MLP expansion	2.00 / 2.00
GroupNorm groups / neural dropout	8 / 0
Greenness-analysis widths / heads	$[40, 80, 160]$ / 10
Boundary fractional limit η	0.250
Greenness log-residual limit	2.00
Optimizer	AdamW, weight decay 0.0500
Learning rate / minimum	5.00×10^{-4} / 1.00×10^{-6}
Schedule / maximum epochs	Cosine / 300
Effective training batch size	4 (one sample on each of four GPUs)
Gradient clipping / DIC-loss weight	1.00 / 0.0500
Main-experiment seeds	0, 1, 42
Selection	Lowest validation CHL NRMSE

C.2 Transport Coefficients and Solver

The coefficient network combines adjacent-frame environmental drivers, dense DIC, sparse chlorophyll, and observation masks. Separate non-negative, bounded outputs parameterize horizontal and vertical diffusivity for the organic and inorganic pools. These coefficients describe effective unresolved redistribution rather than molecular diffusion.

The solver backtracks trajectories using the resolved velocity, applies masked monotonic interpolation, and then performs horizontal and implicit vertical diffusion. Longitude is periodic, while land and closed boundaries prevent outward diffusive flux. A flux limiter stabilizes the explicit update. Non-negative clipping and a differentiable inventory correction preserve each pool's global inventory; a fallback handles degenerate provisional states. This correction does not make the advection scheme locally conservative. Carbon arithmetic and diagnostic accumulation retain higher precision than neural autocast, and the logger records pre-correction changes, limiter activity, and fallback events.

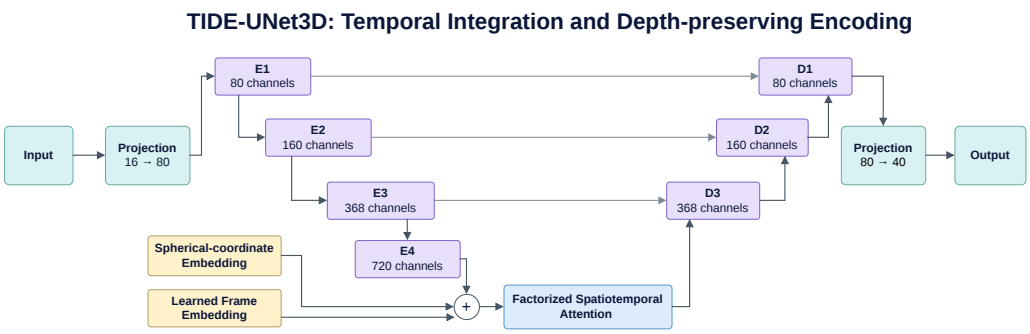


Figure 5: **TIDE-UNet3D.** Depth-preserving spatial encoding and axial temporal attention produce context for the carbon and pigment modules.

C.3 Sparse Ratio Analysis and the Bounded Decoder

At observed cells, the analysis network forms a log chlorophyll-to-carbon ratio using separate positive floors for pigment and carbon. Mask-normalized convolution accounts for available support, so missing observations are not interpreted as measured zero ratios. Context, sparse innovations, and carbon features condition a bounded adjustment around a training-derived reference. A constrained activation enforces the elemental pigment-carbon limit before converting the ratio and organic carbon into chlorophyll. The unbounded ablation removes this upper constraint while retaining the unit conversions.

Transport trajectories are computed before the analysis corrections; interconversion, boundary exchange, and pigment estimates do not feed back into subsequent transport. The model therefore performs contextual reconstruction rather than autonomous forecasting.

D EXPERIMENT DETAILS

D.1 DATASET, UNITS, AND OBSERVATION PROTOCOL

The dataset uses monthly MOM6 ocean output coupled to COBALTv2 biogeochemistry (Adcroft et al., 2019; Stock et al., 2020; Dunne et al., 2020). The common interval is January 1980–December 2022. Fields are regridded to a regular $1^\circ \times 1^\circ$ grid, with 180 latitude centers from -89.5° to 89.5° and 360 longitude centers from 0.5° to 359.5°. The 12 retained depth centers are

$$\{2.5, 10, 20, 32.5, 51.25, 75, 100, 125, 156.25, 200, 250, 312.5\}\ \mathrm{m}. \tag{27}$$

Every sample contains four consecutive monthly frames and the actual elapsed seconds between them. Windows advance by one month and never cross chronological split boundaries. The 1980–2010, 2011–2016, and 2017–2022 partitions contain 369, 69, and 69 windows, respectively. Reported test metrics include all 69 windows; overlapping occurrences of a month are evaluated in their respective window contexts rather than treated as independent temporal observations.

Table 7: Inputs to CALM. Surface-only fields are repeated over depth. There are 16 input channels per monthly frame and 64 per four-month window.

Group	Variables	Input representation
Nutrients	NO_3, PO_4, SiO_4	Dense 3D
Hydrography	Temperature, salinity	Dense 3D
Circulation	Ocean u, v, w	Dense 3D
Surface forcing	Wind u, v, shortwave radiation	Depth-repeated 2D
Upper-ocean state	Mixed-layer depth, sea-surface height	Depth-repeated 2D
Inorganic carbon	DIC	Dense 3D
Chlorophyll	CHL values, observation mask	Sparse 3D + binary mask

Wet domain and supervision. The solver domain is the intersection of the four DIC masks. Data preparation verifies that it covers all valid chlorophyll labels. Supervision and reconstruction metrics use the common CHL/DIC label domain. Conservation and greenness audits use the solver's wet domain. These domains should be distinguished: a domain used to test a physical constraint need not be identical to the set of labeled voxels. Missing velocities do not shrink the wet tracer domain.

Sampling. For each calendar month and depth, valid chlorophyll voxels receive deterministic 64-bit hash scores based on seed, month, depth, and spatial index. Selecting the lowest $\mathrm{round}(rN)$ scores gives a depth-stratified fraction r. Masks remain fixed across epochs, methods, and repeated occurrences of the same calendar month. The nominal fractions 0.1%, 1%, and 10% correspond to effective test fractions 0.1001074%, 0.9999691%, and 10.0001326%. The test set contains 124,893,864 valid voxel instances; the lowest density retains 125,028. Sampling signatures are checked when collecting results.

Units. CHL is expressed as chlorophyll a equivalent in $\mu\mathrm{g\,kg^{-1}}$, carbon pools in mol C $\mathrm{kg^{-1}}$, and greenness in mol Chl per mol C. Chlorophyll a has molar mass $893.49\ \mathrm{g\,mol^{-1}}$ and 55 carbon atoms per molecule. Hence $A = 893.49 \times 10^6\ \mu\mathrm{g\,mol^{-1}}$, $\hat{y} = AOg$, and the pigment-carbon fraction is $55g$. The carbon-to-pigment inequality assumes that the represented organic pool contains pigment carbon.

Standardization. Mean and standard deviation are estimated using training data only, separately by variable, within-window frame, and depth. CALM uses the saved standard deviations exactly:

$$z_{t,d} = (x_{t,d} - \mu_{t,d})/\sigma_{t,d}, \qquad x_{t,d} = \mu_{t,d} + \sigma_{t,d} z_{t,d}. \tag{28}$$

Invalid inputs are zero-filled after normalization and accompanied by masks. Grid-cell volumes are derived from spherical-shell geometry; constant factors in the volume weights cancel in normalized losses and RMSE.

D.2 Parameters and Computational Efficiency

Table 8 compares model size, computational costs, and reconstruction error. NRMSE and RMSE are three-seed means ± sample SD at 0.1% sampling; RMSE is in $10^{-2}\ \mu\mathrm{g\,kg^{-1}}$. Kriging is excluded from this GPU table because its implementation is CPU-only, has no neural parameters, and has no optimizer-based training stage.

Table 8: **Parameter, efficiency, and performance comparison of different models.** CALM uses four training GPUs; $(\times 4)$ marks its training time and per-device memory.

Method	Params	Training		Inference		CHL error	
		s/epoch	GiB	s/test set	GiB	NRMSE	RMSE
4DVarNet	2.06M	274	13.3	1.99	3.51	$0.275_{\pm 0.00142}$	$5.05_{\pm 0.0348}$
Marble	3.02M	256	19.8	1.23	3.77	$0.630_{\pm 0.00404}$	$10.3_{\pm 0.121}$
FNO	22.6M	256	18.6	2.00	6.37	$0.268_{\pm 0.000861}$	$4.42_{\pm 0.0245}$
UNet3D	71.7M	306	36.5	15.2	13.5	$0.199_{\pm 0.00408}$	$3.08_{\pm 0.0410}$
UNet3D (Factorized)	71.7M	313	36.5	15.2	13.5	$0.204_{\pm 0.00147}$	$3.21_{\pm 0.0354}$
ViViT	144M	323	6.13	0.360	3.49	$0.428_{\pm 0.000485}$	$5.64_{\pm 0.0154}$
AtmoRep	525M	249	10.9	1.57	5.11	$0.356_{\pm 0.000982}$	$4.82_{\pm 0.0395}$
Prithvi WxC	2.37B	696	27.0	39.6	15.3	$0.233_{\pm 0.00189}$	$3.62_{\pm 0.00818}$
CALM	68.0M	304 $(\times 4)$	48.0 $(\times 4)$	48.1	54.9	$0.186_{\pm 0.00100}$	$2.99_{\pm 0.0483}$

Training time. Training time is the mean wall-clock duration per epoch, averaged first within each run and then equally across nine runs covering three sampling rates and three seeds. We exclude the initial validation-only entry (epoch zero). Epoch durations include training and validation and exclude the subsequent checkpoint write. CALM uses four GPUs with one sample per GPU; the neural baselines use one GPU with batch size four. The $(\times 4)$ annotations indicate four devices: the displayed time is wall-clock seconds per epoch, and multiplying this time by four gives GPU-seconds.

Training memory. Training memory is reported as peak allocated GPU memory, including forward and backward computations and optimizer state, under each model's training batch size and precision. For CALM, we report the maximum per-device peak across four DDP ranks.

Inference measurement. All models are measured sequentially on the same RTX 6000D using their original project runners and compatible virtual environments. CALM, both UNet3D models, and ViViT use the `6000` environment; 4DVarNet, Marble, FNO, AtmoRep, and Prithvi use their respective environments. FNO uses FP32; the other neural models use BF16 autocast, with CALM's physical computations retained in FP32. We use batch size four and perform three warm-up forwards before measuring all 69 test windows in a fixed order: 17 batches of four and a final batch of one, with no dropped or padded windows. Each forward is bracketed by CUDA synchronization and a wall-clock timer; the table reports the sum of those measured intervals. CUDA-event measurements are also retained. Model loading, disk reads, batch preparation, and error aggregation are excluded; native model computations, including CALM's internal diagnostics and 4DVarNet's iterative solver, are included. Peak allocated memory includes model weights, prepared inputs, intermediate tensors, and outputs; peak reserved memory is retained separately in the measurement artifacts. Inference costs use the 0.1%, seed-42 checkpoint of each model. The NRMSE and RMSE columns report the corresponding 0.1% three-seed means and sample standard deviations ($n = 3$, $\mathrm{ddof} = 1$) from Table 1. The 4DVarNet solver normalizes its variational gradient across the batch, so the fixed evaluation order and batch composition are part of its inference protocol.

D.3 Baseline Adaptation and Training Protocols

Kriging uses three-dimensional local ordinary interpolation separately for each month, with exponential or spherical variograms. Horizontal positions are embedded on the sphere and depth enters the distance metric. It has no neural training stage and runs on CPU. Neural baselines use 56 dense channels (14 per month), four sparse chlorophyll channels, and four masks. Model-specific embeddings may reorganize these inputs. Table 1 reports total adapted parameter counts, verified against evaluation metadata and training logs, including frozen weights. Prithvi WxC has 2,368,350,912 total

parameters: 87,024,832 trainable and 2,281,326,080 frozen. The two UNet3D variants (Çiçek et al., 2016) have 71,667,012 and 71,667,916 parameters, respectively; both round to 71.7M in the table but are ordered using their exact totals. Kriging has no neural model parameters and is shown with a dash.

UNet3D (Factorized) uses widths $[112, 224, 448, 896]$, 64 inputs, and 12 outputs, totaling 71,667,916 trainable parameters. Four outputs represent positive OC, four represent positive greenness, and four represent normalized DIC. Its reconstruction is $\hat{y} = OG$. The public G is a mass chlorophyll-to-carbon ratio; it is divided by A to obtain the molar coefficient g. The audit checks DIC predictions for non-negativity after denormalization from the direct unconstrained head. The training configuration uses batch size 4, learning rate 0.005, 300 epochs, and DIC loss weight 0.05. Architectural and optimizer differences mean this is a mechanistic baseline comparison rather than a single-component controlled ablation.

D.4 Evaluation and Existing Baseline Conventions

Table 1 counts trainable and frozen parameters, uses three significant digits, underlines the strongest baseline in each column, and shades CALM. Promotion is the relative reduction from the lowest baseline mean, computed before rounding. Table 2 marks learned diffusion, interconversion, and boundary sources by K, q, and s; the greenness bound remains enabled in every row.

CALM accumulates weighted squared errors and valid weight sums over the complete test set before taking square roots; distributed evaluation does not pad the 69 windows. Means and extrema in the diagnostic logger use their stated reductions rather than averaging maxima across batches. Three-seed tables report sample standard deviation ($n - 1$ denominator), not standard error. The seeds jointly vary masks and training randomness, so their variability should not be attributed solely to one source.

D.5 Hyperparameter Setting

Table 9 lists the hyperparameters of every baseline in the main comparison. They are taken from the archived experiment configurations and checked against the adaptation code, including inherited pretrained settings. For each method, the architecture and core training settings agree across all nine sampling-rate/seed runs. CALM's configuration is given in Table 6 and Appendix C.

Shared inputs and optimization. Each neural method receives four months on a $12 \times 180 \times 360$ depth–latitude–longitude grid: 56 dense channels, four sparse chlorophyll channels, and four observation masks. ViViT uses 16 channels per frame; Prithvi folds depth into its input channels before the horizontal patch embedding. Direct-output models reconstruct four chlorophyll volumes. UNet3D (Factorized) predicts four volumes each of OC, greenness, and DIC, with DIC-loss weight 0.05 and no greenness upper bound.

All neural models use AdamW (Loshchilov & Hutter, 2017), $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, gradient-norm clipping at 1, and an effective batch size of four. The baseline batch is four on one GPU; CALM uses one sample per device on four GPUs. Checkpoints minimize validation CHL NRMSE. The configured budget is 300 epochs except Prithvi's 100. AtmoRep uses a fixed learning rate. Other methods use cosine decay to 10^{-6}; ViViT first applies ten linear-warmup epochs from 0.01 times its peak rate, followed by 290 cosine epochs. Neural autocast uses BF16 except for FNO, which uses FP32. CALM's transport, carbon arithmetic, and losses remain FP32.

Table 9: Hyper-parameters used for all baseline models.

Baseline	Hyper-param.	Meaning	Value
Kriging (Exp., Sph.)	Covariance	Two separately evaluated covariance kernels	Exponential / spherical
	Neighbors	Nearest observations per local solve	8
	Range	Practical range in the rescaled spatial metric	8.00
	Depth scaling	Metres per degree-equivalent depth coordinate	25.0
	Sill	Observation variance per frame, floored at 10^{-6}	Automatic
	Nugget	Additive diagonal covariance regularization	1.00×10^{-6}
	Jitter	Sill-relative diagonal solver stabilization	1.00×10^{-8}

Continued on next page

Table 9: Hyper-parameters used for all baseline models (continued).

Baseline	Hyper-param.	Meaning	Value
4DVarNet	Iterations	Learned gradient iterations	3
	Initializer	Hidden width / convolutional layers	32 / 3
	Prior	Hidden width / residual prior blocks	96 / 2
	Prior kernel / dropout	Prior convolution kernel size / dropout	3 / 0
	Gradient update	Recurrent hidden width / kernel size	128 / 3
	Objective weights	Observation / prior weights inside the solver	1.00 / 1.00
	Update scale / clip	Learned-update scale / gradient value clip	1.00 / 10.0
	Learning rate	Initial AdamW learning rate	1.00×10^{-4}
	Weight decay	AdamW decoupled weight decay	1.00×10^{-4}
Marble (GridMix)	Encoder widths	3D convolutional encoder channels	[16, 32, 64, 128]
	Latent width	Global conditioning code dimension	128
	Implicit decoder	SIREN hidden width / number of layers	48 / 4
	Frequency	Initial and hidden sinusoidal frequency	30.0
	Hypernetwork	Hidden width / number of layers	128 / 2
	Modulation grid	Horizontal grid resolution / number of bases	24 × 48 / 16
	Modulation	Shift enabled; scale disabled; bases summed	Yes / No / Yes
	Learning rate	Initial AdamW learning rate	0.00100
	Weight decay	AdamW decoupled weight decay	0.00500
FNO	Fourier stack	Hidden width / Fourier layers	48 / 4
	Fourier modes	Configured modes in depth, latitude, longitude	[6, 24, 32]
	Channel ratios	Lifting / projection / channel-MLP expansion	2.00 / 2.00 / 0.500
	Precision	Fourier-block arithmetic	FP32
	Learning rate	Initial AdamW learning rate	0.00500
	Weight decay	AdamW decoupled weight decay	1.00×10^{-4}
UNet3D	Feature widths	Encoder feature widths	[112, 224, 448, 896]
	Pooling	Stride in depth, latitude, longitude	[1, 2, 2]
	Normalization	GroupNorm groups / dropout	1 / 0
	Outputs	Direct chlorophyll output channels	4
	Learning rate	Initial AdamW learning rate	0.00500
	Weight decay	AdamW decoupled weight decay	0.0500
UNet3D (Factorized)	Feature widths	Encoder feature widths	[112, 224, 448, 896]
	Pooling / groups	Same horizontal stride and GroupNorm as UNet3D	[1, 2, 2] / 1
	Outputs	Four channels each for OC, greenness, and DIC	12
	DIC loss	Auxiliary DIC supervision weight	0.0500
	Learning rate	Initial AdamW learning rate	0.00500
	Weight decay	AdamW decoupled weight decay	0.0500
ViViT	Patch size	Depth, latitude, longitude per monthly frame	[12, 18, 18]
	Embedding	Transformer embedding width	768
	Layers / heads	Spatial layers / temporal layers / heads per layer	8 / 4 / 12
	MLP / dropout	Feed-forward expansion / dropout probability	4.00 / 0.100
	Decoder widths	Initial width followed by convolutional widths	[256, 128, 64, 32]
	Warmup	Linear warmup epochs / initial LR multiplier	10 / 0.0100
	Learning rate	Initial AdamW learning rate	5.00×10^{-4}
	Weight decay	AdamW decoupled weight decay	0.0500
AtmoRep (adapted)	Patch size	Depth, latitude, longitude after stacking months	[3, 15, 30]
	Embedding	Transformer embedding width	2048
	Layers / heads	Adapted Transformer blocks / attention heads	10 / 16
	MLP / dropout	Feed-forward expansion / residual dropout	2.00 / 0.0500
	Attention dropout	Dropout on attention weights	0
	Schedule	Fixed learning rate for all configured epochs	Constant / 300
	Learning rate	Initial AdamW learning rate	5.00×10^{-4}
	Weight decay	AdamW decoupled weight decay	0.00200
Prithvi WxC	Patch / window	Horizontal patch / local attention window	2 × 2 / 30 × 30
	Embedding	Pretrained transformer embedding width	2560
	Encoder / decoder	Actual transformer layers: $2n + 1$ for each stack	25 / 5
	Heads / MLP	Attention heads / feed-forward expansion	16 / 4
	Dropout / drop path	Transformer dropout / stochastic depth	0 / 0
	Trainable subset	Input adapter, output head, last decoder block	Yes / Yes / Yes
	Epochs	Maximum configured training epochs	100
	Learning rate	Initial AdamW learning rate	3.00×10^{-4}
	Weight decay	AdamW decoupled weight decay	5.00×10^{-4}

Interpolation and pretrained adaptations. The two kriging baselines (Oliver & Webster, 1990) share their neighborhood and scaling parameters and differ only in the exponential or spherical covariance. Their distance metric combines a spherical chord embedding of longitude/latitude (radius $180/\pi$) with depth divided by 25 m, so the range is expressed in degree-equivalent coordinates. They use sparse chlorophyll only and require no optimizer.

The AtmoRep adaptation loads compatible attention and feed-forward tensors from ten pretrained encoder blocks into a single task-specific Transformer stack. Patch embeddings, positions, layer normalizations, and the reconstruction head are initialized for this task; all parameters are trainable. A single reconstruction head replaces the original decoder and ensemble heads; per-head query/key normalization and post-feed-forward GELU are omitted. Prithvi retains its pretrained backbone, overrides the local window to 30×30, and trains only the input adapter, output head, and final decoder Transformer. The backbone configuration specifies 12 encoder and two decoder local–global pairs; the implementation adds one local block to each stack, giving 25 and five actual Transformer layers. Depth is represented through channels, rather than through Prithvi's horizontal patch dimensions.

D.6 Component, Backbone, and Input Ablation Protocols

The six component/objective ablations in Table 2 share seed 42, 0.1% sampling, TIDE-UNet3D, and the 300-epoch schedule. They set learned diffusion to zero, remove carbon conversion, remove boundary corrections, disable DIC supervision, or set its weight to 0.01 or 0.10. The full bounded reference uses $\lambda_D = 0.05$. The greenness-bound comparison is separate from these six component runs.

For the backbone comparison, the complete model parameter counts are 68,010,226 (TIDE-UNet3D), 67,842,570 (UNet3D), 68,081,618 (ViViT), and 67,957,514 (3D CNN). All modules outside the backbone account for the same 4,101,674 parameters. The alternative UNet3D uses widths $[80, 160, 368, 976]$ and residual convolutions with skip connections. The 3D CNN uses widths $[80, 160, 368, 1040]$ and interpolating convolutional decoding without encoder–decoder skips. Both concatenate the four months along the input channel axis and retain depth through horizontal pooling. The replacement ViViT uses patches $(3, 15, 15)$, width 544, eight spatial and four temporal transformer layers, and eight heads. Each backbone maps 64 input channels to four 40-channel context volumes, while all carbon modules, objectives, and training settings remain fixed.

Input ablations retain every model parameter and input dimension. For each selected variable, all four monthly channels are zeroed after normalization at the TIDE-UNet3D input and after physical scaling at the diffusivity-network input. Thus neither neural pathway receives the removed variable. Ocean velocities u, v, w, dense DIC, sparse CHL, observation masks, wet-domain masks, and supervision remain available. Wind u, v are distinct from the ocean velocities driving transport. The ten configurations remove NO_3, PO_4, SiO_4, temperature, salinity, wind u, wind v, shortwave radiation, mixed-layer depth, or sea-surface height. For each configuration, the checkpoint with the lowest validation CHL NRMSE is evaluated on the test set. Table 3 reports NRMSE only, to compare the design choices on a common metric.

D.7 Paired Field-Level Audit and Overlap Metrics

The paired audit uses the bounded CALM and UNet3D (Factorized) best checkpoints at sampling rate 0.001 and seed 42, with BF16 neural inference and FP32 CALM physical operations. Carbon inventories and audit statistics are accumulated in FP64. The two original samplers are checked for identical observations, labels, and wet domains in each test window.

The 69 four-month windows span 72 calendar months. Only the 66 interior months have four reconstructions, one at each within-window position; the first and last three months are excluded from the main overlap and stitched-series diagnostics. Inventory comparisons use the fixed intersection of all paired wet domains. BCR retains the per-window wet-voxel definition below. Chlorophyll overlap metrics exclude observed voxels; carbon and greenness metrics use the common wet domain.

For monthly field x with four contextual reconstructions $x^{(k)}$, let $\bar{x} = \frac{1}{4}\sum_k x^{(k)}$, $\sigma_x^2 = \frac{1}{4}\sum_k (x^{(k)} - \bar{x})^2$, and $\langle x \rangle_V = \sum_i V_i M_i x_i / \sum_i V_i M_i$. Relative spread is $\langle \sigma_x \rangle_V / \langle |\bar{x}| \rangle_V$, averaged equally over the 66 months. Inventory CV instead computes the population standard deviation of the four inventories divided by their mean magnitude.

The initial-frame unaccounted variation compares $(I^+_{n,t} - B_{n,t})$ to $(I^+_{n,1} - B_{n,1})$, takes absolute relative differences for $t = 2, 3, 4$, and averages over windows and positions. DIC inventory-change error compares each predicted change from the first frame with the corresponding true DIC change, normalized by the first-frame true inventory. Stitched variation first averages the four inventories for each interior month and then measures absolute relative changes from the first such month. Its source-adjusted version subtracts the averaged declared source corrections. It includes independent-window initialization differences and cannot be interpreted as source-free long-range rollout drift. The paired analysis uses seed 42, with overlapping windows drawn from the same test time series.

D.8 What Can Be Compared Across Architectures

Table 4 reports the matched seed-42 field-level audit. BCR and maximum pigment-carbon fraction are directly comparable after converting both decoders to molar greenness. CALM's internal transport and conversion residuals are defined for its explicit transport and conversion operations; the paired comparison reports diagnostics shared by both architectures.

For a shared accounting diagnostic, define $I^+_{n,t} = \mathcal{I}[O^+_{n,t} + D^+_{n,t}]$ and $B_{n,t} = \mathcal{I}[s^O_{n,t} + s^D_{n,t}]$. Let $B = 0$ for a model that declares no source correction. Define

$$E_{\text{unacc}} = \text{mean}_{n,t<T} \frac{|(I^+_{n,t+1} - I^+_{n,t}) - (B_{n,t+1} - B_{n,t})|}{\max(|I^+_{n,1} - B_{n,1}|, \epsilon_I)}. \tag{29}$$

For CALM this is a source-adjusted total-carbon trajectory diagnostic. For UNet3D it measures changes in predicted total carbon not attributed to an explicit source model. We report this quantity as unaccounted inventory variation. A separate raw-inventory variation metric omits the B terms and includes changes due to external exchange.

The paired field-level audit uses the same wet masks, volumes, windows, and density convention for both models; checks predictions for non-finite values and negative carbon; converts UNet3D's G to $g = G/A$; and accumulates per-window inventories before computing absolute relative changes. Absolute relative changes are computed within each window before averaging, avoiding cancellation between signed changes. CALM's transport drift averages separate poolwise transition errors and is reported separately from Equation (29).

We separate global transport accounting, local interconversion, and final-state source accounting. With $I^C_{n,t} = \mathcal{I}[C^{\text{tr}}_{n,t}]$ for window n, transport drift is

$$E_{\text{tr}} = \text{mean}_{n,t,C} \frac{|I^C_{n,t+1} - I^C_{n,t}|}{\max(|I^C_{n,t}|, \epsilon_I)}, \qquad C \in \{O, D\}. \tag{30}$$

It measures each pool separately, which is stronger than checking only their sum. Local conversion error and source-accounted total-carbon error are

$$R_{\text{cv}} = \max_{n,t,\mathbf{x}} |O^{\text{cv}} + D^{\text{cv}} - O^{\text{tr}} - D^{\text{tr}}|, \tag{31}$$

$$E_{\text{bud}} = \max_{n,t} \frac{|\mathcal{I}[O^+ + D^+] - \mathcal{I}[O^{\text{tr}} + D^{\text{tr}}] - \mathcal{I}[s^O + s^D]|}{\max(|\mathcal{I}[O^{\text{tr}} + D^{\text{tr}}]|, \epsilon_I)}. \tag{32}$$

For greenness, the biochemical consistency rate is the fraction of valid wet voxels satisfying $0 \le 55g \le 1$:

$$\text{BCR} = 100 \ \text{mean}_n \frac{\sum_{t,i} M^\Omega_{n,i} \, \mathbf{1}\{0 \le 55g_{n,t,i} \le 1\}}{\sum_{t,i} M^\Omega_{n,i}} \ \%. \tag{33}$$

This is the sample-averaged voxel fraction used by the existing diagnostic logger, not a volume-weighted accuracy metric. Greenness is non-negative by construction, so the stored upper-bound violation fraction determines BCR exactly. We additionally report the maximum pigment-carbon fraction $f_{\max} = \max 55g$.

Table 10: **CALM carbon and biological audits.** Transport drift is averaged over three seeds; the other columns report the worst value across seeds. R_{cv} is in mol C kg^{-1}; both inventory errors and $f_{\max}$ are dimensionless. BCR is in percent.

Sampling	$E_{\mathrm{tr}} \downarrow$	$R_{\mathrm{cv}} \downarrow$	$E_{\mathrm{bud}} \downarrow$	BCR $\uparrow$	$f_{\max}$
0.1%	3.20×10^{-8}	2.38×10^{-7}	1.93×10^{-7}	100	3.83×10^{-5}
1%	2.92×10^{-8}	2.38×10^{-7}	2.08×10^{-7}	100	4.01×10^{-5}
10%	3.02×10^{-8}	2.38×10^{-7}	1.94×10^{-7}	100	4.13×10^{-5}

E Additional Experimental Results

E.1 Predictive Uncertainty: Protocol and Metrics

We freeze the selected CALM, UNet3D, and UNet3D (Factorized) reconstructions and fit a separate 6,913-parameter scale head around each predictive mean. A positive output defines a marginal Gaussian distribution; the reconstruction weights and their deterministic carbon accounting remain unchanged. Features include environmental and DIC fields, the frozen prediction, available sparse chlorophyll and its mask, local observation statistics, observation distance, and coordinates. Unobserved targets are not input features. Feature normalization and the fixed-depth residual-scale control use training data only. The head minimizes Gaussian negative log-likelihood, with validation-only checkpoint selection and no test-set rescaling.

The splits, sampling rates, and seeds match the simulation benchmark. Primary scores use unobserved valid chlorophyll and cell-volume weights. Negative log-likelihood (NLL) scores the assigned target density in standardized coordinates; continuous ranked probability score (CRPS) integrates the squared difference between the predictive cumulative distribution and the target step function, in physical chlorophyll units. Lower values are better for both. Prediction interval coverage is the weighted fraction of targets inside an interval; it should approach its nominal level rather than simply increase. Mean interval width measures sharpness and must be interpreted with coverage. The area under the sparsification error curve (AUSE) compares error after retaining low-uncertainty locations with an oracle ordering by actual error; lower values indicate better error ranking. Its normalized form uses a shared stratified sample with inverse-inclusion volume weights. Three-seed error bars show sample standard deviation, not confidence intervals.

CALM has the lowest CRPS, standardized NLL, and normalized AUSE in all nine matched rate/seed settings. Its nominal 95% intervals attain 92.10%, 92.36%, and 94.13% coverage at 0.1%, 1%, and 10% sampling. The learned head improves CRPS over the fixed-depth control in all 27 model/rate/seed experiments. Residual undercoverage remains. The marginal Gaussian distribution neither separates aleatoric and epistemic uncertainty nor defines jointly conservative carbon-state samples; intervals can extend below zero.

Local trajectory selection. The site is the valid grid center nearest a fixed geographic target in the validation data: 30.5°N, 150.5°W. We display January 2018–December 2019 at 2.5, 75, and 156.25 m, using 0.1% sampling and seed 42. Each month uses the second frame of its reconstruction window, avoiding repeated contextual estimates of a month. All displayed targets are withheld from the sparse inputs. Dark and light bands denote $\mu \pm \sigma$ and $\mu \pm 2\sigma$, with nominal Gaussian coverage approximately 68.3% and 95.4%. The location was selected without inspecting local test errors or coverage. At the displayed 75 m site, the two-standard-deviation bands cover 19, 22, and 23 of 24 months for CALM, UNet3D, and UNet3D (Factorized), respectively. This illustrative local coverage does not replace the full-test scores.

E.2 Argo-Derived Product Adaptation

The observation source is a gridded Argo-derived chlorophyll product, rather than untreated individual float profiles. Extremely sparse valid source values are mapped vertically and converted from mg m^{-3} to μg kg^{-1} using the fixed seawater-density convention. To support stable fine-tuning, local ordinary kriging expands each month's valid support separately at each depth and within connected wet regions. It uses up to eight original neighbors, an exponential covariance with an 800 km scale, and interpolation in log-transformed chlorophyll space. Values are constrained to the neighboring value range; no values are borrowed from another month, depth, or split.

Train/validation/test years are 1997–2018, 2019–2020, and 2021–2022, with 261, 21, and 21 four-month windows. The target support expansion is fifty-fold where wet-domain capacity permits. In the test set, 98,903 original gridded values become 4,919,811 valid values, approximately 98% of which are interpolated (counts deduplicated by month). Original and interpolated labels have equal status in input sampling and the volume-weighted objective. Models initialize from their simulation checkpoints and use 10% input sampling with seed 42; validation physical RMSE selects the adapted checkpoint. A physical-unit bridge preserves pretrained model coordinates while using the Argo

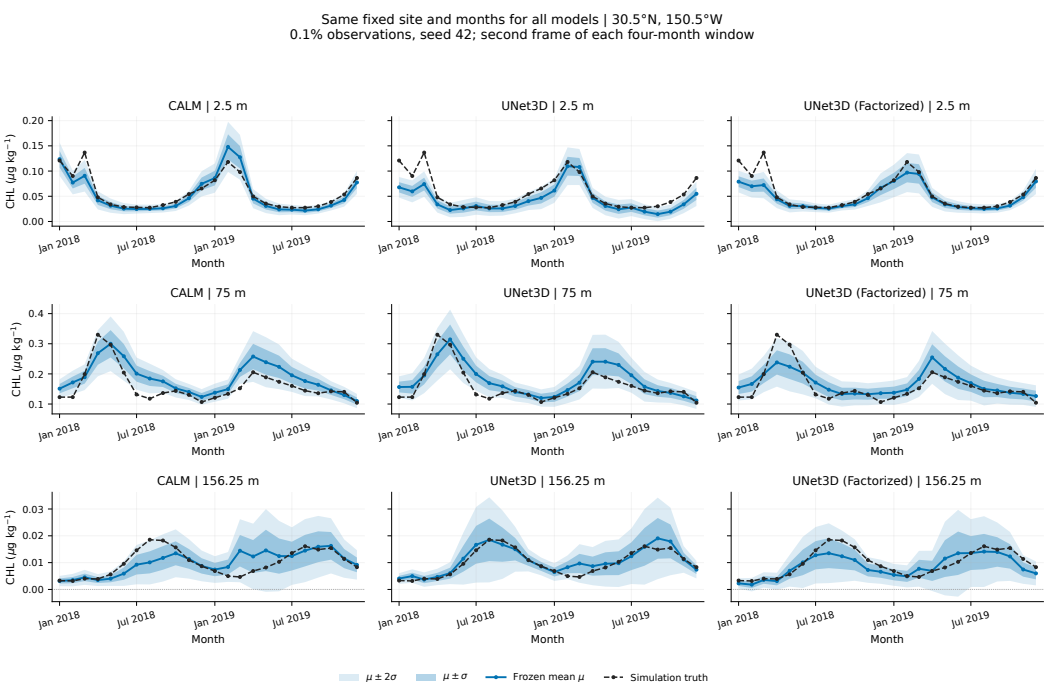


Figure 6: **Local uncertainty across models and depths.** Frozen predictions, simulation truth, and one-/two-standard-deviation bands at the same site and months.

target statistics. Observation-derived and simulated chlorophyll have different means and variances, so they do not share normalization statistics.

Table 5 reports volume-weighted NRMSE and RMSE on unobserved valid positions. NRMSE uses training-only channel/depth statistics and is not directly comparable between the simulation and Argo datasets. The 21 test windows contain 17,163,844 valid label occurrences, including 1,716,318 inputs; repeated months are counted in their respective window contexts. Both original and interpolated points enter the held-out score.

Scope of the observation experiment. Validation and test fields are interpolated using all original points in their own months before drawing the input subset. The scores therefore evaluate reconstruction of the interpolated product; Appendix F discusses the scope of this evaluation.

E.3 Qualitative Error Maps at All Depths

Figures 7–18 show all 11 methods at the 12 modeled depths. They use October 2019, the first frame of the October 2019–January 2020 test window, with 10% observations and seed 42. The same valid CHL domain is used for all methods at each depth. Errors are prediction minus ground truth in μg kg^{-1}; red indicates underestimation, blue overestimation, and gray invalid cells. RMSE annotations use the original cell-volume weights.

Selection and display protocol. The existing visualization case was selected from 276 window/frame slices by first ranking the number of depths where CALM improves over UNet3D, then the relative reduction in full-volume RMSE, with chronological tie-breaking. It is a deliberately favorable case, not a random or representative estimate of average performance. The main figure extracts 10, 100, and 200 m from this same case. Each appendix depth uses a symmetric color scale shared across all methods, based on the pooled 99th percentile of absolute error; extended colorbars indicate saturation beyond those limits. The main figure instead pools only UNet3D and CALM for its depth-specific limits. Maps are regenerated from the frozen predictions without interpolation or spatial smoothing.

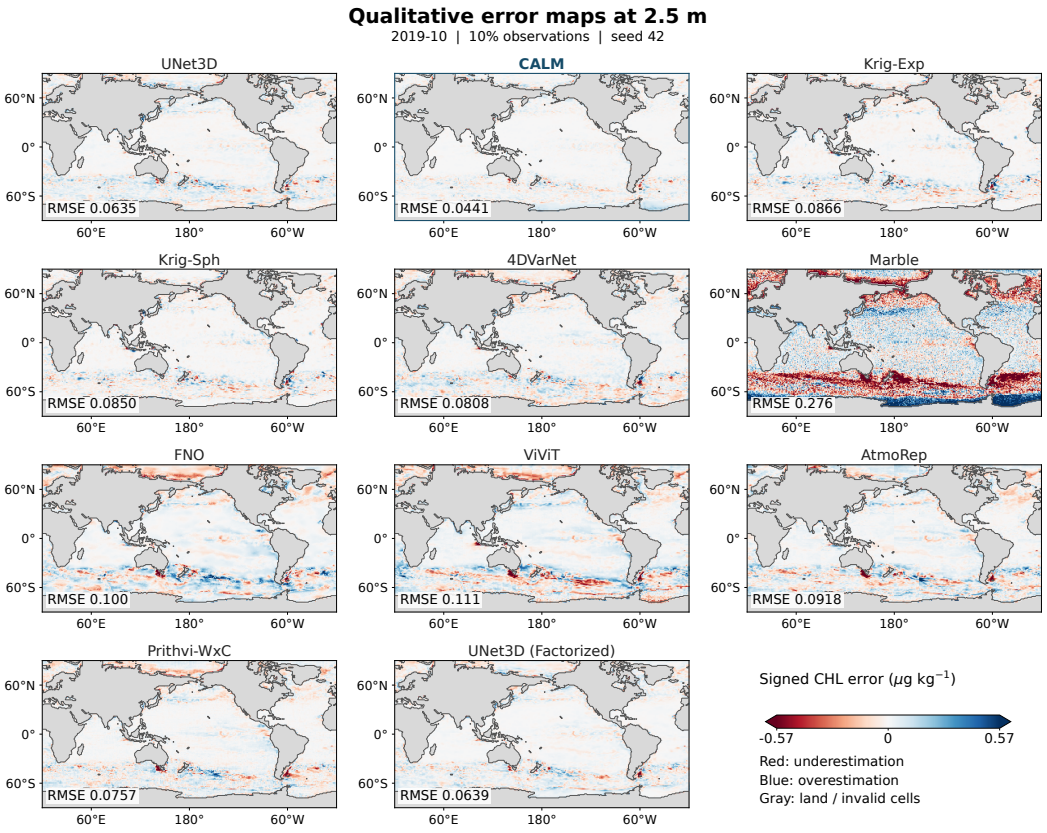


Figure 7: **Qualitative error maps at 2.5 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

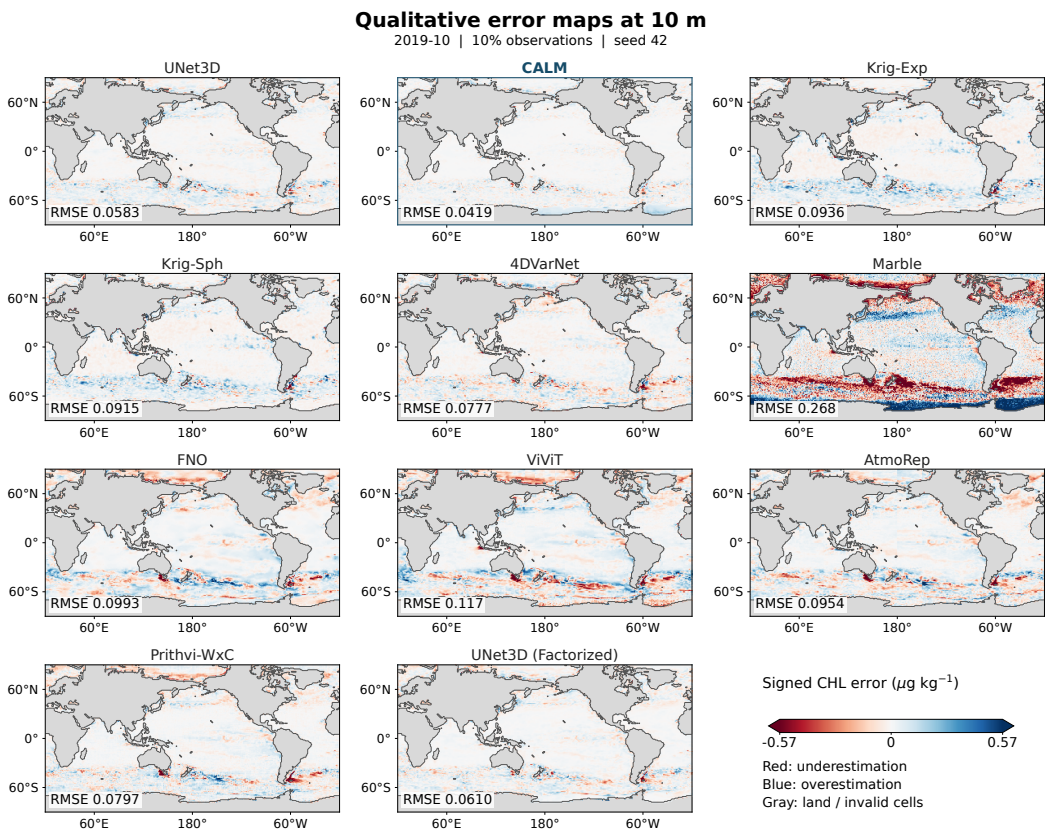


Figure 8: **Qualitative error maps at 10 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

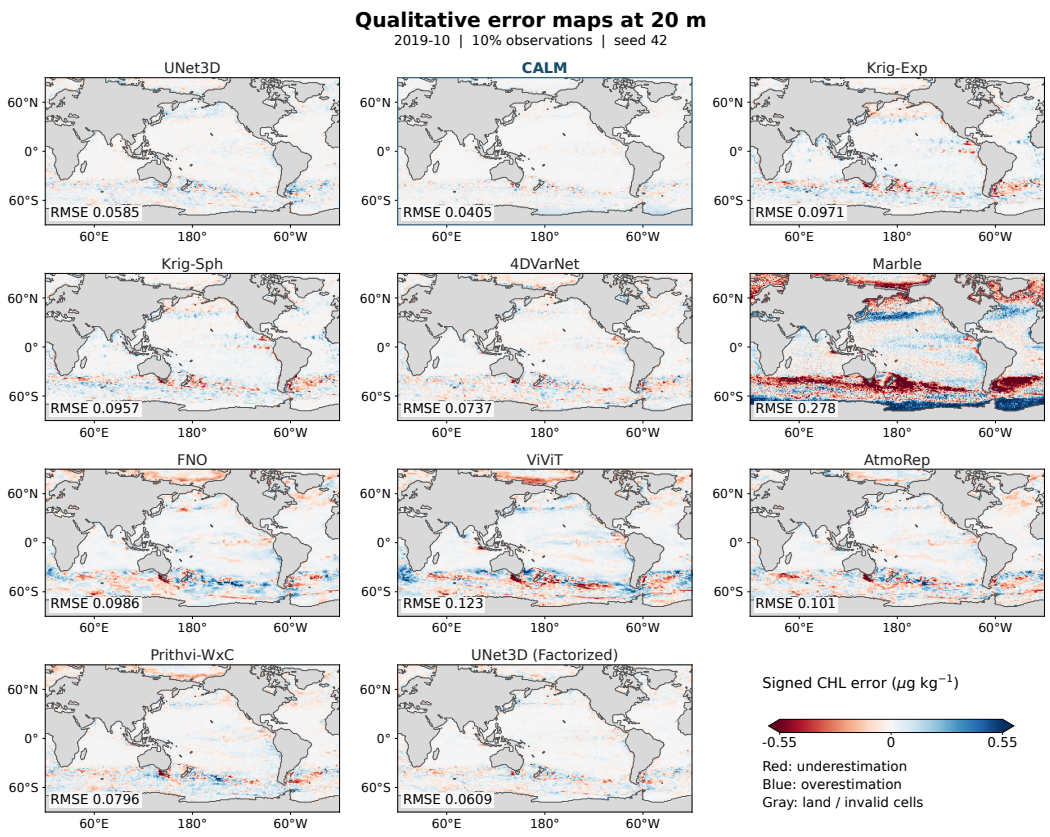


Figure 9: **Qualitative error maps at 20 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

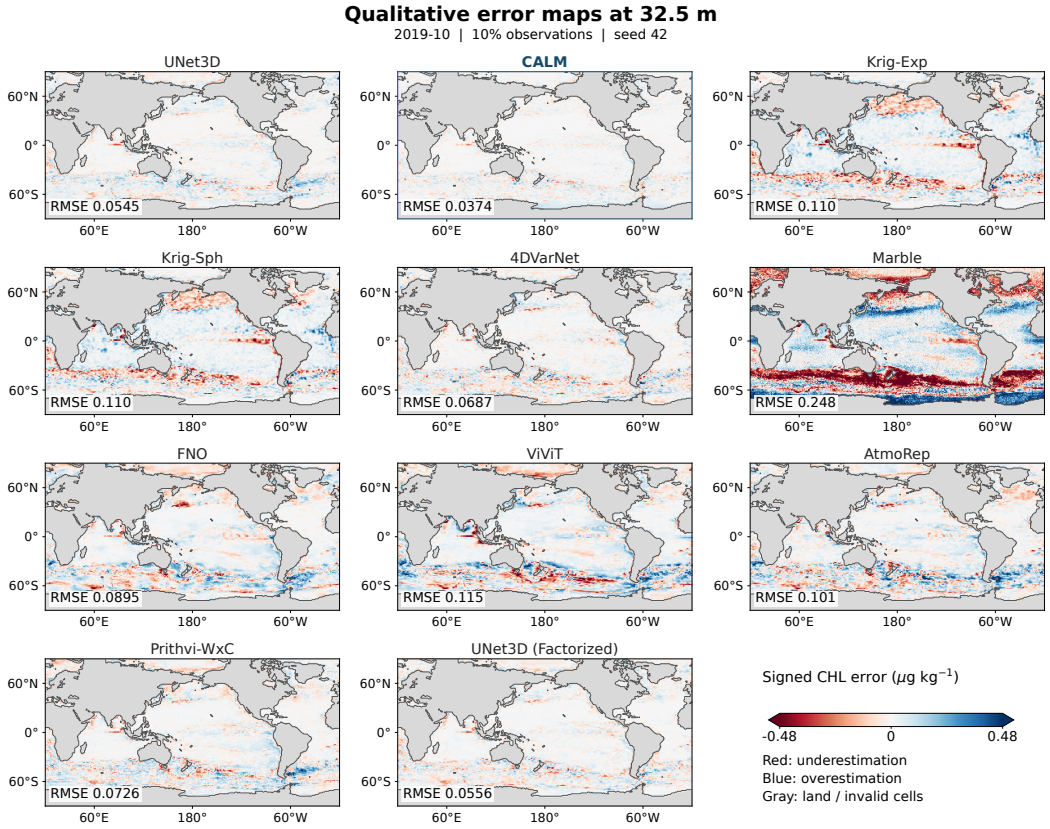


Figure 10: **Qualitative error maps at 32.5 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

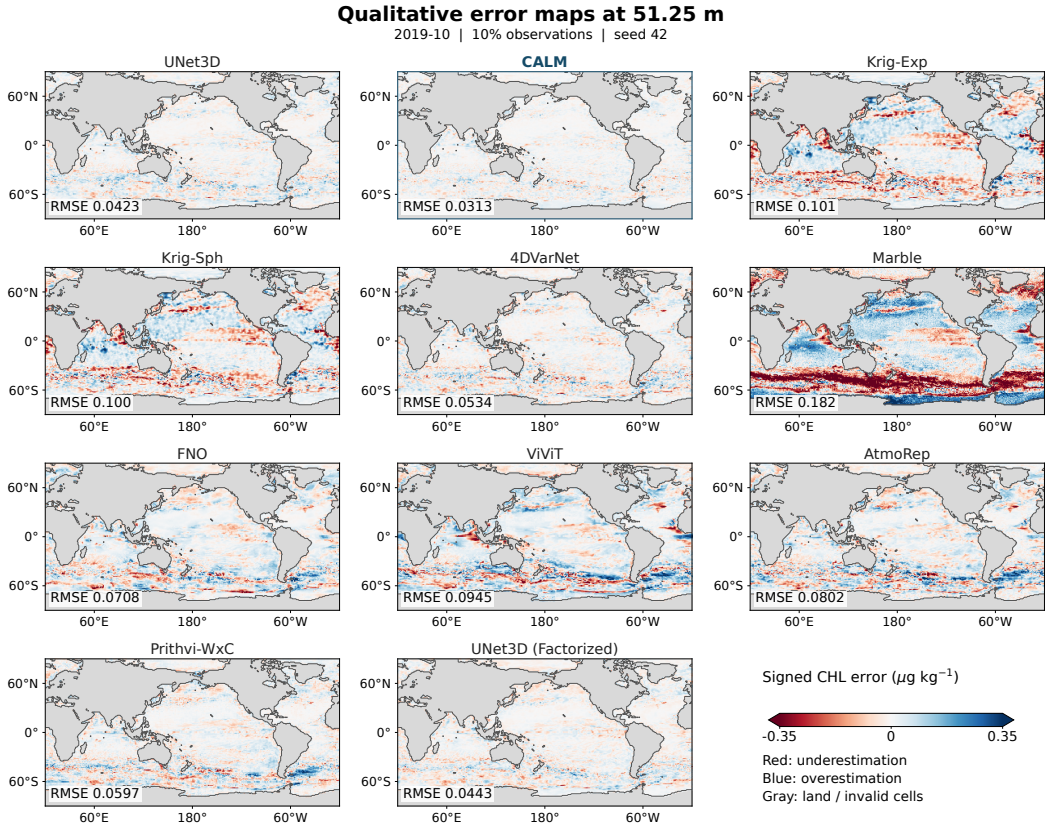


Figure 11: **Qualitative error maps at 51.25 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

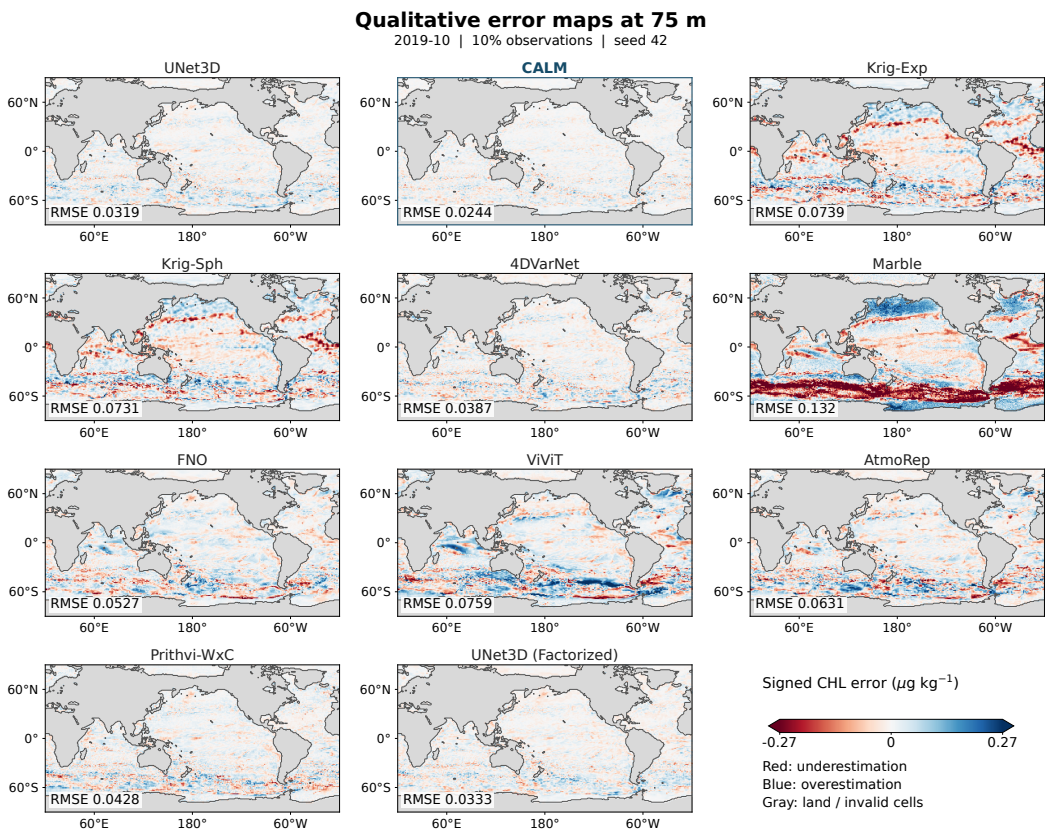


Figure 12: **Qualitative error maps at 75 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

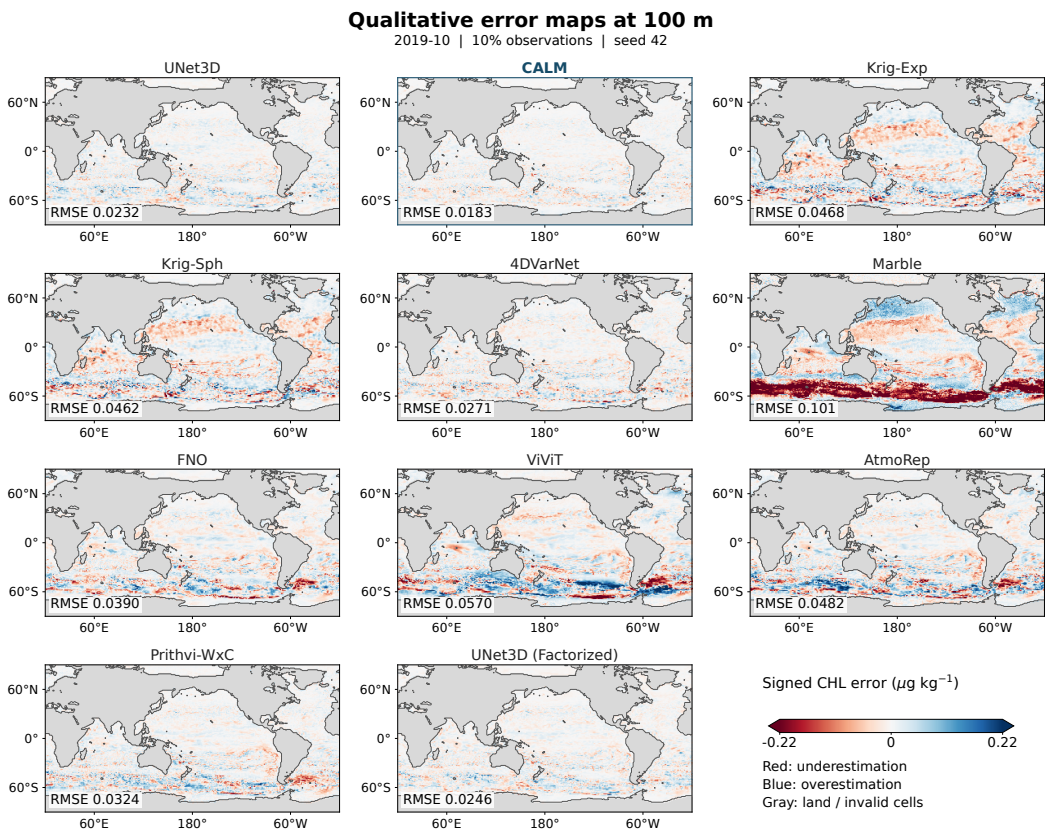


Figure 13: **Qualitative error maps at 100 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

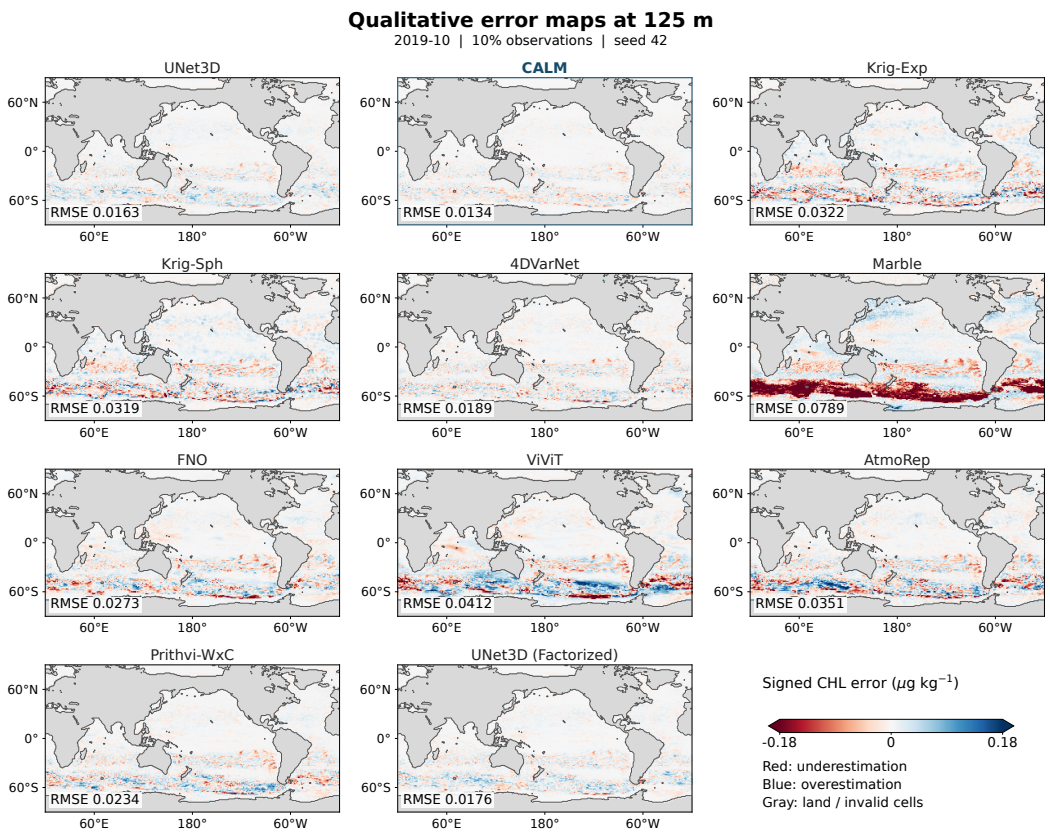


Figure 14: **Qualitative error maps at 125 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

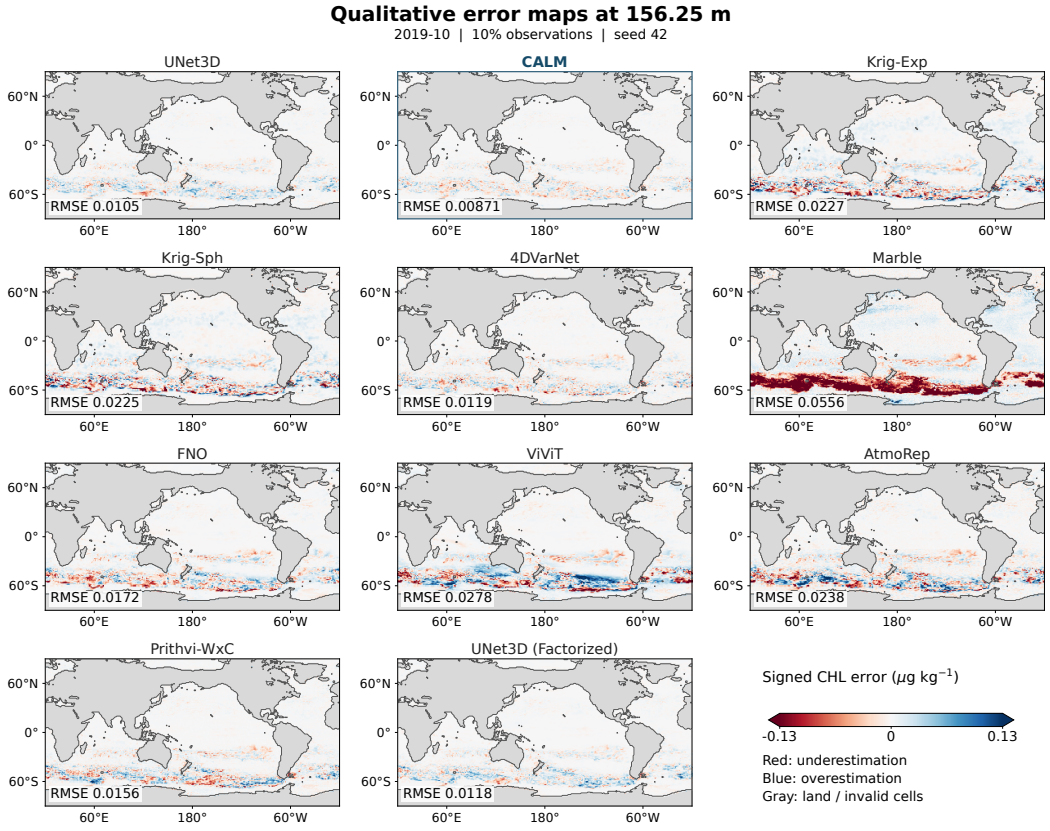


Figure 15: **Qualitative error maps at 156.25 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

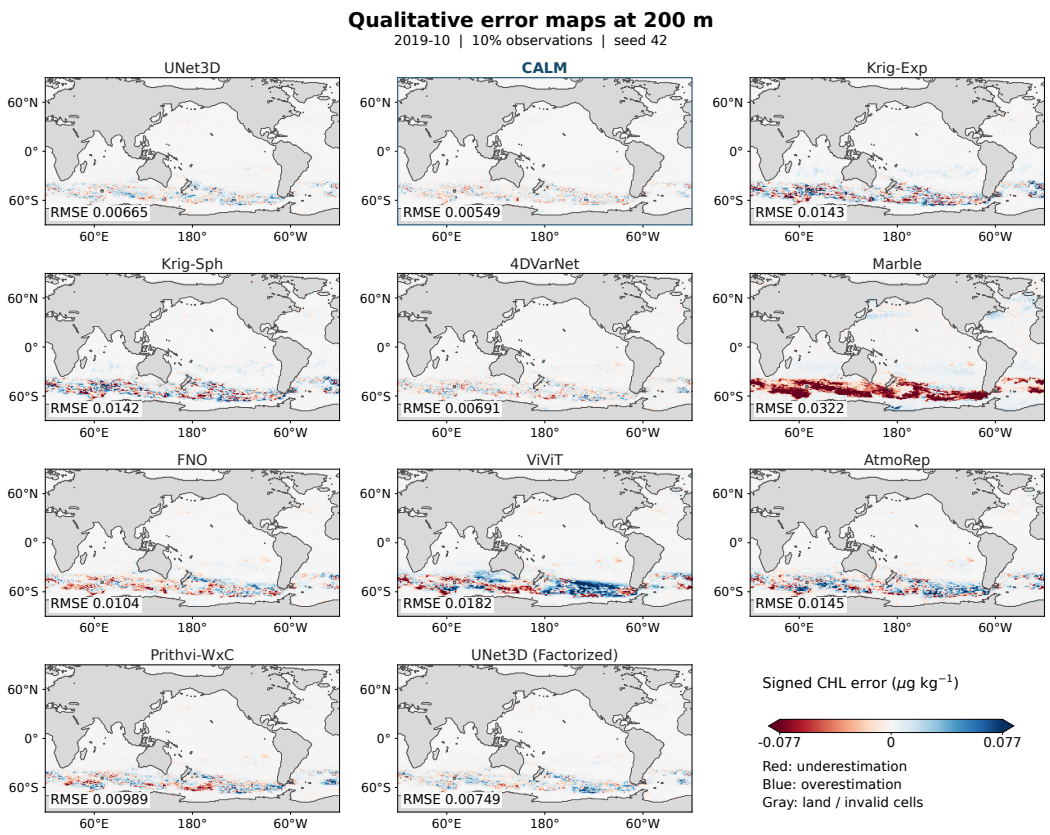


Figure 16: **Qualitative error maps at 200 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

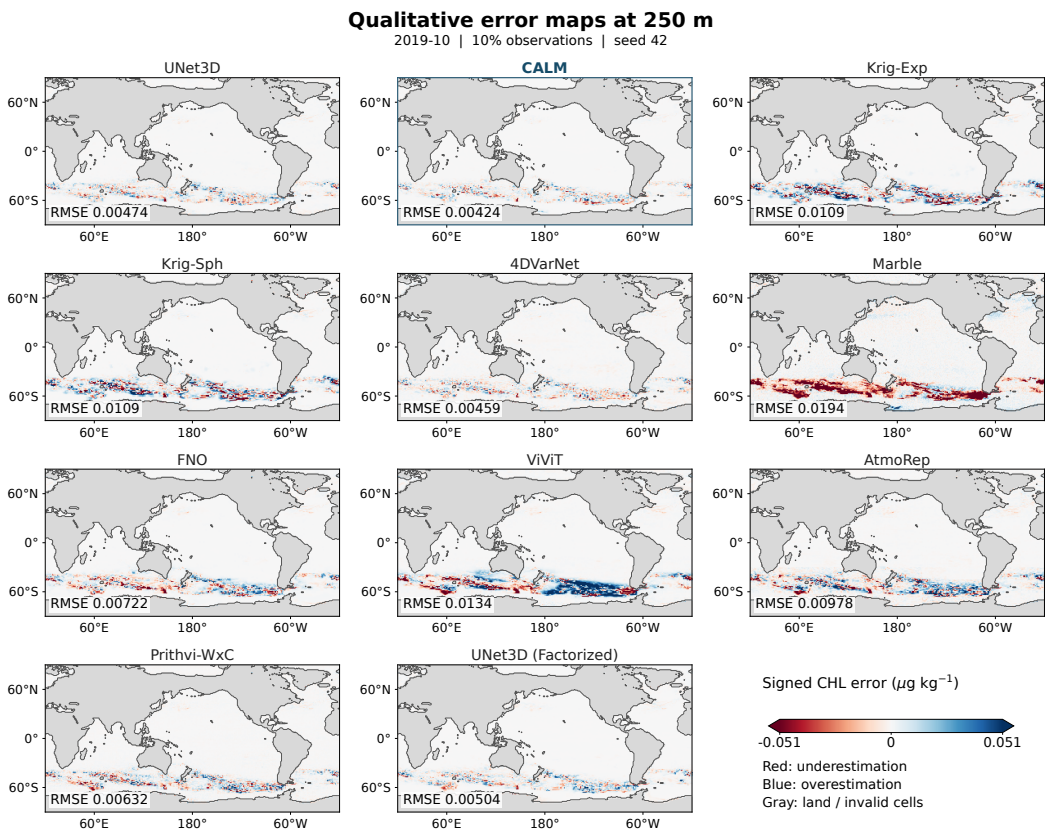


Figure 17: **Qualitative error maps at 250 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

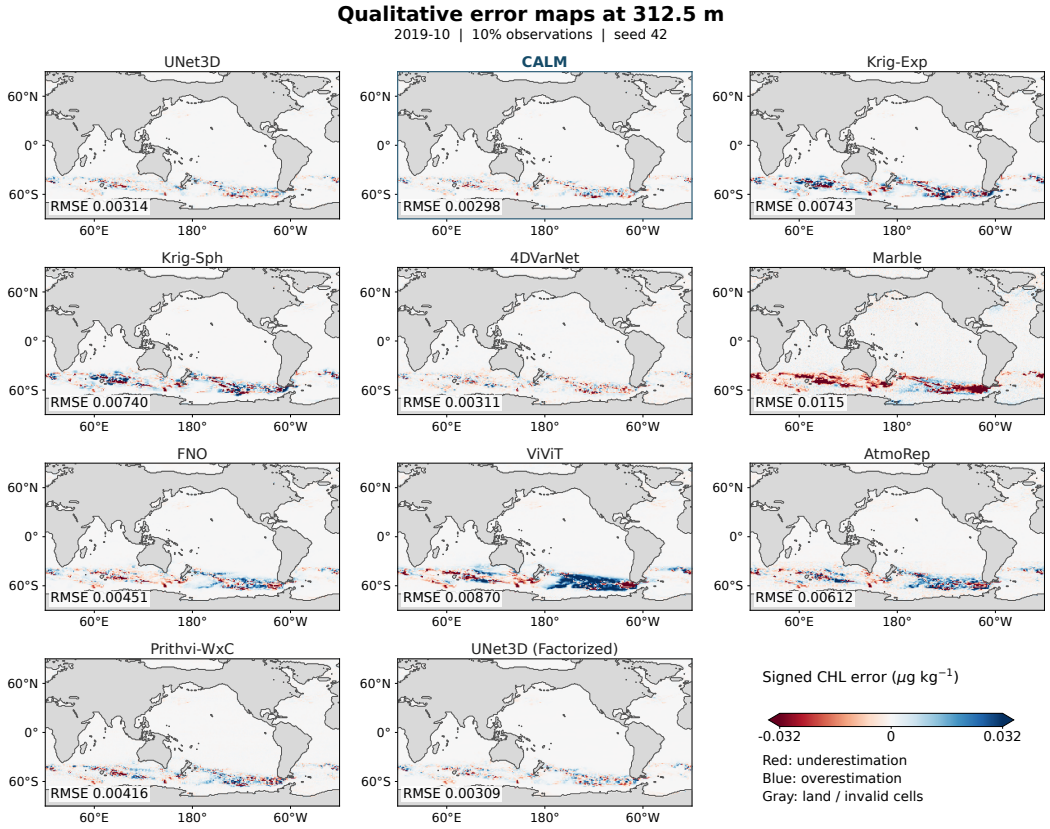


Figure 18: **Qualitative error maps at 312.5 m.** All 11 methods for the October 2019 case. Display and selection protocols are specified in Appendix E.

F LIMITATIONS

Attribution of reconstruction gains. The current experiments do not isolate the carbon-aware mechanisms as the cause of CALM's improved reconstruction. Backbone design, contextual inputs, and optimization also contribute, and the component ablations do not show a uniform accuracy gain from every carbon constraint. More tightly controlled comparisons across multiple seeds are needed to determine how much of the improvement is attributable specifically to carbon awareness.

Carbon-budget closure and carbon-state validity. Closing the represented carbon budget does not establish that the reconstructed carbon states are scientifically reliable. OC lacks direct supervision, and changes in its magnitude can be compensated by the learned chlorophyll-to-carbon ratio. DIC supervision and conservation constrain the solution but do not uniquely identify OC or validate its ecological interpretation. Windowwise budget closure also does not guarantee agreement between carbon states initialized independently in overlapping windows. Independent OC observations are needed to turn the accounting guarantees into evidence for trustworthy carbon states.

Argo-derived product evaluation. The Argo experiment evaluates reconstruction of an interpolated product. Because interpolation precedes input sampling, held-out observations can influence interpolated inputs, limiting conclusions about generalization to independent raw profiles.