

Ref: essd-2026-375

FDU-BTR: a physics-guided ensemble learning reconstruction of global surface-ocean fCO₂ (1982–2024) with uncertainty diagnostics

Zhenguo Wang¹, Weiwei Fu^{1,2}

¹Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai, 200438, China

² Institute of Eco-Chongming (IEC), 1050 Baozhen, Lühua Town, Chongming District, Shanghai 202151, China

Response Letter to Reviewers' Comments

We sincerely thank the Editor and both reviewers for their careful reading of the manuscript and for their constructive and detailed comments. Both reviewers raised the same overarching concern: the manuscript overstated the generalizability of the method because its only genuinely independent comparison was with the HOT and BATS time series, which—as noted by McKinley et al. (2026)—are comparatively easy to reproduce, while the remaining “independent” data (NCEI moorings) were in fact part of SOCAT. We fully accept this diagnosis. Our revision addresses this concern on three levels: (i) we corrected the independence error by removing the NCEI comparison and restructuring our evaluation into internal OOF diagnostics and external observational assessments; (ii) we added spatial and spatiotemporal OOF evaluations, together with five external datasets, to test transfer to undersampled regions, and we report performance stratified by coverage density; and (iii) we now state plainly in the Methods and Discussion that the fixed six-product climatological background remains SOCAT-informed. We also removed the SOCAT-constrained pH/DIC predictors from the final feature set. Thus, held-out SOCAT years do not constitute a fully independent test of the complete system. Reviewer #2's specific point—that the seasonal signal emphasized by McKinley et al. (2026) is already encoded in our prior—is now explicitly acknowledged. We hope these changes demonstrate that we have rebuilt the evidentiary basis of the paper rather than merely softened its wording.

Reviewer #1:

This manuscript presents FDU-BTR, a global monthly $1^\circ \times 1^\circ$ reconstruction of surface-ocean pCO₂ for 1982–2024. The proposed Background–Thermal Residual framework decomposes observed pCO₂ into a multi-product background climatology, an explicit thermal-anomaly term, and a learned residual, with basin-specific CatBoost models and boundary blending. The paper addresses an important problem for ocean carbon-cycle research: how to reconstruct surface-ocean pCO₂ in the presence of strong spatial and temporal sampling biases, while accounting for the physics. The manuscript is timely, potentially useful for the community, and the attempt to combine

physical structure, machine learning, product intercomparison, and coverage diagnostics is valuable.

However, I think the manuscript requires a lot of work before publication. My main concern is that several claims are stronger than what the validation and uncertainty analysis currently support, especially regarding global spatial skill, dominant sources of product uncertainty, and readiness for air–sea pCO₂ flux assessment. The method is promising, but the validation framework needs to be strengthened, and the claims should be more carefully qualified. Some “independent” validation is actually not independent as the data is included in SOCAT, so this raises concerns about the stated independence of the validation dataset and should be clarified carefully

Major comments:

1. Independent validation does not fully test the central gap-filling problem

The paper correctly emphasizes that global pCO₂ reconstruction is fundamentally a gap-filling problem, especially in regions with sparse SOCAT coverage. However, much of the independent validation focuses on time-series performance at well-observed locations, especially HOT and BATS. These are useful benchmarks, but they mainly test temporal variability at long-term subtropical stations. They do not provide a strong independent test of spatial extrapolation into poorly sampled regions, which is where the product is most needed and where uncertainty is expected to be largest.

The NCEI buoy validation is actually part of SOCAT, so definitely not an independent validation. Besides, it is not necessarily a useful addition, especially because it includes coastal, and coral-reef environments which is not expected to excel in this context. This validation is still based on point time series matched to a 1° monthly product, and the results show clear degradation in coastal and reef environments. The paper should therefore be more cautious in presenting the product as globally validated. Good agreement at HOT/BATS and open-ocean buoys does not necessarily demonstrate skill in the Southern Ocean, high-latitude winters, the South Pacific, the South Atlantic, marginal seas, coastal upwelling zones, or other regions where the product performs the most gap filling.

The authors should either temper claims about global spatial skill or add a stronger spatially independent validation. Possible options include spatial block cross-validation, leave-one-region-out or leave-one-basin-out tests, withholding data from under-sampled regions, validation stratified by SOCAT coverage density, and/or comparison with independent station compilations across broader regions. The SPOTS compilation from Lange et al. (2024), for example, could provide a more convincing independent evaluation across different regions and observing regimes, although still only a temporal independent validation.

Response: We thank the reviewer for identifying this weakness, which prompted the most substantial revision of the paper. We agree on both points: the original design leaned too heavily on well-observed time-series sites, and the NCEI moorings—being part of SOCAT—should not have been called independent.

We removed the NCEI buoy comparison and introduced Temporal, Spatial ($10^\circ \times 10^\circ$ blocks), and Joint spatiotemporal out-of-fold (OOF) evaluations. These test transfer of the learned residual model under withheld observations, but we no longer describe them as independent validation of the complete reconstruction because the fixed climatological background inherits information from existing products. We also expanded the separate external-observation assessment to LDEO, GLODAP, SPOTS, HOT, and BATS. LDEO records coincident with SOCAT at the grid-cell-month level were screened; GLODAP and BATS $f\text{CO}_2$ values were calculated from carbonate-system observations. The breadth and remaining source-level limitations of these comparisons are now stated more carefully.

HOT and BATS are presented as targeted tests of temporal variability at well-observed open-ocean stations, not as evidence that poorly sampled regions have been independently validated. Regional and seasonal OOF performance is provided in Fig. S1, and the fixed-threshold coverage analysis is reported in Table 4. Crucially, the OOF results now quantify the extrapolation penalty anticipated by the reviewer: RMSE increases from $9.13 \mu\text{atm}$ in the high-coverage stratum to $13.58 \mu\text{atm}$ in the low-coverage stratum (Sect. 4.2, Table 4).

Changes in the revised manuscript:

external datasets and processing, Sect. 2.4 (lines 119–135), we added:

“We compiled five external observational datasets to evaluate reconstructed surface-ocean $f\text{CO}_2$ beyond the SOCAT training samples: the Global Ocean Surface Water Partial Pressure of CO_2 Database from Lamont–Doherty Earth Observatory (LDEO; Takahashi et al., 2019), the Synthesis Product for Ocean Time Series (SPOTS; Lange et al., 2024), the Global Ocean Data Analysis Project (GLODAP; Lauvset et al., 2024), the Bermuda Atlantic Time-series Study (BATS; Steinberg et al., 2001), and the Hawaii Ocean Time-series (HOT; Karl and Lukas, 1996). These datasets provide complementary constraints across spatial and temporal scales. LDEO provides extensive underway surface-ocean CO_2 observations, GLODAP provides quality-controlled discrete carbonate-system measurements, and SPOTS, BATS, and HOT provide long-term time-series observations. HOT records represented in SPOTS were excluded from the SPOTS subset to avoid duplication with the dedicated HOT dataset.

Source-specific preprocessing was applied before harmonization. For LDEO, quality-controlled surface-ocean $p\text{CO}_2$ observations were retained, and records coincident with SOCAT training observations in the same year, month, and 1° grid cell were excluded.”

OOF design, Sect. 3.3.4 (lines 325–344), we added:

“Model generalization to samples not used for training was evaluated using three complementary out-of-fold (OOF) validation schemes designed to test temporal, spatial, and joint spatiotemporal prediction. In each outer OOF fold, feature selection and hyperparameter optimization were repeated using only the corresponding training samples, separately for each modeling region.

Hyperparameters were selected by inner cross-validation using contiguous temporal blocks. The outer test samples were excluded from both model training and model selection and were used only to generate and evaluate the final OOF predictions. Bias, MAE, RMSE, mean absolute percentage error (MAPE) and the coefficient of determination (R^2) were calculated from paired observations and predictions.”

OOF results, Sect. 4.1 (lines 400–428) and Fig. S1, we added:

“The transferability of the learned residual relationships under different forms of withheld observational support was evaluated using Temporal, Spatial, and Joint spatiotemporal OOF schemes (Fig. 3). Because the BTR background is a fixed multi-product monthly climatology constructed in advance, these experiments specifically assess the out-of-fold generalization of the learned residual component rather than providing fully independent validation of the complete FDU-BTR system.

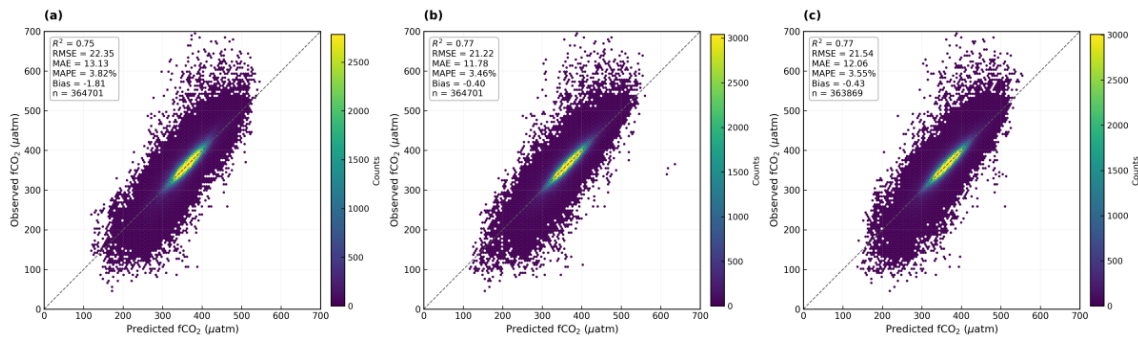


Figure 3. Out-of-fold (OOF) validation of FDU-BTR under different withholding schemes. Observed versus predicted surface-ocean fCO_2 for (a) temporal OOF, (b) spatial OOF, and (c) joint spatiotemporal OOF. Colors indicate sample density, and the dashed line denotes the 1:1 relationship.

Aggregate predictive performance remains similar across the three withholding schemes. RMSE ranges from 21.22 to 22.35 μatm , MAE from 11.78 to 13.13 μatm , and R^2 from 0.75 to 0.77. Spatial and Joint OOF yield RMSE values of 21.22 and 21.54 μatm , respectively, while Temporal OOF is moderately higher at 22.35 μatm . Predicted and observed fCO_2 values remain broadly aligned with the 1:1 relationship under all three schemes. The relatively small differences among the three schemes indicate that the learned residual relationships remain broadly transferable when temporal, spatial, or combined spatiotemporal support is withheld, rather than depending strongly on nearby training samples.

This aggregate stability, however, masks pronounced regional and seasonal heterogeneity (Fig. S1).

NH-MH consistently shows the largest regional RMSE, whereas SH-ST exhibits the lowest errors across the three OOF schemes. Semi-enclosed seas (ESE) show particularly strong sensitivity to the withholding strategy: Temporal OOF produces a bias of approximately $-8.81 \mu\text{atm}$, whereas Spatial OOF remains close to zero and Joint OOF shows a substantially smaller negative bias, indicating greater difficulty in temporal transfer within this region. Seasonal performance shows a similarly structured contrast. RMSE increases consistently from approximately $18.73 \mu\text{atm}$ in DJF to $25.62 \mu\text{atm}$ in SON, while biases under Spatial and Joint OOF remain comparatively small. The seasonal deterioration is therefore expressed mainly as increased prediction scatter rather than a systematic shift in the reconstructed mean state.

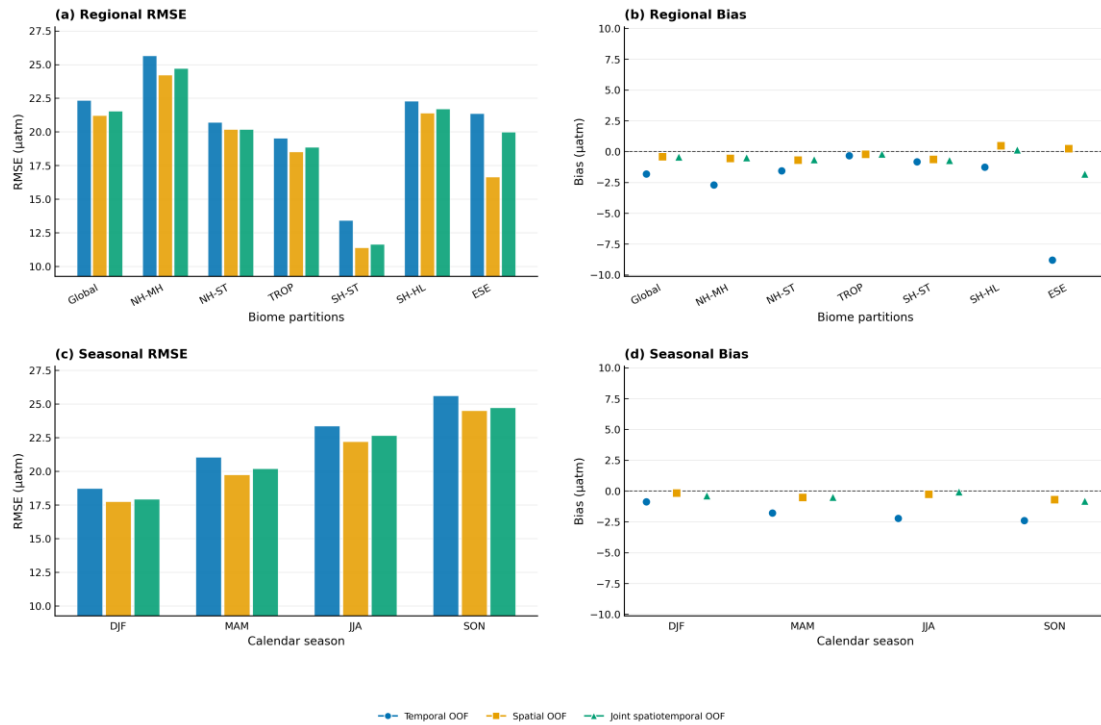


Figure S1. Regional and seasonal performance of FDU-BTR under temporal, spatial, and joint spatiotemporal out-of-fold (OOF) validation.

These OOF experiments show that broadly stable aggregate performance can coexist with substantial regional and seasonal differences in reconstruction difficulty. The largest departures are concentrated in specific regimes, particularly NH-MH in overall error and ESE in temporal transfer, rather than being distributed uniformly across the global ocean. This distinction motivates the subsequent examination of whether differences in observational support alone can explain the spatial variation in prediction skill. Biome-specific feature combinations and feature importance are presented in Fig. S2.”

external observations, Sect. 5.2 (lines 519–538), we added:

“FDU-BTR was evaluated against external LDEO and GLODAP observations to assess its performance beyond the SOCAT training samples (Fig. 6). For LDEO, FDU-BTR achieved an RMSE of 28.89 μatm , an MAE of 17.47 μatm , a bias of 0.11 μatm , and an R^2 of 0.67 across 16,072 matched observations. Against GLODAP, the corresponding values were 30.75 μatm , 17.78 μatm , -2.67 μatm , and 0.65 across 10,204 matched observations. The LDEO comparison shows little systematic offset, whereas the GLODAP comparison exhibits a modest negative bias together with higher RMSE and lower R^2 , indicating larger discrepancies for the conditions represented by that dataset.

For both datasets, most observations are concentrated near the 1:1 relationship over the central fCO_2 range, while larger departures occur at the low and high ends of the observed distribution. Reconstructed values tend to be compressed toward the central range under these more extreme conditions, indicating that anomalously low or high fCO_2 values are less fully reproduced than the bulk of the observations. This error structure suggests that the main limitation is associated more with the amplitude of extreme or locally anomalous variability than with a uniform systematic bias. These comparisons show that FDU-BTR reproduces the central range of externally observed surface-ocean fCO_2 with broadly consistent performance across the two observational datasets, while reconstruction errors increase under more extreme conditions. The difference in performance between LDEO and GLODAP further indicates that predictive skill depends on the oceanographic conditions and sampling characteristics represented by the validation data rather than being spatially uniform across the global ocean.

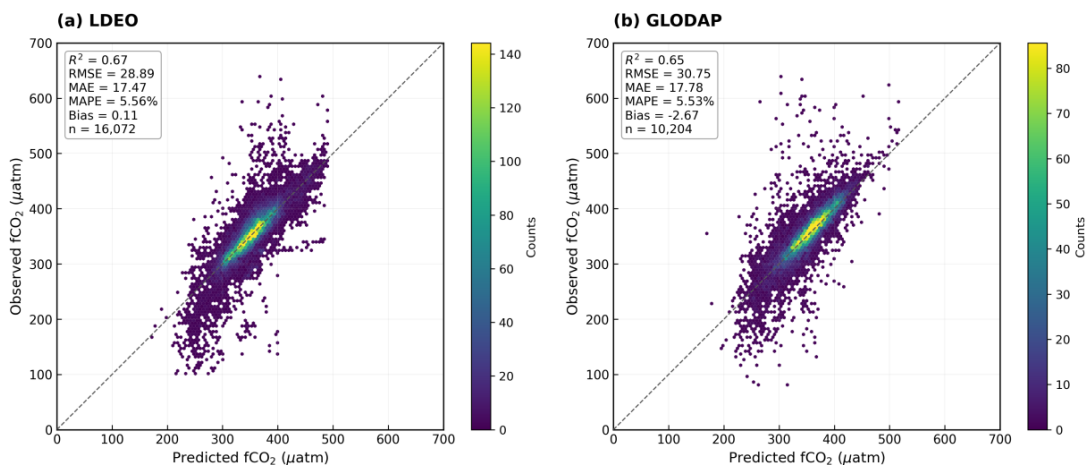


Figure 6. Comparison of FDU-BTR reconstructed fCO_2 with external LDEO and GLODAP observations.”

HOT/BATS, Sect. 5.3 (lines 539–560), we added:

Evaluation at the HOT and BATS time-series stations provides a site-specific assessment of how

well FDU-BTR represents surface-ocean $f\text{CO}_2$ variability across different temporal scales (Fig. 7). At both sites, FDU-BTR reproduces the pronounced seasonal cycle and sustained long-term increase observed over the measurement period, with generally close agreement at the monthly scale. Some short-term excursions and periods of enhanced variability are less fully resolved, indicating that reconstruction fidelity varies across temporal scales.

This scale dependence becomes clearer after removal of the seasonal cycle. FDU-BTR retains the broad interannual evolution and long-term increase at both stations, but smooths the amplitude of some individual year-to-year anomalies. Despite these differences in short-term variability, the reconstructed long-term trends remain statistically consistent with the observational estimates. Thus, the agreement in long-term evolution persists even where individual interannual excursions are not fully reproduced.

Overall, the HOT and BATS records show that, at these two well-observed open-ocean sites, FDU-BTR captures seasonal structure and long-term change more consistently than the amplitude of individual interannual anomalies. These evaluations therefore provide targeted evidence for the representation of temporal variability at long-term open-ocean stations, rather than a general assessment of reconstruction performance across the global ocean.

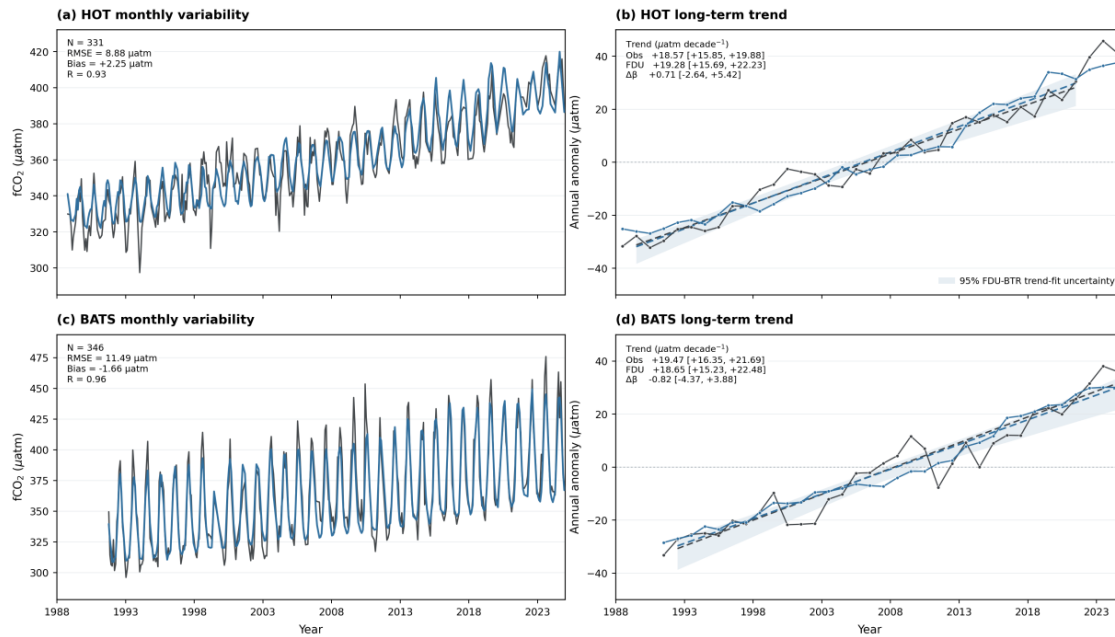


Figure 7. Validation of monthly variability and long-term trends in surface-ocean $f\text{CO}_2$ at the HOT and BATS time-series stations. (a, c) Comparison of monthly FDU-BTR $f\text{CO}_2$ estimates with station observations at HOT and BATS, respectively. (b, d) Annual deseasonalized $f\text{CO}_2$ anomalies and fitted long-term trends at HOT and BATS, respectively. The shaded band represents the pointwise 95% bootstrap interval of the fitted FDU-BTR trend, estimated from 2,000 paired circular moving-block resamples of annual deseasonalized anomalies using five-year blocks.

2. Validation of air–sea CO₂ flux applications

The manuscript repeatedly motivates the product as a basis for air–sea pCO₂ flux assessment. However, the validation focuses primarily on pCO₂ RMSE, correlation, and bias. These are useful but not sufficient for the stated flux application. Air–sea pCO₂ flux depends on the air–sea pCO₂ difference, not only on ocean pCO₂ in isolation. A relatively small pCO₂ bias can still affect flux estimates if it changes the sign, seasonal phasing, or regional magnitude of $\Delta p\text{CO}_2$.

The authors should evaluate whether the product preserves the correct ocean–atmosphere pCO₂ gradient. This could include validation of $\Delta p\text{CO}_2$ regional and seasonal bias in $\Delta p\text{CO}_2$, the frequency of incorrect source/sink classification, and ideally implied air–sea pCO₂ fluxes using a consistent atmospheric CO₂ field and gas-transfer formulation. If the product is intended to support flux assessment, this should be part of the validation, or the flux-related claims should be softened.

Response: We appreciate the reviewer’s request for a flux-specific assessment. We agree that skill in reconstructing oceanic fCO₂ alone does not establish the accuracy of the air–sea gradient or the resulting CO₂ flux.

We now compute $\Delta f\text{CO}_2$ and air–sea CO₂ fluxes for FDU-BTR and the comparison products using a common atmospheric CO₂ field, common auxiliary inputs, and a single gas-exchange formulation. Fig. 10 compares the global annual estimates, Figs. S4–S5 provide the eight-region breakdown, and Fig. 11 places the global sink alongside GCB2025. After the riverine and warm-layer/cool-skin adjustments, the 2015–2024 mean is 3.26 Gt C yr^{–1}, close in magnitude to the GCB2025 merged estimate of 3.2 ± 0.4 Gt C yr^{–1}. We are careful to describe these as harmonized inter-product and budget comparisons, rather than as an independent observational validation of the flux or of local source–sink classification, and we have moderated the flux-related claims accordingly. The global riverine adjustment is applied to the integrated annual sink, whereas the warm-layer/cool-skin adjustment is treated separately when period means are compared with the GCB merged estimate. Changes: Sect. 7 (lines 670–753), Figs. 10–11, and Figs. S4–S5, we added:

“The global mean $\Delta f\text{CO}_2$ and air–sea CO₂ flux derived from FDU-BTR both show an overall strengthening during 1982–2024, superimposed on pronounced interannual and decadal variability (Fig. 10). The external surface-ocean CO₂ products show broadly similar long-term evolution, including several common low-frequency peaks and troughs, although substantial differences remain in the absolute magnitude of the estimated sink. Agreement among products is therefore stronger in the direction and timing of long-term change than in the magnitude of ocean CO₂ uptake.

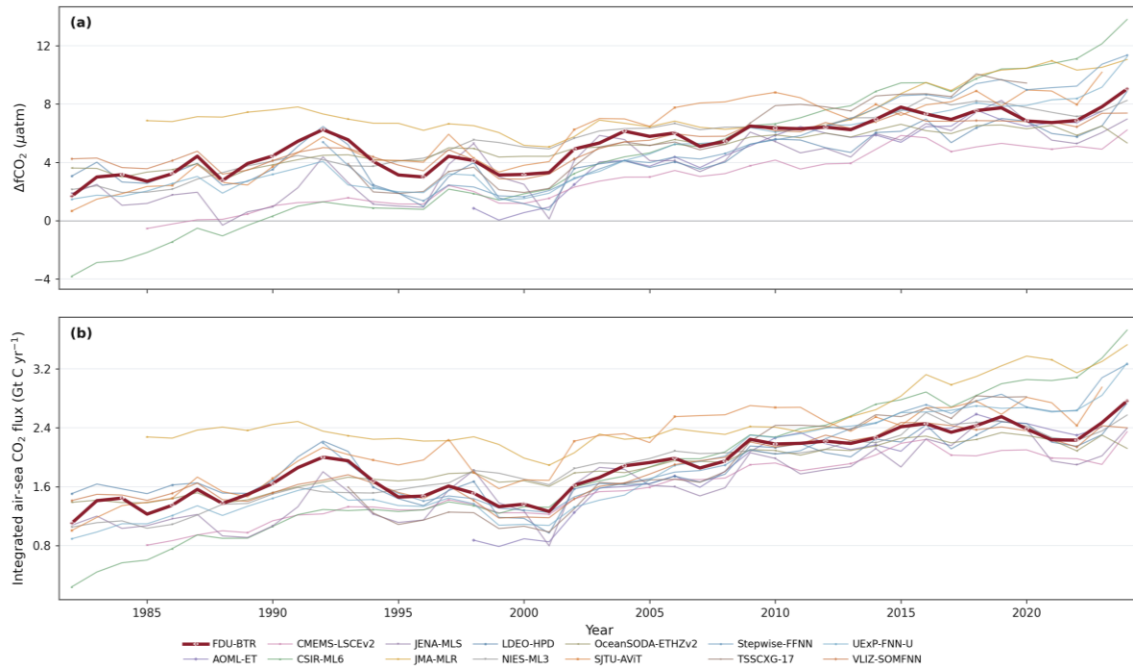


Figure 10. Global annual $\Delta f\text{CO}_2$ and air–sea CO_2 flux from FDU-BTR and 13 external products (1982–2024).

Whether this global strengthening reflects synchronous changes across ocean regions was then examined using the regional flux estimates (Figs. S4 and S5). The regional results show that it does not. The Arctic, North Atlantic, North Pacific, and Indian Ocean generally exhibit strengthening CO_2 uptake over the study period, whereas the South Atlantic evolves from relatively weak toward stronger uptake. The South Pacific remains close to long-term source–sink balance, with interannual variability large relative to its mean exchange. The Southern Ocean exhibits pronounced decadal variability, including weakening from the late 1990s to the early 2000s followed by subsequent recovery and strengthening. Enclosed and semi-enclosed seas (ESE) show the largest inter-product spread, including periods in which products differ in the inferred sign of net air–sea CO_2 exchange. FDU-BTR indicates a long-term transition in ESE from net CO_2 release toward net uptake, whereas most external products do not show a comparable transition. A comparable source-to-sink transition has previously been reported for the Mediterranean Sea, which shifted from a net CO_2 source in the 1960s to a net sink in the 1990s (Louanchi et al., 2009). Thus, the long-term strengthening of the globally integrated ocean carbon sink emerges from distinct regional trajectories rather than a synchronous increase in CO_2 uptake across all ocean basins.

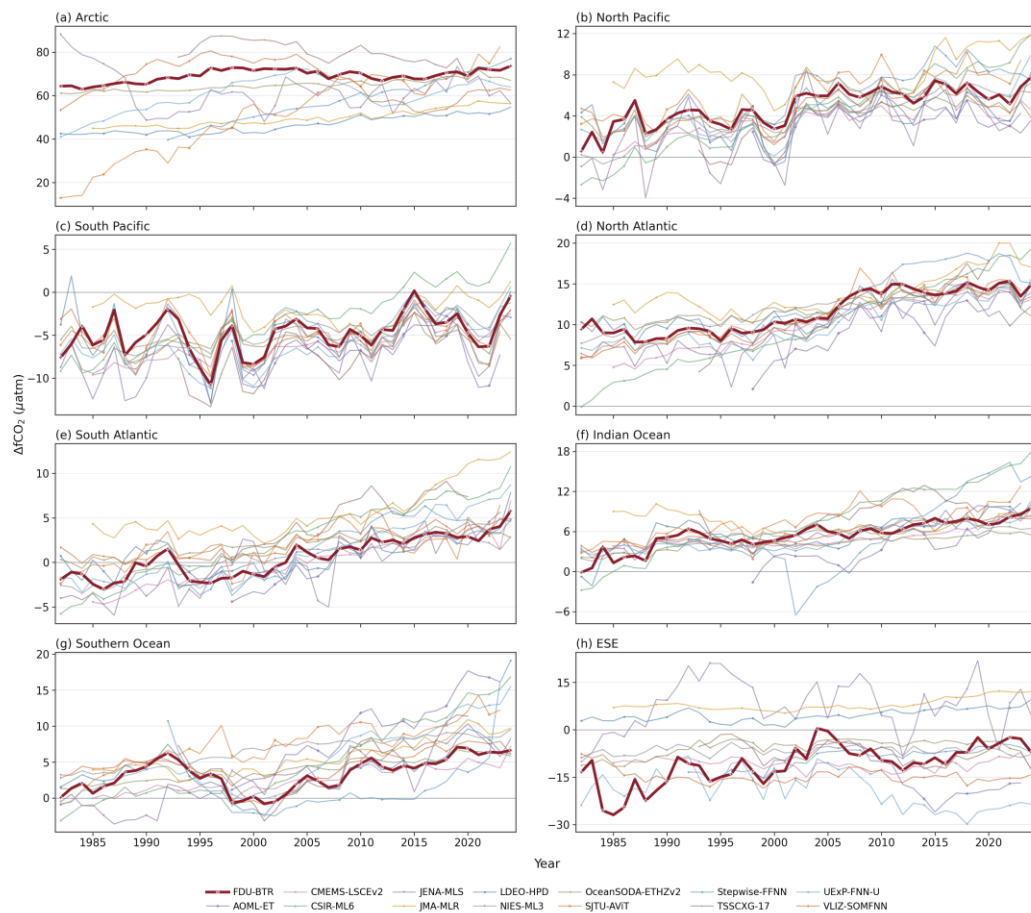


Figure S4. Regional annual $\Delta f\text{CO}_2$ across eight ocean regions from FDU-BTR and 13 external products (1982–2024). Positive values indicate CO_2 uptake by the ocean.

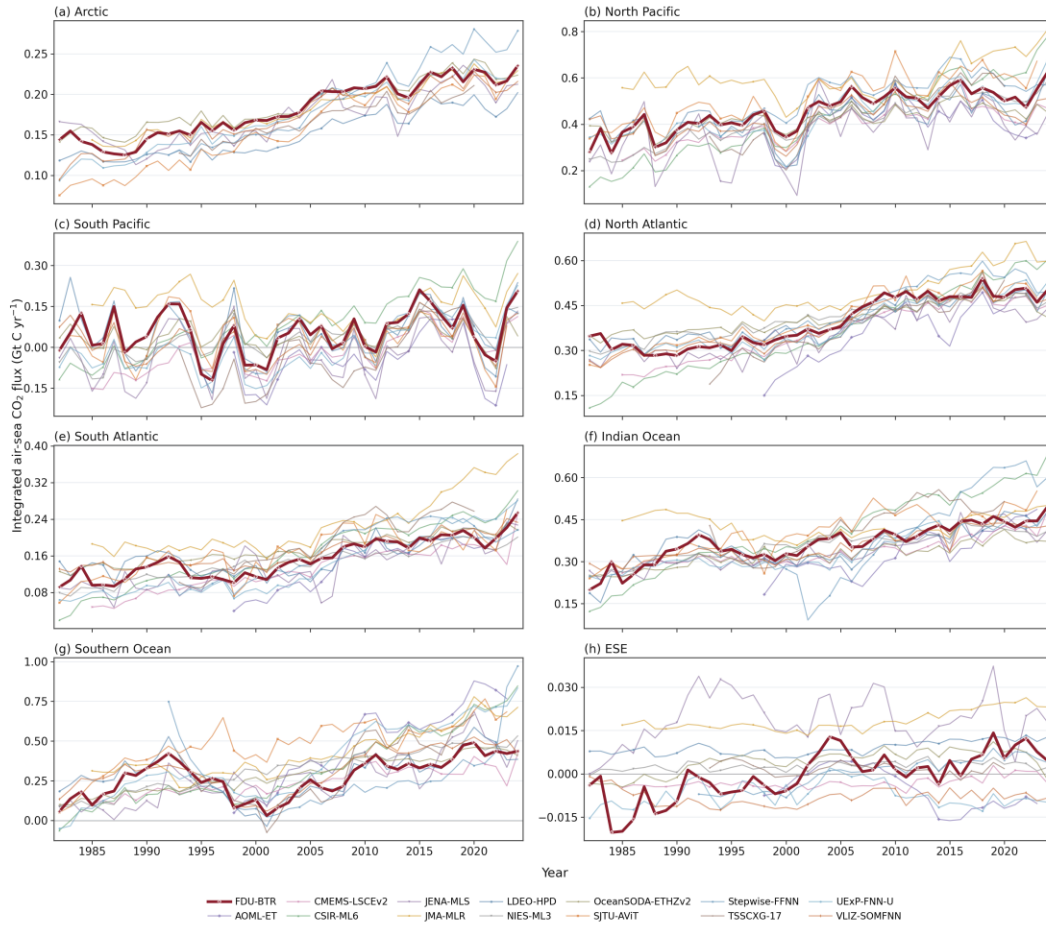


Figure S5. Regional annual air–sea CO₂ flux across eight ocean regions from FDU-BTR and 13 external products (1982–2024). Positive values indicate CO₂ uptake by the ocean.

The position of the FDU-BTR ocean carbon sink within the broader GCB2025 framework was further examined by comparing its temporal evolution with the GCB2025 mean of the surface-ocean fCO₂ data products (Fig. 11). FDU-BTR reproduces the major low-frequency features evident in the data-product mean, including strengthening in the early 1990s, subsequent weakening toward the late 1990s and early 2000s, and a more sustained increase after the early 2000s. Although the two estimates differ in absolute magnitude, particularly after 2000 when FDU-BTR is generally slightly lower than the data-product mean, they show broadly consistent long-term evolution. The GOBMs (global ocean biogeochemistry models) mean remains systematically lower over much of the later period, resulting in a persistent separation between the observation-based and model-based estimates of ocean CO₂ uptake.

After applying the global riverine correction of +0.65 Gt C yr⁻¹, the FDU-BTR ocean carbon sink is 2.20, 2.17, 2.75, and 3.08 Gt C yr⁻¹ for 1982–1994, 1995–2004, 2005–2014, and 2015–2024,

respectively (Fig. 11). For comparison with the GCB2025 merged ocean-sink estimate, an additional global warm-layer/cool-skin adjustment of $+0.18 \text{ Gt C yr}^{-1}$ was applied to the FDU-BTR period means (Ford et al., 2024). The corresponding adjusted estimates are 2.38, 2.35, 2.93, and $3.26 \text{ Gt C yr}^{-1}$, respectively. The adjusted 2015–2024 mean of $3.26 \text{ Gt C yr}^{-1}$ is close to the GCB2025 merged ocean-sink estimate of $3.2 \pm 0.4 \text{ Gt C yr}^{-1}$ (Friedlingstein et al., 2026).

FDU-BTR and the GCB2025 data-product mean show broadly consistent long-term evolution and major low-frequency variations, although differences remain in the absolute magnitude of the sink. The recent corrected FDU-BTR estimate is also consistent in magnitude with the GCB2025 merged estimate.

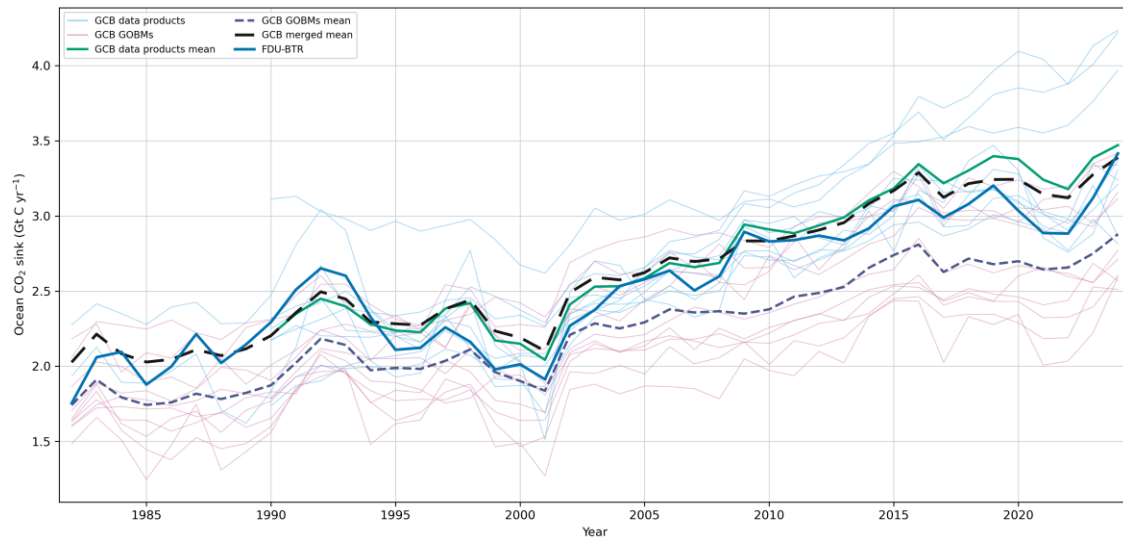


Figure 11. Global ocean CO₂ sink estimates from FDU-BTR and GCB2025 during 1982–2024. Thin lines show the individual GCB data products and GCB GOBMs. Bold lines show the GCB data products mean, GCB GOBMs mean, GCB merged mean, and FDU-BTR estimate. The FDU-BTR curve represents the study’s physical air–sea CO₂ sink with the $+0.65 \text{ Gt C yr}^{-1}$ river correction applied once. Positive values indicate net CO₂ uptake by the ocean from the atmosphere.

”

3. Decomposition framework

The BTR method decomposes pCO₂ into a multi-product background climatology, a thermal-anomaly term, and a residual term. This is a central methodological contribution. However, the validation mainly evaluates the final reconstructed pCO₂ field. It would be useful to assess the behavior of each component separately.

For example, how biased is the multi-product background climatology relative to independent observations? Does the thermal-anomaly term improve or degrade performance in different regions or seasons? Is the learned residual correcting real non-thermal variability, or is it compensating for errors in the prior climatology and thermal correction? This is especially important because the

authors acknowledge that biases in the background climatology or thermal term can propagate into the residual. A component-level evaluation would make the physical interpretation of the BTR framework much more convincing.

Response: We agree that a component-level evaluation is needed when the decomposition is central to the method. Evaluating only the final reconstruction left the contribution of each stage unclear.

We therefore added a stepwise component analysis using the same Spatial OOF samples. We evaluate the climatological background alone (B), the background plus the explicit thermal component (B + T), and the complete reconstruction after residual learning (B + T + $\hat{R}$). Table 5 reports RMSE, MAE, bias, and R^2 for the three stages. We also added regional and seasonal diagnostics showing the changes in RMSE when the thermal term and residual correction are introduced separately.

The results clarify that the background establishes the large-scale reference state (RMSE = 30.41 μatm), the thermal term provides a secondary adjustment whose effect varies by region and season (ΔRMSE ranging from an increase of 4.23 μatm to a reduction of 9.33 μatm), and the learned residual provides the largest and most consistent reduction in error (final RMSE = 21.22 μatm). Following the reviewer’s caution, we now state explicitly that the residual should not be interpreted as a unique “non-thermal process component”; it is a predictable correction that absorbs unresolved physical, biogeochemical, and prior-error contributions.

Changes: Sect. 3.1.3 (lines 199–211); Sect. 4.3 (lines 451–487); Table 5; Fig. 4, we added:

“The contribution of individual BTR components to reconstruction skill was evaluated by comparing three successive prediction stages using the same Spatial OOF samples: the monthly climatological background, the background plus the explicit thermal term, and the complete reconstruction after residual learning (Table 5). The background provides the large-scale spatial–seasonal reference state. Adding the explicit thermal term produces only a modest change in global error, whereas the largest reduction occurs after the learned residual correction is introduced. Relative to the background alone, the complete reconstruction reduces Spatial OOF MSE by approximately 51%. A more informative contrast, however, is provided by the regional and seasonal variation in the thermal and residual contributions.

Table 5. Incremental Spatial OOF skill of the BTR prediction stages.

Stage	RMSE (μatm)	MAE (μatm)	Bias (μatm)	R^2
Background B	30.41	22.15	−10.96	0.59
B+T	29.85	21.16	−8.47	0.59

$B + T + \hat{R}$	21.22	11.78	-0.40	0.77
.....				

The contribution of the explicit thermal term is strongly dependent on region and season (Fig. 4a). At the global scale, RMSE increases slightly by 0.28 and 0.10 μatm in DJF and MAM, respectively, but decreases by 1.51 μatm in JJA and 0.97 μatm in SON. NH-ST and SH-ST show error reductions in all four seasons, with the largest improvements occurring in NH-ST during JJA (3.60 μatm) and in SH-ST during DJF (2.13 μatm). The strongest seasonal improvement occurs in ESE, where RMSE decreases by 9.33 μatm in JJA. By contrast, the tropical region shows increased RMSE in DJF, MAM, and SON and almost no change in JJA. Explicitly representing the first-order temperature response therefore does not provide a uniform predictive benefit across the global ocean; its contribution depends strongly on regional and seasonal context.

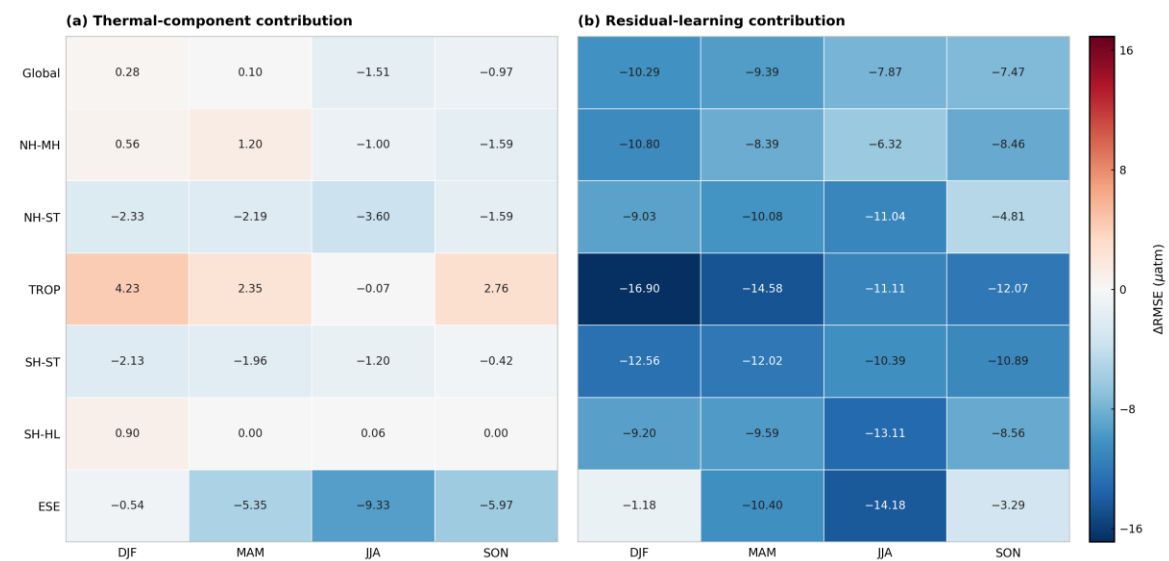


Figure 4. Regional and seasonal changes in RMSE associated with successive components of the BTR framework. (a) RMSE change after introducing the explicit thermal component; (b) RMSE change after introducing the learned residual correction. Negative values indicate reduced error relative to the preceding reconstruction stage.”

4. Uncertainty framework

The uncertainty diagnostics are a valuable part of the manuscript, but the current framework appears closer to an empirical diagnostic than a rigorous grid-cell-level uncertainty estimate. Mapping uncertainty is estimated from residuals at validation observations, but observations are not uniformly distributed in space or process regime. As a result, the inferred uncertainty may still be biased toward better-observed regions. This is important because the paper makes strong claims about characterizing spatial patterns and dominant sources of product uncertainty.

The authors should clarify that the uncertainty product is diagnostic and partly observation-location dependent. They should also test whether the reported uncertainties are calibrated. For example, do

regions with higher predicted uncertainty actually show larger independent residuals? Are uncertainty estimates reliable when stratified by basin, latitude, season, coastal/open-ocean class, and SOCAT coverage density? The phrase “dominant sources of product uncertainty” should be softened unless the authors can more clearly separate model structural uncertainty, prior-field uncertainty, observational uncertainty, and sampling/extrapolation uncertainty.

Response: We thank the reviewer for distinguishing an uncertainty diagnostic from a calibrated prediction interval; we agree that the former observation-location residual map could not justify grid-cell uncertainty statements in poorly sampled regions.

We replaced that analysis with CatBoost conditional quantile models followed by nested spatial-block conformal calibration to construct nominal 67% and 90% prediction intervals. The spatially held-out calibration procedure is described in Sect. 3.4. Importantly, calibration using SOCAT-derived samples does not guarantee nominal coverage for other observational sources or under arbitrary spatial shifts.

We therefore assessed the intervals against the five external observational datasets. Wider PI90 intervals generally correspond to larger absolute errors (sample-level Spearman $\rho = 0.39$), but empirical PI90 coverage across width groups ranges from 75.6% to 82.8%, below the nominal 90%. We now present interval width as a useful measure of relative prediction risk while explicitly reporting this undercoverage. The shallow-water and deep-water comparison further describes where the intervals differ; it does not attribute those differences uniquely to specific uncertainty sources.

We have also removed the previous language concerning “dominant sources of product uncertainty” and no longer attempt to decompose total uncertainty into observational, structural, prior, and sampling components.

Changes in the revised manuscript:

interval construction, Sect. 3.4 (lines 346–380) , we added:

“Conditional uncertainty in FDU-BTR predictions was characterized using conformalized quantile regression (CQR), following Bennington et al. (2022a), together with nested spatial-block conformal calibration. This procedure was used to construct 67% and 90% prediction intervals for surface-ocean $f\text{CO}_2$. Both the central model and the quantile models were implemented using CatBoost. The quantile models were fitted to the BTR residual,
.....”

external evaluation and depth contrast, Sect. 6.1 (lines 583–618), Fig. 8, we added:

“FDU-BTR predictive uncertainty exhibits pronounced spatial differences across the global ocean (Fig. 8). The mean full width of PI90 over 1982–2024 is relatively large in high-latitude regions and parts of

the tropics, but generally smaller in the subtropical open ocean. The 67% prediction interval (PI67) shows a spatial pattern broadly consistent with the 90% prediction interval (PI90), but is narrower overall (Fig. 8a, b). Across valid mapped cells, the mean full widths are 47.2 μatm for PI90 and 25.8 μatm for PI67. This indicates systematic regional differences in the degree of predictive constraint.

To examine whether the prediction intervals reflect actual prediction errors, we further evaluated them using five external observational datasets: LDEO, SPOTS, GLODAP, HOT, and BATS (Fig. 8c). All matched samples were divided into five groups according to increasing PI90 width, with Q1 representing the narrowest intervals and Q5 the widest. PI90 width is positively associated with absolute prediction error, with a sample-level Spearman correlation coefficient of 0.39 and a 95% bootstrap confidence interval of 0.35–0.44. The dataset-balanced MAE rises overall from 14.3 μatm in Q1 to 39.1 μatm in Q5. Although Q3 has a slightly lower MAE than Q2, errors are substantially larger in Q4 and Q5, indicating that interval width helps distinguish relative error risk rather than ranking every group strictly monotonically.

The empirical PI90 coverage (C90) across the five groups ranges from 75.6% to 82.8%, below the nominal 90% level. Thus, PI90 provides useful information about relative error risk but does not achieve nominal coverage for these external observations.

Predictive uncertainty shows an even clearer contrast between shallow and deeper waters (Fig. 8d). Using 200 m water depth as the threshold, the PI90 half-width in shallow-water regions is mostly approximately 39–56 μatm , whereas values in deeper-water regions are concentrated at approximately 18–21 μatm . PI67 shows a consistent depth-dependent contrast. Uncertainty in shallow-water regions also exhibits stronger month-to-month variability, whereas uncertainty in deeper-water regions is generally lower and more stable. Here, the PI half-width is defined as half the difference between the upper and lower prediction-interval bounds.

Overall, FDU-BTR predictive uncertainty shows two clear characteristics: first, larger actual prediction errors are generally associated with wider prediction intervals, indicating that interval width provides a useful measure of relative prediction risk; second, uncertainty is substantially higher in shallow waters than in deeper waters, indicating that accurate reconstruction in shallow-water and shelf environments remains particularly challenging. The pronounced shallow-water contrast also calls for greater caution when interpreting local variability in heterogeneous shallow-water environments.

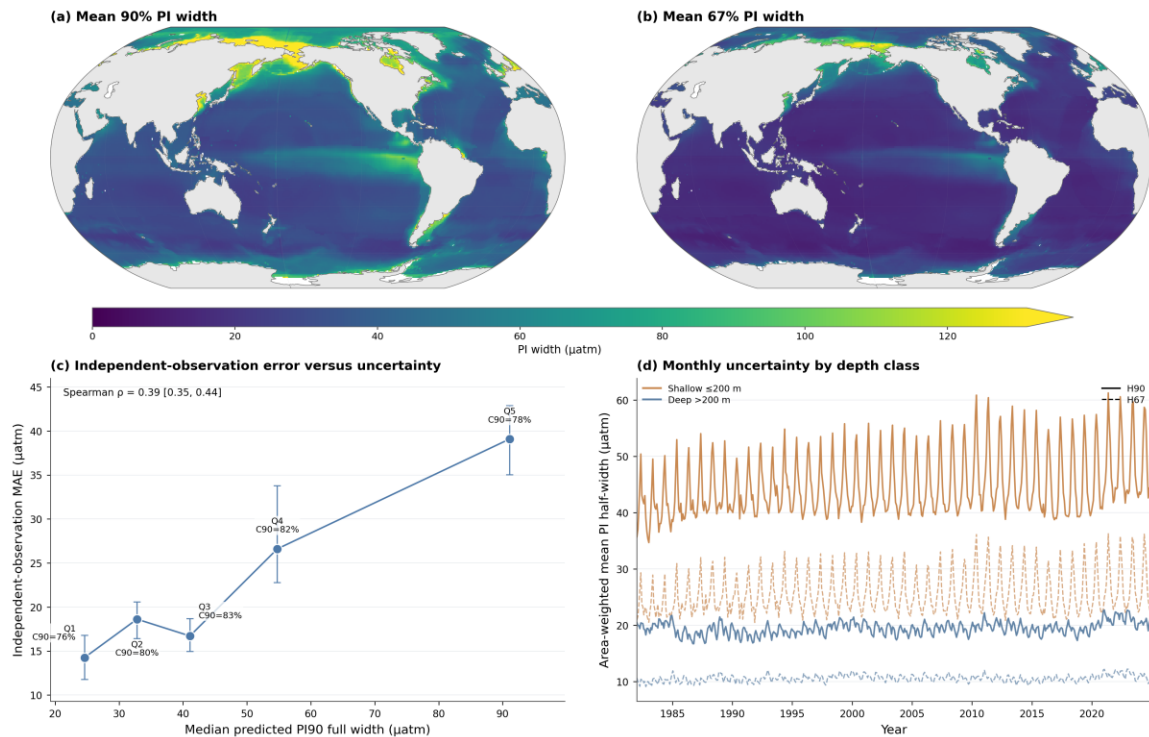


Figure 8. Spatial distribution and evaluation of prediction uncertainty in the FDU-BTR fCO_2 product (1982–2024). (a, b) Mean spatial distributions of the 90% and 67% prediction-interval widths, respectively; wider intervals indicate greater estimated uncertainty. (c) Dataset-balanced MAE versus median PI90 width for five groups of external observations, from the narrowest intervals (Q1) to the widest (Q5). Dataset-balanced MAE is the unweighted mean of the dataset-specific MAEs within each group. Labels show the empirical PI90 coverage (C90) and sample size; the reported Spearman correlation is calculated at the individual-sample level between PI90 width and absolute prediction error. (d) Monthly area-weighted mean prediction-interval half-widths in shallow and deeper waters.”

5. Use of other reconstructed carbonate-system products as predictors

The model uses pH and DIC from CMEMS/LSCE products as predictors. These variables may improve predictive skill, but they also raise questions about independence and circularity. If pH and DIC are themselves reconstructed fields that use related observations, assumptions, or machine-learning methods, then FDU-BTR is not purely constrained by independent environmental predictors. It may partly inherit structure from existing carbon products.

The authors should discuss whether using reconstructed pH and DIC introduces dependence on another pCO_2 /carbon-system product, whether it affects the independence of validation and product intercomparison, and how sensitive FDU-BTR is to these predictors. A useful sensitivity test would be to train and validate a version without pH and DIC predictors, or at least quantify how much they affect performance, trends, and uncertainty patterns (that may propagate through the full prediction).

Response: We agree that reconstructed carbonate-system predictors created a potential dependence on existing SOCAT-constrained products and complicated the interpretation of both validation and inter-product comparison.

We removed pH, DIC, alkalinity, and related reconstructed carbonate-system predictors from the final predictor set and retrained the models. This changed pointwise RMSE by only 0.16 μatm , confirming that these predictors were not the principal source of skill. We are careful not to overstate the fix: it does not remove the separate dependence introduced by the six selected external products used to define the fixed climatological background; we address that explicitly in our response to Reviewer 2 and in the revised Discussion.

6. The held-out-year validation still does not test spatial extrapolation

The fixed-year validation design is useful for testing temporal generalization, but it does not prevent spatial information from nearby or repeatedly sampled locations from entering the training set. Since SOCAT contains repeat cruise tracks and long-term sampling locations, holding out complete years may still leave the model well constrained spatially. This design therefore does not fully test the model's ability to extrapolate into regions with little or no data.

The authors should add spatially blocked validation, preferably with blocks large enough to avoid local spatial autocorrelation. They could also report performance separately for high-coverage and low-coverage regions. This would align the validation more closely with the paper's central argument that spatial coverage is the dominant limitation.

Response: We agree that withholding years cannot establish spatial transferability because nearby and repeatedly sampled locations remain in the training set. We replaced the fixed held-out-year design with three complementary OOF schemes: Temporal OOF, Spatial OOF using $10^\circ \times 10^\circ$ blocks, and Joint spatiotemporal OOF. We also stratified Spatial OOF performance by historical coverage. Aggregate performance remains similar across the three schemes, with RMSE values of 21.22–22.35 μatm ; Spatial and Joint OOF yield RMSEs of 21.22 and 21.54 μatm , respectively.

Changes in the revised manuscript:

results, Sect. 4.1 (lines 400–428);

fixed-threshold coverage comparison, Sect. 4.2 (lines 429–450), we added:

“Spatial OOF samples were stratified according to observational density over 1982–2024 to examine how historical SOCAT coverage affects spatial generalization. Observation density was defined as the number of distinct observation years for each calendar-month $\times 1^\circ$ grid cell, normalized to a 0–1 range, and classified as low, intermediate, or high coverage using thresholds of 0–1/3, 1/3–2/3, and 2/3–1, respectively. To ensure a comparable spatial evaluation across coverage classes, five $10^\circ \times 10^\circ$ Spatial OOF blocks were randomly selected from each class. The same selected blocks were then used for the matched weighted and unweighted experiments, with weighting applied only during model training and all OOF metrics calculated from the original unweighted held-out samples.

Prediction error increases systematically as historical observational coverage decreases (Table 4). Under the weighted experiment, RMSE increases from 9.13 μatm in the high-coverage group to 11.40 μatm in the intermediate-coverage group and 13.58 μatm in the low-coverage group. This progression indicates that spatial generalization becomes increasingly difficult as historical observational support weakens,

even when the number of evaluated Spatial OOF blocks is held constant across coverage classes. Sample weighting has only a modest effect on prediction error, with small RMSE reductions in the intermediate- and low-coverage groups and little change in the high-coverage group. Relative to the unweighted experiment, weighting reduces RMSE by 0.12 μatm in the intermediate-coverage group and by 0.11 μatm in the low-coverage group, whereas the high-coverage group changes little (+0.02 μatm). This pattern is consistent with the intended role of the weighting scheme: to increase the influence of samples from poorly constrained conditions rather than to improve performance uniformly across all regions. The effect remains small compared with the much larger error increase associated with declining observational coverage, but it provides a measurable improvement where the reconstruction is most weakly constrained.

Table 4. Coverage-stratified Spatial OOF performance

Observational coverage	$w_{i,UT}$ RMSE (μatm)	No-weight RMSE (μatm)
High	9.13	9.11
Intermediate	11.40	11.52
Low	13.58	13.69

”

7. The claim that spatial sparsity, not algorithm choice, is the dominant limitation is too strong

The thinning experiments are interesting and support the importance of spatial coverage. However, they are conducted within a fixed model framework. They show that, for this model, reducing spatial coverage produces a larger error penalty than reducing temporal coverage under comparable sample-size conditions. They do not prove that algorithm choice is secondary in general, nor that spatial sparsity is more important than all methodological choices.

The authors should rephrase this conclusion more carefully. A more defensible statement would be that, within the FDU-BTR framework and the thinning experiments performed here, spatial representativeness has a stronger effect on independent-year RMSE than temporal continuity. Broader claims about algorithm choice would require controlled comparisons across substantially different algorithms under the same coverage perturbations.

Response: We appreciate this distinction. A thinning experiment with one model architecture cannot establish that algorithm choice is generally less important than observational coverage.

We therefore redesigned the thinning analysis and substantially narrowed the conclusion. The revised experiment is deliberately separated from the production BTR framework: a single global

CatBoost model is trained directly on SOCAT fCO₂ without the BTR residual formulation. Fifty matched temporal–spatial thinning pairs are constructed, with identical retained sample counts within each pair, and all realizations are evaluated against the same fixed external observations.

The revised conclusion is now limited to what the experiment directly demonstrates: both temporal and spatial loss of observational support degrade prediction, while spatially organized loss produces a larger average pooled RMSE penalty at matched sample sizes, particularly under sparse coverage. We no longer state that spatial sparsity is more important than algorithm choice or that it is universally the “dominant” limitation.

Changes in the revised manuscript: thinning methods and results, Sect. 6.2 (lines 619–669), Fig. 9; revised claims, Abstract (lines 8–21), Discussion (lines 754–809), and Conclusions (lines 810–829).

“The coverage-stratified Spatial OOF comparison showed higher prediction errors in the lower-coverage groups within the selected spatial blocks. This contrast suggests that regions with weaker historical observational support tend to present greater reconstruction difficulty, although observational density is unlikely to be the only factor involved. The prediction-interval analysis further revealed substantial spatial variation in predictive constraint. Together, these findings raise a more direct question: when the amount of training data is held comparable, does the temporal or spatial organization of observational support itself affect extrapolation skill? To address this question, we conducted controlled temporal–spatial thinning experiments in which training-sample size was matched between the two thinning strategies.

The experiments used unique SOCAT cell-month observations from 1985–2024. A single global CatBoost model was trained directly on surface-ocean fCO₂ using a fixed set of environmental and spatiotemporal predictors. To isolate the effect of observational coverage, model hyperparameters were optimized once using the FULL dataset and then held fixed across all thinning realizations. We generated 50 matched temporal–spatial thinning pairs. Temporal thinning removed complete calendar years, whereas spatial thinning removed complete spatial blocks throughout the record. For each temporal realization, a spatial realization with the same number of retained cell-month observations was constructed, allowing the effects of temporal and spatial coverage loss to be compared at matched training-sample sizes.

The distribution of observational support was characterized using Shannon effective coverage. For the temporal or spatial dimension d , Shannon entropy was calculated as

$$H_d = - \sum_{i=1}^{N_d} p_{d,i} \ln p_{d,i}, \quad (24)$$

and converted to effective coverage as

$$C_d^{\text{eff}} = \frac{\exp(H_d)}{N_d}, \quad (25)$$

where $p_{d,i}$ is the fraction of retained observations in bin i , and N_d is the total number of bins. Thus,

C_t^{eff} and C_s^{eff} describe the effective temporal and spatial extent of observational support, respectively. The total retained training coverage was quantified separately as

$$C_{\text{OCC}} = \frac{N_{\text{obs}}}{N_t N_s}, \quad (26)$$

where N_{obs} is the number of retained unique grid-cell–month observations, N_t is the number of monthly time steps, and N_s is the number of ocean grid cells within the analysis domain. Thus, $N_t N_s$ represents the total number of possible grid-cell–month combinations over the study period.

All realizations were evaluated against the same frozen external dataset comprising LDEO, GLODAP, HOT, BATS and SPOTS observations. Prediction skill was evaluated using pooled RMSE and a dataset-balanced macro-RMSE that assigned equal weight to the five external datasets.

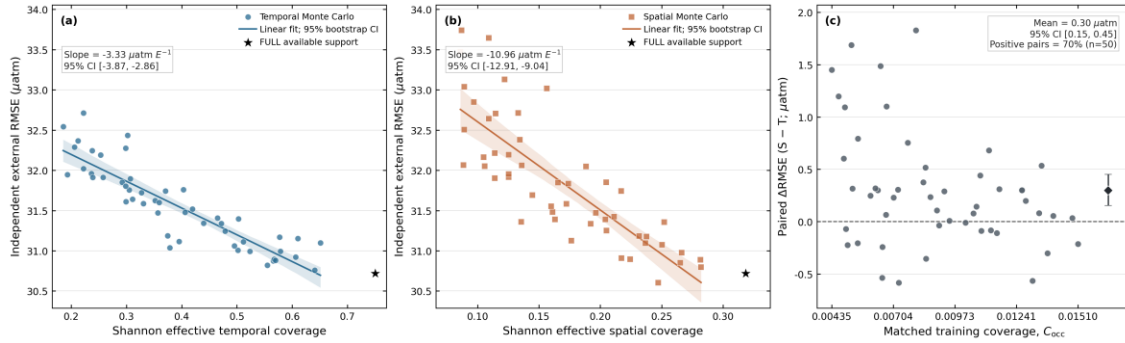


Figure 9. Effects of temporal and spatial observational thinning on external fCO_2 prediction skill. (a) External RMSE versus Shannon effective temporal coverage. (b) External RMSE versus Shannon effective spatial coverage. Solid lines and shaded bands show linear fits and 95% bootstrap confidence intervals; black stars denote the FULL model. (c) Paired RMSE differences between row-matched spatial and temporal thinning realizations as a function of matched training coverage C_{OCC} . Gray points show individual pairs, and the black line and gray band indicate the paired mean and its 95% bootstrap confidence interval.

Both temporal and spatial thinning show increasing external RMSE as effective coverage decreases (Fig. 9a, b). Spatial thinning also produces substantially greater variability in prediction error under low-coverage conditions. Realizations with similar effective spatial coverage can yield markedly different RMSE values, indicating that effective coverage alone does not uniquely determine predictive performance on external observations. This interpretation is consistent with Heimdal et al. (2024), who showed that adding a modest number of observations during Southern Hemisphere winter and along meridional gradients in the Southern Ocean substantially reduced reconstruction bias and RMSE, demonstrating that strategically placed observations can provide disproportionate gains in reconstruction skill.

The matched experiments provide a direct comparison between the two thinning strategies at equivalent training-sample sizes (Fig. 9c). Spatially organized observational loss produces a larger average pooled RMSE penalty than temporally organized loss, with a mean paired RMSE difference of 0.30 μatm (95% CI: 0.15–0.45 μatm). This difference is largest under sparse training coverage and decreases as observational support increases. By contrast, the corresponding difference in dataset-balanced macro-RMSE is much smaller, indicating that the effect of spatial thinning is not uniform across the external

observational datasets.

These thinning experiments therefore show that prediction skill depends on both the amount and the spatiotemporal distribution of observational support. At matched sample sizes, temporal and spatial reductions in coverage do not produce equivalent prediction errors, with spatially organized loss imposing the larger average penalty, particularly under sparse coverage.

”

8. Boundary blending may improve numerical continuity but could also smooth real gradients

The boundary blending scheme is a reasonable solution to discontinuities between regional models. However, the evaluation focuses on reduced differences across regional boundaries. Reduced discontinuity is not necessarily equivalent to improved physical realism, because real gradients may occur across fronts, water-mass boundaries, and biogeochemical regimes. The authors acknowledge this point, but the analysis would be stronger if they evaluated whether boundary blending improves agreement with observations near boundaries, rather than only reducing cross-boundary contrast. It would also be useful to assess whether boundary blending damps real variability or trends in frontal regions.

Response: We agree that reduced numerical discontinuity does not establish greater physical realism and that blending final predictions could smooth real fronts.

We therefore changed the regional-boundary treatment. The revised framework no longer blends final predictions across biome boundaries. Instead, each biome retains its own core prediction domain, while a directional ocean-connectivity training buffer of up to 5° allows nearby observations from an adjacent biome to contribute to model fitting near a shared boundary. Thus, neighboring observations strengthen the training constraint, but the final prediction is not obtained by averaging two regional-model outputs across the boundary.

The revised procedure no longer averages final predictions across biome boundaries. This removes blending-induced smoothing, although the training-buffer approach does not by itself prove that all true boundary gradients are resolved. The directional training buffer instead strengthens observational constraints near biome boundaries by allowing nearby observations from adjacent ocean-connected regions to contribute to model fitting. A related ocean pCO₂ precedent added observations from a boundary area to the training sets of adjacent provinces while leaving validation assignments and prediction domains unchanged; this precedent does not directly validate the directional ocean-connectivity buffer used here (Zhong et al., 2022).

Changes in the revised manuscript: biome partitioning and directional training buffer, Sect. 3.2.2 (lines 253–266), Fig. 2; product assembly without prediction blending, Sect. 3.5 (lines 381–398).

“For each biome, an independent residual-learning model was trained. When regional models are fitted separately, the training samples can change abruptly across biome boundaries, even though the

underlying ocean environment is generally more continuous between adjacent regions. To strengthen the training constraint near these boundaries, a directional ocean-connectivity training buffer was constructed around each biome core region. A training buffer with a maximum width of 5° was used, allowing the target biome model to incorporate observations from directly adjacent biomes that lie close to the shared boundary. Core samples from the target region were always retained, and the final training set for each regional model therefore consisted of its core samples together with neighboring samples that satisfied the buffer criterion (Zhong et al., 2022). After regional partitioning, boundary buffering, data matching, and quality control, 364,701 valid grid-cell-month samples were retained for training the regional residual models.

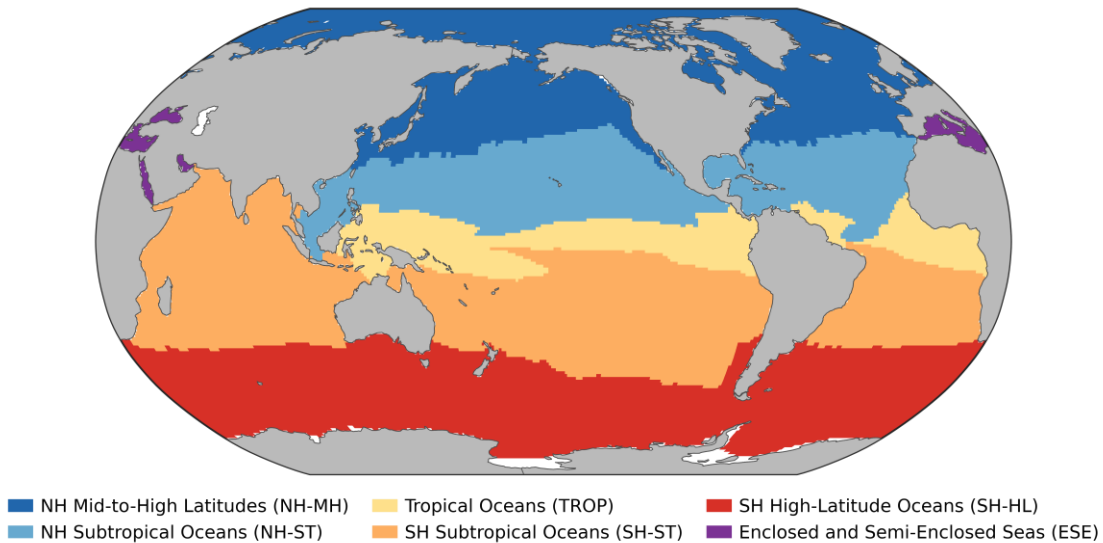


Figure 2. Global distribution of the six large-scale ocean biome regions used in this study: Northern Hemisphere mid-to-high latitudes (NH-MH), Northern Hemisphere subtropical oceans (NH-ST), tropical oceans (TROP), Southern Hemisphere subtropical oceans (SH-ST), Southern Hemisphere high-latitude oceans (SH-HL), and enclosed and semi-enclosed seas (ESE). ”

9. Claims about coastal and marginal regions should be more cautious

The paper identifies coastal, shelf, marginal-sea, coral-reef, and high-latitude regions as areas of elevated uncertainty, and the buoy validation indicates poorer performance in coastal and reef environments. This is not surprising for a 1° monthly product, but it should be reflected consistently in the abstract, conclusions, and data-use recommendations. The product may be appropriate for basin-scale and open-ocean analyses, but users should be warned more explicitly against interpreting local coastal or reef-scale variability.

Response: We agree with the reviewer and therefore made this limitation explicit throughout the revised manuscript. The uncertainty analysis now shows substantially wider prediction intervals in shallow waters (≤ 200 m), where the PI90 half-width is approximately 39–56 μatm , compared with approximately 18–21 μatm in deeper waters (> 200 m). The Discussion links this contrast to both weaker observational constraints and the scale mismatch between a global 1° monthly product and high-frequency local processes. Explicit data-use guidance—that the product is intended for basin-

scale and open-ocean analyses and should not be used to interpret local coastal or reef variability—now appears in the Abstract, Conclusions, and Data Availability sections.

Changes in the revised manuscript: shallow/deep interval contrast, Sect. 6.1 (lines 583–618), Fig. 8d; scale limitations, Sect. 8 (lines 779–791); and Data availability (lines 830–848).

“Predictive uncertainty shows an even clearer contrast between shallow and deeper waters (Fig. 8d). Using 200 m water depth as the threshold, the PI90 half-width in shallow-water regions is mostly approximately 39–56 μatm , whereas values in deeper-water regions are concentrated at approximately 18–21 μatm . PI67 shows a consistent depth-dependent contrast. Uncertainty in shallow-water regions also exhibits stronger month-to-month variability, whereas uncertainty in deeper-water regions is generally lower and more stable. Here, the PI half-width is defined as half the difference between the upper and lower prediction-interval bounds.”

“The prediction-interval analysis further shows that reconstruction reliability is strongly dependent on spatial scale and region. Prediction-interval width increases with external prediction error and therefore provides useful information on relative predictive risk. However, the observed coverage of the nominal 90% interval remains below 90% for fully external observations, indicating that these intervals should not be interpreted as universally exact probabilistic bounds across all ocean environments. Their main utility lies in identifying relative differences in predictive constraint. This distinction is particularly important in shallow and coastal waters, where prediction intervals are substantially wider and more variable than in the deep open ocean. These regions are influenced by riverine inputs, shelf exchange, coastal upwelling, tides, and high-frequency biological variability, many of which occur at spatial and temporal scales finer than the monthly 1° resolution of FDU-BTR. The elevated uncertainty therefore reflects both weaker observational support and a mismatch between the resolution of a global product and the scales of local processes. FDU-BTR is consequently better suited to global, basin-scale, and open-ocean analyses than to resolving local variability in estuaries, coral reefs, narrow coastal zones, or individual events. Improved representation of these environments will likely require denser local observations and reconstruction frameworks that explicitly accommodate multiple spatial and temporal scales.”

“FDU-BTR is primarily intended for global, basin-scale, and open-ocean analyses of monthly-to-interannual surface-ocean fCO_2 variability, long-term change, product intercomparison, and air–sea CO_2 fluxes. Predictive uncertainty is substantially greater in shallow and coastal waters and in other spatially heterogeneous environments. The product should therefore not be used to infer reef-scale, estuarine, narrow-coastal, individual-event, or submonthly variability without additional local observational constraints. Users are encouraged to consider the accompanying 67% and 90% prediction intervals when evaluating regional applications.”

Minor comments:

L64: Please define SOCOM at first use, and other acronyms too.

Response: We have reviewed acronyms throughout the revised manuscript. The unnecessary use of SOCOM in the main text has been removed; acronyms retained for the methods and datasets are expanded on first use.

L80: Since BTR decomposes observed pCO₂ into a background climatology, thermal-anomaly term, and residual term, the authors should validate or at least diagnose each of these components. It is not sufficient to evaluate only the final reconstructed pCO₂ field if the decomposition is presented as a key physical advance.

Response: We agree; Table 5 and Fig. 4 now assess the background, background plus thermal contribution, and full reconstruction on the same Spatial OOF samples. Our response to Major Comment 3 explains the interpretation and its limits.

Changes in the revised manuscript: Sect. 4.3 (lines 451–487), Table 5, and Fig. 4.

L91: The statement that the study “characterizes the spatial patterns and dominant sources of product uncertainty” is too strong given the current validation. The uncertainty analysis is useful, but the independent validation is still limited, especially for spatial extrapolation. This sentence should be softened unless additional independent spatial validation is added.

Response: We agree that the original statement about “dominant sources” of total product uncertainty exceeded the evidence. We removed that claim and now distinguish nominal prediction intervals from demonstrated external coverage. The revised discussion emphasizes interval width as a measure of relative prediction risk and reports the observed undercoverage.

Changes in the revised manuscript: Abstract (lines 8–21), Sect. 6.1 (lines 583–618), and Discussion (lines 754–809).

L91–93: The statement that the product provides a foundation for air–sea CO₂ flux estimation should be supported by validation of Δ pCO₂ or air–sea CO₂ fluxes. Otherwise, the authors should say that the product is potentially useful for flux estimation, pending flux-specific evaluation.

Response: We agree that pointwise fCO₂ validation does not, on its own, validate inferred air–sea fluxes. We added a common Δ fCO₂ and flux calculation and compare its global and regional evolution across products. Because this is not a direct observational flux benchmark, the revised text does not claim that the flux estimates are independently validated.

Changes in the revised manuscript: Sect. 7 (lines 670–753), Fig. 10, and Figs. S4–S5.

“”

Specific to methods:

Please clarify how missing predictors are handled in different periods, especially for variables that do not cover the full 1982–2024 period (Several predictors do not cover the full 1982–2024 period: Chl-a starts in 1997, SLA in 1993, MLD ends in 2022, pH/DIC start in 1985, etc.). CatBoost can handle missing values, but long periods of missingness may still encode time or product-availability information.

Response: We agree that long gaps in predictor coverage require explicit treatment. For SLA and Chl-a, unavailable years are filled with the corresponding grid-cell calendar-month climatology; MLD is represented by its monthly climatological field. Other missing predictor values remain missing and are handled by CatBoost during fitting. The use of climatological fields to extend predictors with incomplete temporal coverage is consistent with previous global surface-ocean CO₂ reconstruction approaches; for example, Zhang et al. (2025) used a climatological Chl-a field outside the period of satellite coverage and climatological MLD as model inputs.

Changes in the revised manuscript: feature construction and climatological filling, Sect. 3.2.1 (lines 220–252), we added:

“.....where $X(g, t)$ is the value at grid cell g and time t , and $X_{\text{clim}}(g, m)$ is the climatological value for the corresponding calendar month m at that grid cell. Using this formulation, anomaly features were constructed for SST, SSS, 10 m wind speed, sea-ice fraction, and mean sea level pressure based on the 1982–2024 climatology. For SLA and chlorophyll concentration (Chl-a), which have incomplete temporal coverage, missing years were filled using the multi-year climatology for the corresponding calendar month at each grid cell to preserve the basic spatial pattern and seasonal cycle. Chl-a was subsequently log-transformed to reduce the influence of its highly skewed distribution on model fitting. For mixed layer depth, the climatological value MLD_{clim} was used to represent the stable background state of upper-ocean mixing.

”

Because the product is intended for long-term variability and trend analyses, the authors should evaluate trend robustness explicitly. Are regional pCO₂ trends sensitive to the background climatology, xCO₂, pH/DIC predictors, and the choice of training years? Please compare trends against independent time series where possible and report trend uncertainty, not only pointwise RMSE.

Response: We agree that long-term applications require evaluation beyond pointwise prediction error. We therefore added trend comparisons based on annual deseasonalized anomalies at HOT and BATS, with bootstrap uncertainty intervals (Fig. 7), together with a regional inter-product trend comparison (Fig. S3). FDU-BTR reproduces the observed temporal evolution at both stations, with monthly Pearson correlation coefficients of $r = 0.93$ at HOT and $r = 0.96$ at BATS, and the reconstructed long-term trends remain statistically consistent with the observational estimates. Regional trends also generally fall within the spread of existing products across most major ocean basins.

We also examined the role of atmospheric $x\text{CO}_2$ and found it to be an important predictor of long-term temporal evolution; removing $x\text{CO}_2$ produced substantial deviations in the earlier part of the reconstruction. This sensitivity test is not included in the main manuscript because of space limitations, but we report the result here in response to the reviewer.

Changes in the revised manuscript: station trends and bootstrap uncertainty, Sect. 5.3 (lines 539–560), Fig. 7; regional inter-product trend comparison, Fig. S3.

“This long-term evolution is broadly consistent with existing surface-ocean CO_2 products (Fig. S3a). FDU-BTR remains within the main inter-product range through most of the record and reproduces the common transition from negative anomalies in the 1980s to increasingly positive anomalies in recent decades, despite differences in magnitude and shorter-term variability among products.

Regional trends are also broadly consistent across most major ocean basins (Fig. S3b). FDU-BTR generally falls within the external-product range for the North and South Pacific, North and South Atlantic, Indian Ocean, and Southern Ocean, whereas larger inter-product differences occur in the Arctic Ocean and ESE region. Thus, agreement is strongest for the broad long-term change in the open ocean, while greater product dependence remains in more weakly constrained regions.

Overall, FDU-BTR resolves two defining features of global surface-ocean $f\text{CO}_2$ variability: strong and persistent spatial contrasts among oceanic regimes and a coherent long-term upward shift in the global mean state. The agreement in large-scale temporal evolution with existing products provides additional context for these reconstructed patterns, while the regional differences identified in Fig. S3 indicate where product estimates remain less consistent. These characteristics provide the baseline for the subsequent external observational validation and time-series evaluation.

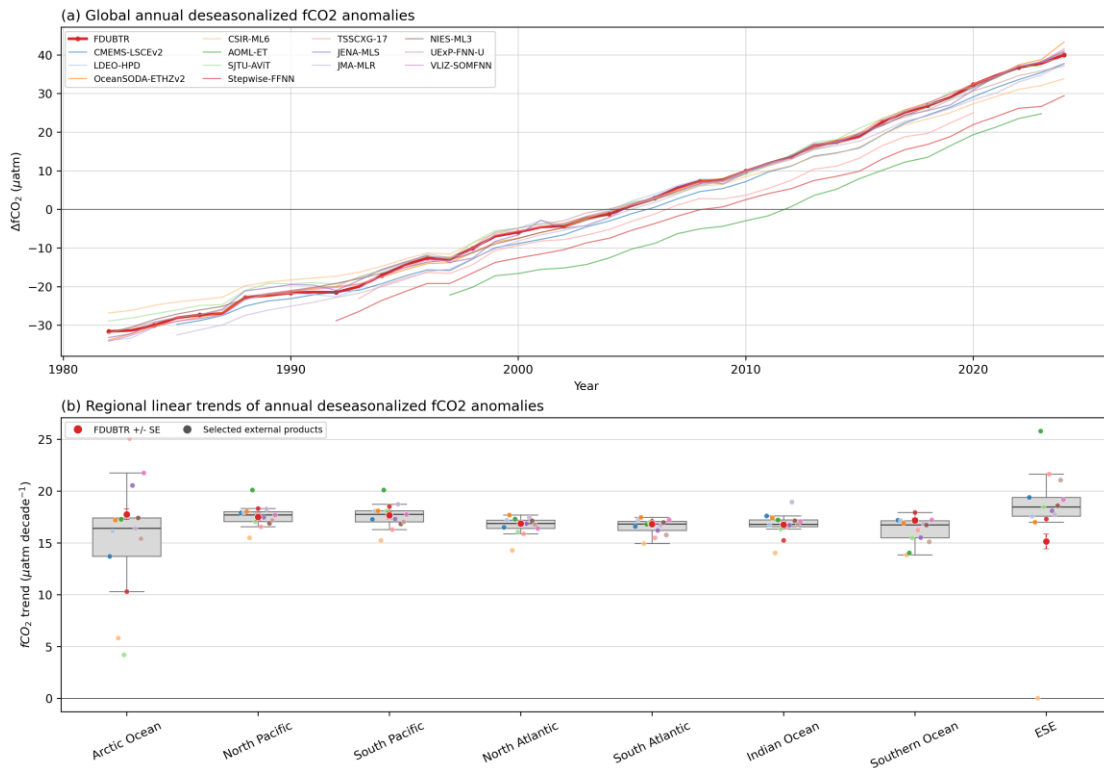


Figure S3. Global evolution and regional trends in deseasonalized surface-ocean fCO₂ anomalies from FDU-BTR and external products (1982–2024). (a) Area-weighted global annual anomalies after removal of each product’s monthly climatology. (b) Ordinary least-squares trends of annual deseasonalized anomalies across eight ocean regions. Dots represent individual external products, and boxplots summarize their distribution. The red marker shows the FDU-BTR trend, with the error bar indicating ± 1 standard error of the fitted slope.

”

Please justify the $p\text{CO}_2 > 700 \mu\text{atm}$ filtering threshold. This may remove extreme but real coastal, upwelling, reef, or marginal-sea values. If the product excludes or smooths such environments, this should be stated clearly. Also, please quantify sensitivity to the fCO₂ to pCO₂ conversion, the 0.0423 K⁻¹ temperature correction, and the 40 μatm correction filter. These choices may disproportionately affect high-latitude, coastal, upwelling, or reef observations and could influence both the training target and validation metrics.

Response: We thank the reviewer for highlighting how the original filters and transformations could affect physically meaningful extremes, especially in heterogeneous coastal and marginal waters.

We therefore removed the 700 μatm upper-threshold filter, the preprocessing conversion of SOCAT fCO₂ to pCO₂, and the associated 40 μatm temperature-correction-difference filter. The SOCAT fCO₂ observations are now used directly as the training target.

We emphasize that the coefficient 0.0423 K^{-1} remains in the revised manuscript only as the explicitly defined first-order thermal component of the BTR framework; it is no longer used as a preprocessing filter or correction applied to the SOCAT target. Its physical basis and relevant references are now stated explicitly.

Changes in the revised manuscript: direct SOCAT $f\text{CO}_2$ target, Sect. 2.1 (lines 74–84); first-order thermal component, Sect. 3.1.2 (lines 181–199).

Because the background climatology is constructed from existing $p\text{CO}_2/f\text{CO}_2$ products, the comparison of FDU-BTR with those products is not fully independent. The authors should discuss this dependence and avoid presenting inter-product agreement as completely independent validation.

Response: We agree that agreement with products contributing to the climatological background cannot be treated as independent validation of the complete framework.

The six selected external surface-ocean CO_2 products are detrended and summarized as fixed calendar-month climatologies before their median is used to define the background. This prior encodes broad spatial and seasonal structure, while the residual model learns departures from the prescribed background and thermal term using SOCAT and environmental predictors. The background therefore does not inject year-specific anomalies from held-out years, but it does retain SOCAT-derived information and is not independent of the products contributing to it.

Changes in the revised manuscript: construction of the six-product background, Sect. 3.1.1 (lines 160–181); interpretation of inter-product comparison, Sect. 5.4 (lines 560–582); prior dependence, Sect. 8 (lines 754–809).

Please provide validation metrics stratified by basin, latitude band, season, open-ocean/coastal class, and observational coverage density. This would make the validation much more relevant to the spatial gap-filling problem.

Response: We agree that a single global score can conceal strong regional and seasonal differences in predictive performance.

We added regional and seasonal OOF diagnostics (Fig. S1) and fixed-threshold observational-coverage strata (Table 4). Fig. 8d separately compares prediction-interval widths between shallow and deeper waters; this uncertainty diagnostic should not be mistaken for a direct coastal skill test. The revised text therefore limits claims about local coastal performance.

Changes in the revised manuscript: Sect. 4.1 (lines 400–428), Fig. S1; Sect. 4.2 (lines 429–450), Table 4; Sect. 6.1 (lines 583–618), Fig. 8d.

Please include a spatial-block or leave-region-out validation experiment. The current held-out-year validation mainly tests temporal generalization.

Response: We agree and have added Spatial OOF based on $10^\circ \times 10^\circ$ blocks, alongside Joint spatiotemporal OOF.

Spatial OOF now uses $10^\circ \times 10^\circ$ spatial blocks assigned to five outer folds, with all test blocks removed from the corresponding model fitting and model-selection procedures.

Changes in the revised manuscript: $10^\circ \times 10^\circ$ Spatial OOF and Joint OOF, Sect. 3.3.4 (lines 325–344); results, Sect. 4.1 (lines 400–428), Fig. 3.

Please test uncertainty calibration against independent residuals. For example, do high-uncertainty grid cells actually have larger validation errors?

Response: We agree that prediction-interval width is useful only if its relationship with realized error and its empirical coverage are evaluated.

We therefore added an external observational calibration analysis. Prediction-interval width is compared directly with absolute error, and the empirical coverage of the nominal 90% interval is reported. The external analysis shows that wider intervals are associated with larger errors but that nominal 90% coverage is not fully achieved. We state this limitation explicitly.

Changes in the revised manuscript: Sect. 6.1 (lines 583–618), Fig. 8c. External undercoverage is stated explicitly rather than treated as successful nominal calibration.

The conclusion that spatial observational sparsity, rather than algorithm choice, is the dominant remaining limit should be softened. The thinning experiments support the importance of spatial coverage within this framework, but do not isolate algorithm choice across reconstruction methods.

Response: We agree that the thinning experiment cannot establish a general ranking of observational coverage versus algorithm choice. We removed that comparison from the conclusions and now state only the difference observed between the two matched thinning strategies.

The manuscript now states only that, under the controlled matched thinning experiment, spatially organized loss of observational support produces a larger average pooled RMSE penalty than temporally organized loss at the same retained sample size.

Changes in the revised manuscript: Sect. 6.2 (lines 619–669), Discussion (lines 754–809), and Conclusions (lines 810–829).

Overall, I find the dataset and framework potentially useful, but the manuscript should better align its claims with the validation performed. The strongest contribution is the physically guided residual-learning framework combined with a careful diagnosis of coverage limitations.

Also, please ensure that the archived dataset contains not only the final pCO₂ field but also the uncertainty, and preferably the background, thermal, and residual components.

Response: We agree that releasing prediction intervals and component fields is important for transparent interpretation and reuse.

The revised product description specifies the central fCO₂ field, the climatological background, explicit thermal contribution, learned residual correction, ensemble spread, and nominal 67% and 90% prediction-interval bounds. We also state the intended scale of application and the undercoverage observed against external observations.

Changes in the revised manuscript: product generation, Sect. 3.5 (lines 381–398); Data availability, Sect. 10 (lines 830–848).

“3.5 Product generation

After training the residual models for each biome, the corresponding ensemble models were applied to the monthly predictor fields over 1982–2024 to generate gridded residual corrections. Each ocean grid cell was predicted using the model associated with its biome and was then recombined with the multi-product monthly climatological background and the explicit thermal component. The final reconstruction therefore consists of the prescribed background state, the thermal adjustment, and the ensemble-mean residual prediction.

Based on this workflow, we release FDU-BTR, a global monthly surface-ocean fCO₂ product covering January 1982 to December 2024 (Wang and Fu, 2026, <https://doi.org/10.5281/zenodo.22994847>). The product contains 516 monthly fields on a regular 1° × 1° latitude–longitude grid, with monthly values referenced to the middle of each month. It follows CF-1.8 and ACDD-1.3 metadata conventions and is distributed under a CC BY 4.0 license.

In addition to the central fCO₂ reconstruction, the product retains the three BTR components separately: the multi-product climatological background, the explicit thermal contribution, and the machine-learned residual correction. This allows the reconstructed field to be traced to the individual stages of the BTR framework rather than being provided only as a final combined estimate. The product also includes the spread among central-model ensemble members and calibrated 67% and 90% conditional prediction intervals. The ensemble spread characterizes disagreement among the central-model members, whereas the calibrated prediction intervals provide the uncertainty ranges associated with the released reconstruction. After generation of the monthly 1° fields, no observation-based spatial interpolation,

post-reconstruction gap filling, or additional spatial smoothing was applied, so the released fields retain the direct gridded output of the reconstruction framework.

”

Reviewer #2:

Major comments:

I note that I was asked to formally review this paper after submitting the following comment on 26 June. This comment was echoed by the first formal reviewer.

The authors use NCEI buoy data (Sutton et al. 2018) to validate their method (Figure 6 in section "5.1 Independent time-series validation"). There is an error here because these data are not actually independent; they are included in SOCAT. The authors need to change this section to reflect the fact that the buoy data are not independent of SOCAT, their training set.

Response: We thank the reviewer for identifying this error and for the earlier comment of 26 June; the NCEI mooring measurements are part of SOCAT and should not have been presented as independent validation.

We have removed that comparison. Sect. 5.2 now presents the broader LDEO and GLODAP observation comparisons; Sect. 5.3 reserves HOT and BATS for targeted time-series evaluation. SPOTS also contributes to the five-source product and uncertainty comparisons, subject to the source-specific screening described in Sect. 2.4.

I concur with the first reviewer that the central problem with the current MS is that it overstates the generalizability of the method given that the only independent comparison made is to BATS and HOT timeseries. As shown in McKinley et al. 2026 (ERL), BATS and HOT are relatively easy to represent with this kind of algorithm. There is no evidence provided that the approach proposed here generalizes globally with skill.

Response: We agree that strong performance at HOT and BATS—both well-observed open-ocean sites—cannot establish global spatial generalization, and we note the point of McKinley et al. (2026) that such sites are comparatively easy to reproduce. We now use HOT/BATS only to assess temporal variability and long-term trends at two open-ocean stations, and we evaluate broader performance with LDEO and GLODAP together with Spatial and Joint spatiotemporal OOF. Language implying global validation from HOT/BATS has been removed throughout.

Changes in the revised manuscript: external observations, Sect. 5.2 (lines 519–538); station-specific scope, Sect. 5.3 (lines 539–560); limits of extrapolation, Sect. 8 (lines 754–809).

“5.2 External observational validation

FDU-BTR was evaluated against external LDEO and GLODAP observations to assess its performance beyond the SOCAT training samples (Fig. 6). For LDEO, FDU-BTR achieved an RMSE of 28.89 μatm , an MAE of 17.47 μatm , a bias of 0.11 μatm , and an R^2 of 0.67 across 16,072 matched observations. Against GLODAP, the corresponding values were 30.75 μatm , 17.78 μatm , -2.67 μatm , and 0.65 across 10,204 matched observations. The LDEO comparison shows little systematic offset, whereas the GLODAP comparison exhibits a modest negative bias together with higher RMSE and lower R^2 , indicating larger discrepancies for the conditions represented by that dataset.

For both datasets, most observations are concentrated near the 1:1 relationship over the central $f\text{CO}_2$ range, while larger departures occur at the low and high ends of the observed distribution. Reconstructed values tend to be compressed toward the central range under these more extreme conditions, indicating that anomalously low or high $f\text{CO}_2$ values are less fully reproduced than the bulk of the observations. This error structure suggests that the main limitation is associated more with the amplitude of extreme or locally anomalous variability than with a uniform systematic bias.

These comparisons show that FDU-BTR reproduces the central range of externally observed surface-ocean $f\text{CO}_2$ with broadly consistent performance across the two observational datasets, while reconstruction errors increase under more extreme conditions. The difference in performance between LDEO and GLODAP further indicates that predictive skill depends on the oceanographic conditions and sampling characteristics represented by the validation data rather than being spatially uniform across the global ocean.

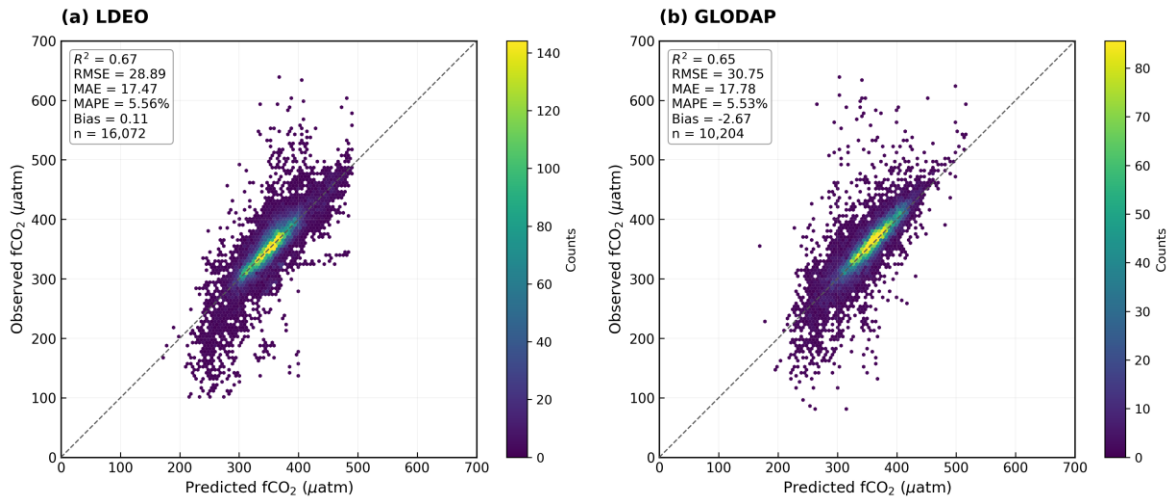


Figure 6. Comparison of FDU-BTR reconstructed $f\text{CO}_2$ with external LDEO and GLODAP observations.

5.3 Time-series evaluation

Evaluation at the HOT and BATS time-series stations provides a site-specific assessment of how well FDU-BTR represents surface-ocean $f\text{CO}_2$ variability across different temporal scales (Fig. 7). At both

sites, FDU-BTR reproduces the pronounced seasonal cycle and sustained long-term increase observed over the measurement period, with generally close agreement at the monthly scale. Some short-term excursions and periods of enhanced variability are less fully resolved, indicating that reconstruction fidelity varies across temporal scales.

This scale dependence becomes clearer after removal of the seasonal cycle. FDU-BTR retains the broad interannual evolution and long-term increase at both stations, but smooths the amplitude of some individual year-to-year anomalies. Despite these differences in short-term variability, the reconstructed long-term trends remain statistically consistent with the observational estimates. Thus, the agreement in long-term evolution persists even where individual interannual excursions are not fully reproduced.

Overall, the HOT and BATS records show that, at these two well-observed open-ocean sites, FDU-BTR captures seasonal structure and long-term change more consistently than the amplitude of individual interannual anomalies. These evaluations therefore provide targeted evidence for the representation of temporal variability at long-term open-ocean stations, rather than a general assessment of reconstruction performance across the global ocean.

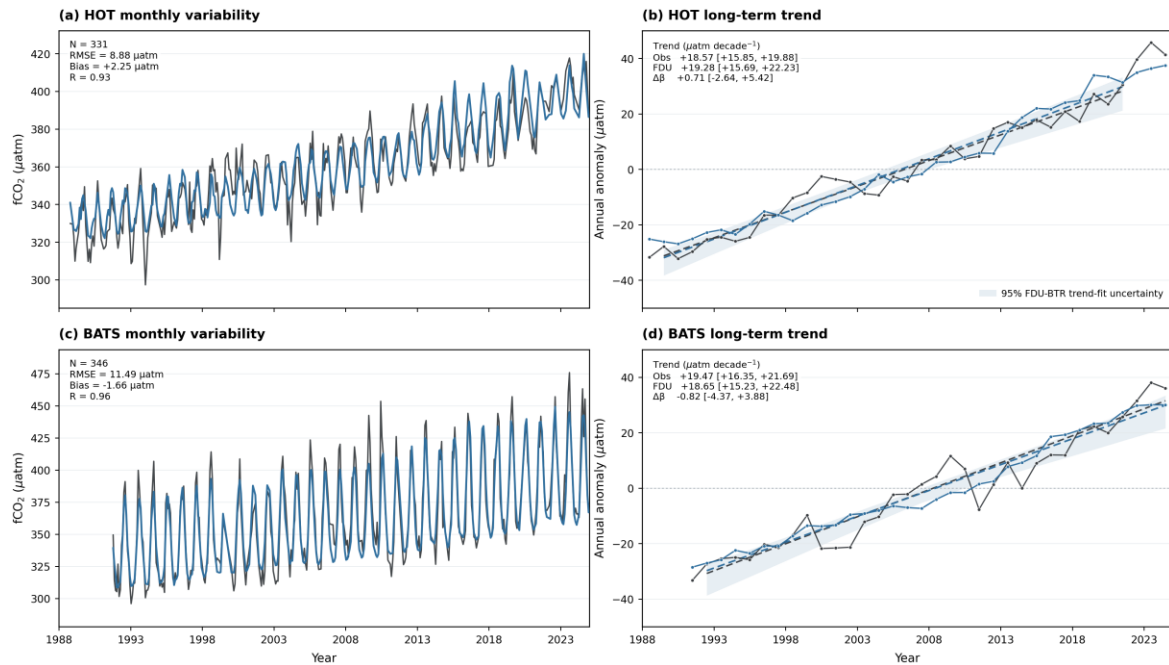


Figure 7. Validation of monthly variability and long-term trends in surface-ocean fCO_2 at the HOT and BATS time-series stations. (a, c) Comparison of monthly FDU-BTR fCO_2 estimates with station observations at HOT and BATS, respectively. (b, d) Annual deseasonalized fCO_2 anomalies and fitted long-term trends at HOT and BATS, respectively. The shaded band represents the pointwise 95% bootstrap interval of the fitted FDU-BTR trend, estimated from 2,000 paired circular moving-block resamples of annual deseasonalized anomalies using five-year blocks.

”

The authors should directly compare both this approach and those released by the Global Carbon Budget (<https://zenodo.org/records/17579108>) to the same independent data (as, for example, in Bennington et al. 2022) – beyond just BATS and HOT. This would allow the community to understand the extent to which this proposed approach is an advance, as the authors claim it to be.

Response: We agree that common-sample evaluation is more informative than comparisons based on different observation sets. FDU-BTR and 13 comparison products are therefore evaluated on the same product-common subset of five external observational datasets. Table 6 reports pooled metrics, and Table S1 provides dataset-specific metrics, permitting like-for-like comparison. On this common subset, FDU-BTR achieves RMSE = 28.21 μatm , MAE = 16.64 μatm , bias = $-0.04 \mu\text{atm}$, and $R^2 = 0.68$. Its RMSE is close to those of OceanSODA-ETHZv2 (27.66 μatm), CMEMS-LSCEv2 (28.01 μatm), and Stepwise-FFNN (28.04 μatm), indicating performance comparable to several existing reconstructions rather than a uniquely superior result. We note that Fig. 6 uses a larger FDU-BTR-only LDEO/GLODAP match and therefore represents a different evaluation sample. For the carbon-sink application, all surface-ocean products are processed using a common atmospheric CO_2 field and gas-exchange framework; the GCB2025 comparison concerns sink magnitude and temporal evolution rather than prediction accuracy against independent flux observations.

Changes in the revised manuscript: Sect. 5.4 (lines 560–582), Table 6, and Table S1; common flux framework and GCB2025 comparison, Sect. 7 (lines 670–753), Figs. 10–11.

“The global mean ΔfCO_2 and air–sea CO_2 flux derived from FDU-BTR both show an overall strengthening during 1982–2024, superimposed on pronounced interannual and decadal variability (Fig. 10). The external surface-ocean CO_2 products show broadly similar long-term evolution, including several common low-frequency peaks and troughs, although substantial differences remain in the absolute magnitude of the estimated sink. Agreement among products is therefore stronger in the direction and timing of long-term change than in the magnitude of ocean CO_2 uptake.

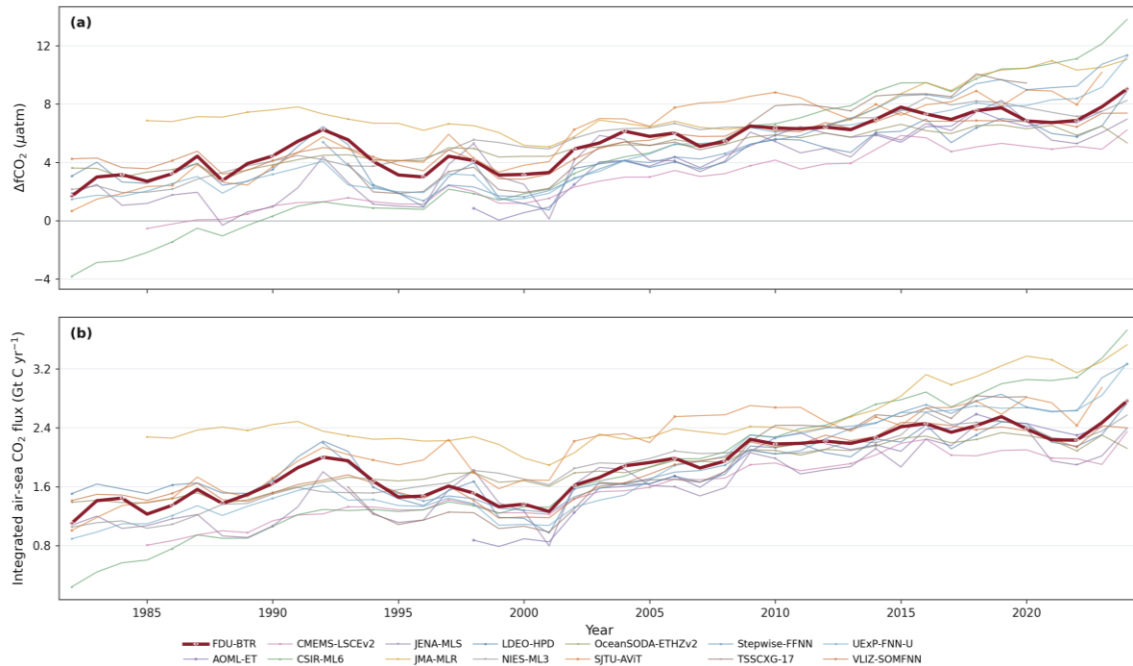


Figure 10. Global annual $\Delta f\text{CO}_2$ and air–sea CO_2 flux from FDU-BTR and 13 external products (1982–2024).

Whether this global strengthening reflects synchronous changes across ocean regions was then examined using the regional flux estimates (Figs. S4 and S5). The regional results show that it does not. The Arctic, North Atlantic, North Pacific, and Indian Ocean generally exhibit strengthening CO_2 uptake over the study period, whereas the South Atlantic evolves from relatively weak toward stronger uptake. The South Pacific remains close to long-term source–sink balance, with interannual variability large relative to its mean exchange. The Southern Ocean exhibits pronounced decadal variability, including weakening from the late 1990s to the early 2000s followed by subsequent recovery and strengthening. Enclosed and semi-enclosed seas (ESE) show the largest inter-product spread, including periods in which products differ in the inferred sign of net air–sea CO_2 exchange. FDU-BTR indicates a long-term transition in ESE from net CO_2 release toward net uptake, whereas most external products do not show a comparable transition. A comparable source-to-sink transition has previously been reported for the Mediterranean Sea, which shifted from a net CO_2 source in the 1960s to a net sink in the 1990s (Louanchi et al., 2009). Thus, the long-term strengthening of the globally integrated ocean carbon sink emerges from distinct regional trajectories rather than a synchronous increase in CO_2 uptake across all ocean basins.

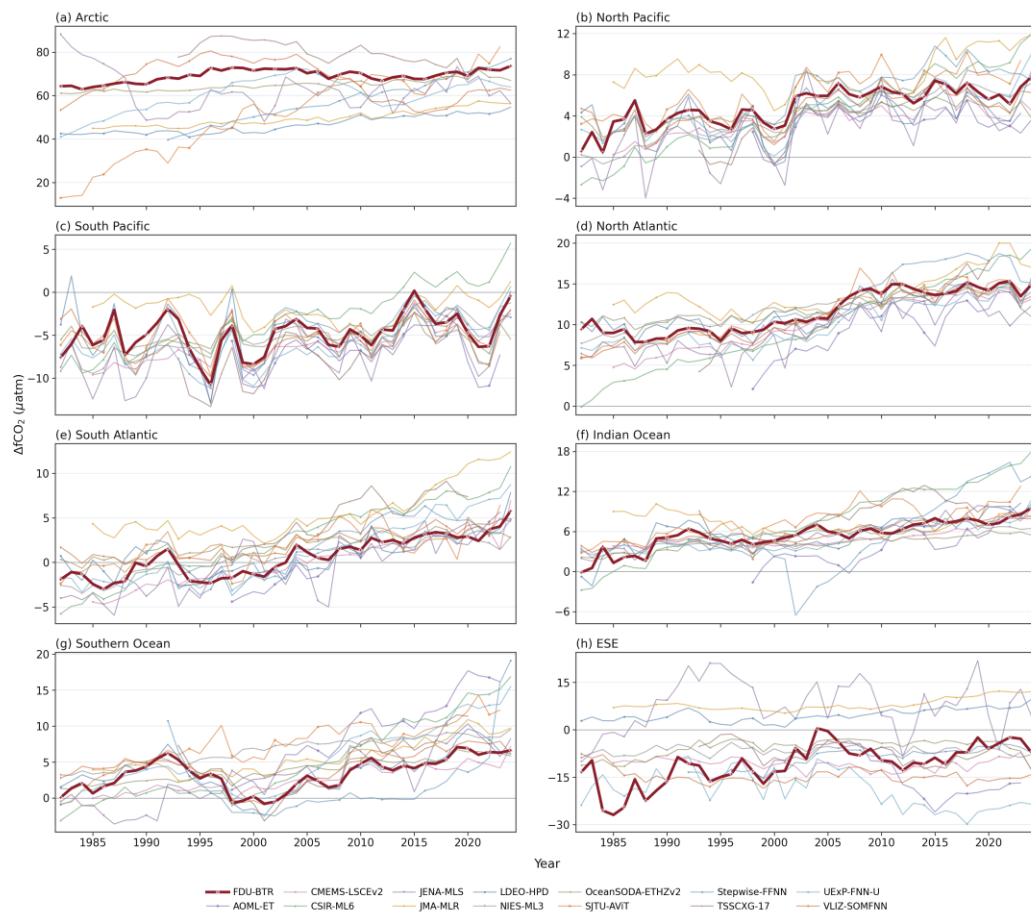


Figure S4. Regional annual $\Delta f\text{CO}_2$ across eight ocean regions from FDU-BTR and 13 external products (1982–2024). Positive values indicate CO_2 uptake by the ocean.

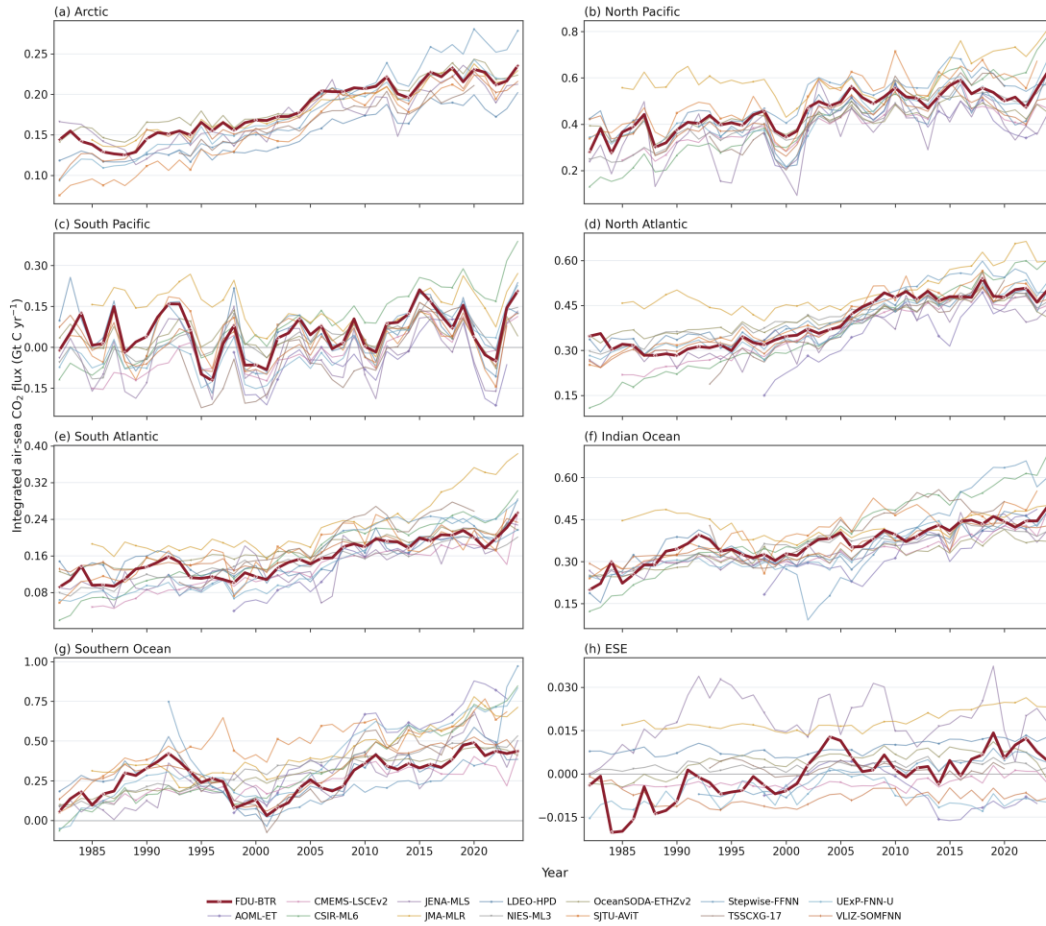


Figure S5. Regional annual air–sea CO₂ flux across eight ocean regions from FDU-BTR and 13 external products (1982–2024). Positive values indicate CO₂ uptake by the ocean.

The position of the FDU-BTR ocean carbon sink within the broader GCB2025 framework was further examined by comparing its temporal evolution with the GCB2025 mean of the surface-ocean fCO₂ data products (Fig. 11). FDU-BTR reproduces the major low-frequency features evident in the data-product mean, including strengthening in the early 1990s, subsequent weakening toward the late 1990s and early 2000s, and a more sustained increase after the early 2000s. Although the two estimates differ in absolute magnitude, particularly after 2000 when FDU-BTR is generally slightly lower than the data-product mean, they show broadly consistent long-term evolution. The GOBMs (global ocean biogeochemistry models) mean remains systematically lower over much of the later period, resulting in a persistent separation between the observation-based and model-based estimates of ocean CO₂ uptake.

After applying the global riverine correction of +0.65 Gt C yr⁻¹, the FDU-BTR ocean carbon sink is 2.20, 2.17, 2.75, and 3.08 Gt C yr⁻¹ for 1982–1994, 1995–2004, 2005–2014, and 2015–2024,

respectively (Fig. 11). For comparison with the GCB2025 merged ocean-sink estimate, an additional global warm-layer/cool-skin adjustment of $+0.18 \text{ Gt C yr}^{-1}$ was applied to the FDU-BTR period means (Ford et al., 2024). The corresponding adjusted estimates are 2.38, 2.35, 2.93, and $3.26 \text{ Gt C yr}^{-1}$, respectively. The adjusted 2015–2024 mean of $3.26 \text{ Gt C yr}^{-1}$ is close to the GCB2025 merged ocean-sink estimate of $3.2 \pm 0.4 \text{ Gt C yr}^{-1}$ (Friedlingstein et al., 2026). FDU-BTR and the GCB2025 data-product mean show broadly consistent long-term evolution and major low-frequency variations, although differences remain in the absolute magnitude of the sink. The recent corrected FDU-BTR estimate is also consistent in magnitude with the GCB2025 merged estimate.

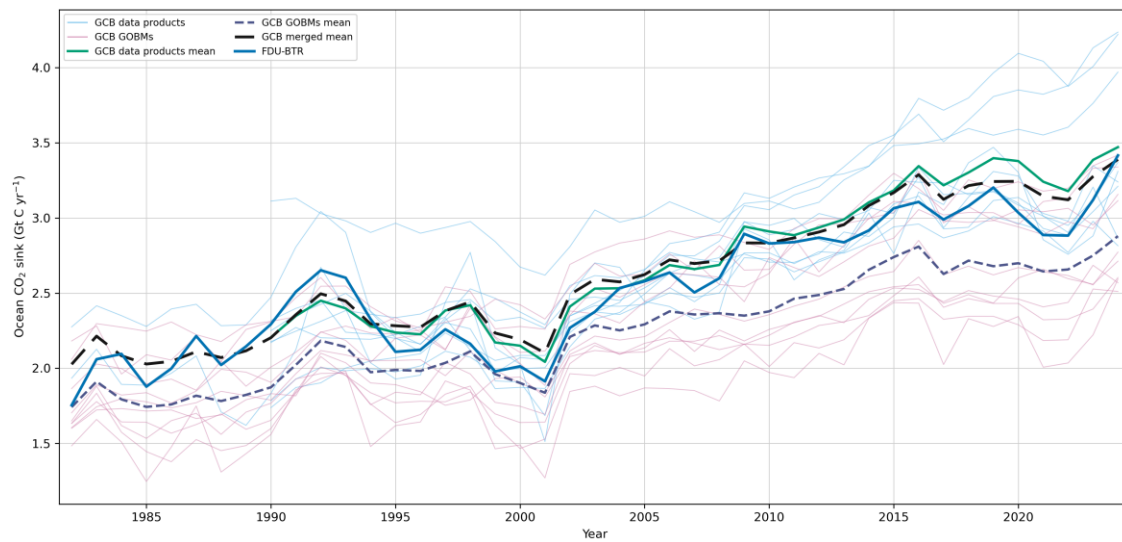


Figure 11. Global ocean CO₂ sink estimates from FDU-BTR and GCB2025 during 1982–2024. Thin lines show the individual GCB data products and GCB GOBMs. Bold lines show the GCB data products mean, GCB GOBMs mean, GCB merged mean, and FDU-BTR estimate. The FDU-BTR curve represents the study’s physical air–sea CO₂ sink with the $+0.65 \text{ Gt C yr}^{-1}$ river correction applied once. Positive values indicate net CO₂ uptake by the ocean from the atmosphere.

”

In addition to using pH and ALK built with ML by Chau et al, the authors use input to their method 6 ML products already built from SOCAT data. This is an additional circularity that needs to be assessed. This should mean that any data held out in training has already been used in training the average of these 6 ML products, thus limiting what held-out training data could tell us and further emphasizing the need for independent assessment of generalizability, as mentioned above. Please discuss this in reference to the held-out years from SOCAT (Line 221). McKinley et al. 2026 (ERL) show that most of the signal in these products is in the seasonality, and this seasonality has already been assigned in this method

for these years, so this does not appear to be an independent test as the authors suggest it is.

Response: We thank the reviewer for pointing out that the six selected external surface-ocean CO₂ products used for the background partly inherit information from SOCAT. They cannot serve as an independent prior.

We detrend each product and retain only its fixed monthly climatology before taking the median, so the prior supplies broad spatial and seasonal structure but not the held-out year's specific anomaly. We agree, however, that this fixed seasonal signal remains available in every OOF fold; consequently, the OOF experiments do not independently validate the complete FDU-BTR system but assess transfer of the learned residual conditional on a prescribed, SOCAT-informed prior. We have made this dependence explicit rather than treating withheld SOCAT rows as a fully external test, and we now direct skill assessment to the external LDEO/GLODAP/SPOTS/HOT/BATS evaluations, acknowledging that these still cannot remove all source-level overlap.

Sect. 8 (lines 755–765), we added:

“The results of this study suggest that the main value of the BTR framework lies not in the decomposition itself, but in how the reconstruction task is distributed among different sources of information. The multi-product climatological background constrains the stable large-scale spatial and seasonal structure, the explicit thermal component represents the known first-order temperature response, and machine learning focuses on the remaining nonlinear variability and departures not fully represented by the prescribed components. This division of the learning task reduces the amount of structure that must be inferred directly from sparse observations and provides a useful constraint for global extrapolation. At the same time, the quality of the background state remains important because systematic errors in the prior must ultimately be compensated by the residual model. The present background is constructed from existing surface-ocean CO₂ products and should therefore be regarded as an informed large-scale prior rather than an independent physical reference state. Agreement between FDU-BTR and products contributing to this background cannot therefore be interpreted as fully independent validation. Future improvements may arise not only from more flexible residual models, but also from more accurate and physically independent prior information.

”

The sample weighting (3.2.3) may be the most useful innovation here. It would be very interesting to see the impact of not including sample weighting and including it so that the impact could be better understood by the community.

Response: We agree that the effect of the sample-weighting scheme should be evaluated directly. We therefore added a matched weighted-versus-unweighted Spatial OOF experiment using the same selected spatial blocks and evaluation procedure. Weighting reduces RMSE from 11.52 to

11.40 μatm in the intermediate-coverage group and from 13.69 to 13.58 μatm in the low-coverage group, while the high-coverage group changes only slightly from 9.11 to 9.13 μatm . These results indicate that the weighting scheme provides a modest but targeted benefit in more weakly observed regions, consistent with its intended purpose, rather than a uniform improvement across all coverage classes.

Changes in the revised manuscript: sample-weight construction, Sect. 3.3.2 (lines 289–318); matched weighting comparison, Sect. 4.2 (lines 429–450) and Table 4.

“3.3.2 Sample weighting

SOCAT observations are unevenly distributed in space and time. We therefore assigned each training sample a fixed weight that jointly accounts for disagreement among the CORE6 background products and the scarcity of annual observational coverage.

For calendar month m and 1° grid cell g , let $p_{p',m,g}$ denote the detrended monthly climatology of CORE6 product p , calculated from 1982 to 2024. Where at least three CORE6 products were available, their disagreement was quantified using the scaled median absolute deviation:

$$M_{m,g} = 1.4826 \text{median}_p [p_{p,m,g} - \text{median}_{p'}(p_{p',m,g})]. \quad (12)$$

Valid $M_{m,g}$ values were pooled across all calendar months and ocean grid cells, and their 50th and 90th area-weighted empirical percentiles, Q_{50} and Q_{90} , were calculated using $\cos(\text{latitude})$ as the area weight. Product disagreement was then normalized as

$$U_{m,g} = \min \left[1, \max \left(0, \frac{M_{m,g} - Q_{50}}{Q_{90} - Q_{50}} \right) \right], \quad (13)$$

where product disagreement was unavailable within the valid ocean domain, $U_{m,g}$ was set to zero without spatial interpolation.

Annual observational scarcity was quantified from the number C_y of distinct observed 1° grid-cell–calendar-month combinations in SOCAT for year y , with multiple observations in the same combination counted once. Scarcity was transformed as $R_y = C_y^{-1/2}$ and normalized across the full 1982–2024 period:

$$T_y = \frac{R_y - \min_{1982 \leq k \leq 2024} R_k}{\max_{1982 \leq k \leq 2024} R_k - \min_{1982 \leq k \leq 2024} R_k}. \quad (14)$$

If all annual counts were identical, T_y was set to zero. No decadal grouping or temporal smoothing was applied.

For training sample i observed in year y_i , calendar month m_i , and grid cell g_i , the final weight was defined as

$$w_{i,UT} = 1 + U_{m_i,g_i} T_{y_i}. \quad (15)$$

Thus, $1 \leq w_{i,UT} \leq 2$, and weights greater than one occur only when both background-product disagreement and annual sampling scarcity are elevated. The complete weight field was constructed

once from the fixed input datasets and was not recalculated separately for individual validation folds. Sample weights were applied only to the model-training loss. For models trained with the MAE objective, the weighted loss is expressed as

$$\mathcal{L}_w = \frac{\sum_{i=1}^N w_{i,UT} |R_i - \hat{R}_i|}{\sum_{i=1}^N w_{i,UT}}, \quad (16)$$

where R_i and $\hat{R}_i$ denote the observed residual and model-predicted residual for training sample i , respectively. The weighting does not alter the SOCAT observations themselves and is not used in the calculation of subsequent model-performance metrics. It therefore affects only the relative contribution of individual training samples during model fitting, whereas model generalization is evaluated using the original unweighted observation–prediction pairs.

”

“4.2 Coverage and weighting sensitivity

Spatial OOF samples were stratified according to observational density over 1982–2024 to examine how historical SOCAT coverage affects spatial generalization. Observation density was defined as the number of distinct observation years for each calendar-month $\times$ 1° grid cell, normalized to a 0–1 range, and classified as low, intermediate, or high coverage using thresholds of 0–1/3, 1/3–2/3, and 2/3–1, respectively. To ensure a comparable spatial evaluation across coverage classes, five 10° $\times$ 10° Spatial OOF blocks were randomly selected from each class. The same selected blocks were then used for the matched weighted and unweighted experiments, with weighting applied only during model training and all OOF metrics calculated from the original unweighted held-out samples.

Prediction error increases systematically as historical observational coverage decreases (Table 4). Under the weighted experiment, RMSE increases from 9.13 μatm in the high-coverage group to 11.40 μatm in the intermediate-coverage group and 13.58 μatm in the low-coverage group. This progression indicates that spatial generalization becomes increasingly difficult as historical observational support weakens, even when the number of evaluated Spatial OOF blocks is held constant across coverage classes.

Sample weighting has only a modest effect on prediction error, with small RMSE reductions in the intermediate- and low-coverage groups and little change in the high-coverage group. Relative to the unweighted experiment, weighting reduces RMSE by 0.12 μatm in the intermediate-coverage group and by 0.11 μatm in the low-coverage group, whereas the high-coverage group changes little (+0.02 μatm). This pattern is consistent with the intended role of the weighting scheme: to increase the influence of samples from poorly constrained conditions rather than to improve performance uniformly across all regions. The effect remains small compared with the much larger error increase associated with declining observational coverage, but it provides a measurable improvement where the reconstruction is most weakly constrained.

Table 4. Coverage-stratified Spatial OOF performance

Observational coverage	$w_{i,UT}$ RMSE (μatm)	No-weight RMSE (μatm)
High	9.13	9.11
Intermediate	11.40	11.52
Low	13.58	13.69

”

The choice of regions being whole basins (Atlantic, Pacific, etc.) that cut across enormous physical and biogeochemical variability from equatorial to subtropical to subpolar. This appears counter to the motivation presented starting at line 265 that considering things such as upwelling dominance and ENSO. Other methods have used biomes (e.g. Landschutzer’s SOM FFN) that are more true to physical and biogeochemical structures. The authors need to discuss why their choice is appropriate and clarify how their choice fits with the motivation starting at line 265. They should also ensure that this is a useful approach, given the significant additional complexity and effort for boundary-blending that it necessitates.

Response: We agree that a single whole-basin partition can combine dissimilar physical and biogeochemical regimes. The revised model uses six large-scale biogeochemical biomes based on Fay and McKinley (2014), with the aggregation informed by Gregor et al. (2019). The partition separates tropical, subtropical, mid-to-high-latitude, Southern Ocean, and enclosed/semi-enclosed environments while keeping regional sample support sufficient for fitting. Final predictions are not blended across region boundaries; a directional training buffer of at most 5° instead expands the training domain near adjacent ocean-connected biomes. This is a pragmatic regionalization, not a claim that each biome is internally homogeneous.

Changes in the revised manuscript: Sect. 3.2.2 (lines 253–266), Fig. 2.

“3.2.2 Biome-based regionalization

The remaining variability in surface-ocean $f\text{CO}_2$ is jointly influenced by multiple processes, including ocean circulation, mixing, sea ice, biological activity, and atmospheric forcing, whose relative contributions vary substantially across the global ocean. To account for this spatial heterogeneity in the controls on residual $f\text{CO}_2$ variability, we followed the global ocean biogeochemical partitioning framework of Fay and McKinley (2014) and further merged the original regions with reference to the aggregation strategy used by Gregor et al. (2019). The resulting regionalization adopted in this study is shown in Fig. 2.

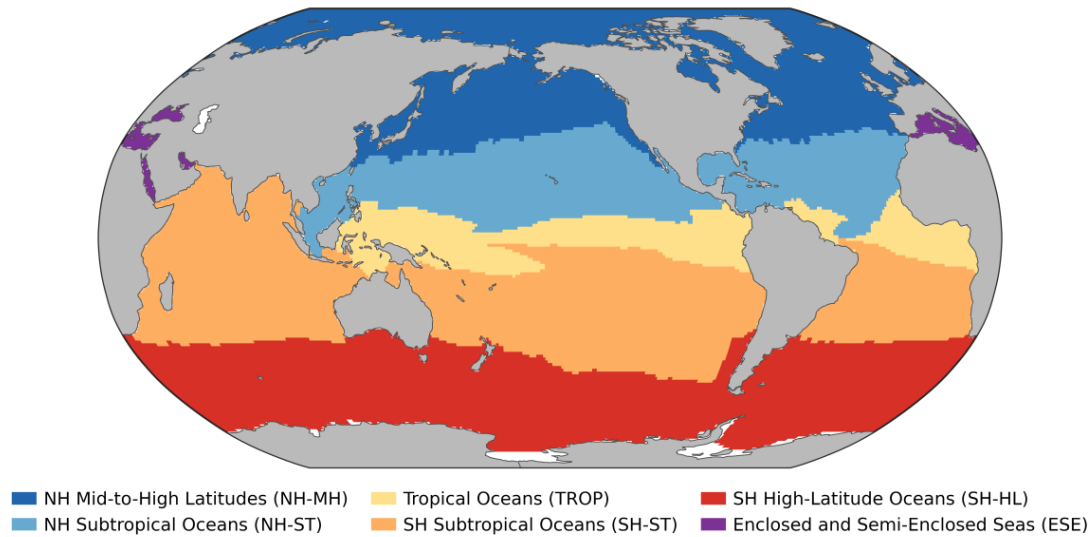


Figure 2. Global distribution of the six large-scale ocean biome regions used in this study: Northern Hemisphere mid-to-high latitudes (NH-MH), Northern Hemisphere subtropical oceans (NH-ST), tropical oceans (TROP), Southern Hemisphere subtropical oceans (SH-ST), Southern Hemisphere high-latitude oceans (SH-HL), and enclosed and semi-enclosed seas (ESE).

For each biome, an independent residual-learning model was trained. When regional models are fitted separately, the training samples can change abruptly across biome boundaries, even though the underlying ocean environment is generally more continuous between adjacent regions. To strengthen the training constraint near these boundaries, a directional ocean-connectivity training buffer was constructed around each biome core region. A training buffer with a maximum width of 5° was used, allowing the target biome model to incorporate observations from directly adjacent biomes that lie close to the shared boundary. Core samples from the target region were always retained, and the final training set for each regional model therefore consisted of its core samples together with neighboring samples that satisfied the buffer criterion (Zhong et al., 2022). After regional partitioning, boundary buffering, data matching, and quality control, 364,701 valid grid-cell-month samples were retained for training the regional residual models.

”

Minor comments:

Line 110. Please provide a reference to justify the temperature correction applied.

Response: We have clarified the temperature-sensitivity coefficient. The original preprocessing correction to the SOCAT target has been removed; 0.0423 K^{-1} is now used only in the explicit thermal component and is discussed with Takahashi et al. (1993) and Bennington et al. (2022a).

Changes in the revised manuscript: Sect. 3.1.2 (lines 181–199).

“This coefficient represents the first-order exponential sensitivity of surface-ocean CO_2 to temperature

changes under approximately constant seawater chemical composition (Takahashi et al., 1993). Similar temperature-sensitivity formulations have been used in residual-based surface-ocean CO₂ mapping to explicitly separate temperature-driven variability, allowing the machine-learning model to focus on the remaining non-thermal physical and biogeochemical signals and reducing the need to disentangle competing controls directly from the observations (Bennington et al., 2022a).
”

Line 157. Cite also Gloege et al. 2022

Response: We have added Gloege et al. (2022) to the Introduction when discussing reconstruction methods that use model-derived prior information and observation-based correction.

The reference appears in the Introduction (lines 24–72), where prior-guided approaches are discussed.

“LDEO-HPD uses a biogeochemical model prior followed by machine-learning correction (Gloege et al., 2022), while the pCO₂-residual approach explicitly separates the first-order temperature signal before residual learning (Bennington et al., 2022a).”

Line 170-180. It would be appropriate to cite Bennington et al. 2022a as the originator of the removal of the pCO₂T term to create residual.

Response: We agree that Bennington et al. (2022a) is an important precedent for explicitly separating the temperature contribution before residual learning. The revised manuscript cites it in the method context rather than presenting that general idea as unique to FDU-BTR.

We now cite Bennington et al. (2022a) when introducing related residual-learning approaches and again when discussing the explicit first-order temperature component.

Changes in the revised manuscript: residual-learning precedent, Sect. 3.1 (lines 135–159), and thermal formulation, Sect. 3.1.2 (lines 181–199).

“Conceptually similar residual-learning approaches have been used in previous surface-ocean CO₂ mapping studies, in which a climatological or model-derived reference state is specified before machine learning is applied to the observation–reference residuals (Bennington et al., 2022a, b; Heimdal et al., 2026; Gloege et al., 2022).
”

“This coefficient represents the first-order exponential sensitivity of surface-ocean CO₂ to temperature changes under approximately constant seawater chemical composition (Takahashi et al., 1993). Similar temperature-sensitivity formulations have been used in residual-based surface-ocean CO₂ mapping to explicitly separate temperature-driven variability, allowing the machine-learning model to focus on the

remaining non-thermal physical and biogeochemical signals and reducing the need to disentangle competing controls directly from the observations (Bennington et al., 2022a).
”

Line 288. Please provide detail on the basin-specific feature sets

Response: We agree that the regional predictor sets require a clearer account. Sect. 3.2.1 lists the common candidate predictors. Fig. S2 shows which predictors appear in each biome-specific model and their relative importance. These importances describe predictive use, not unique physical causation.

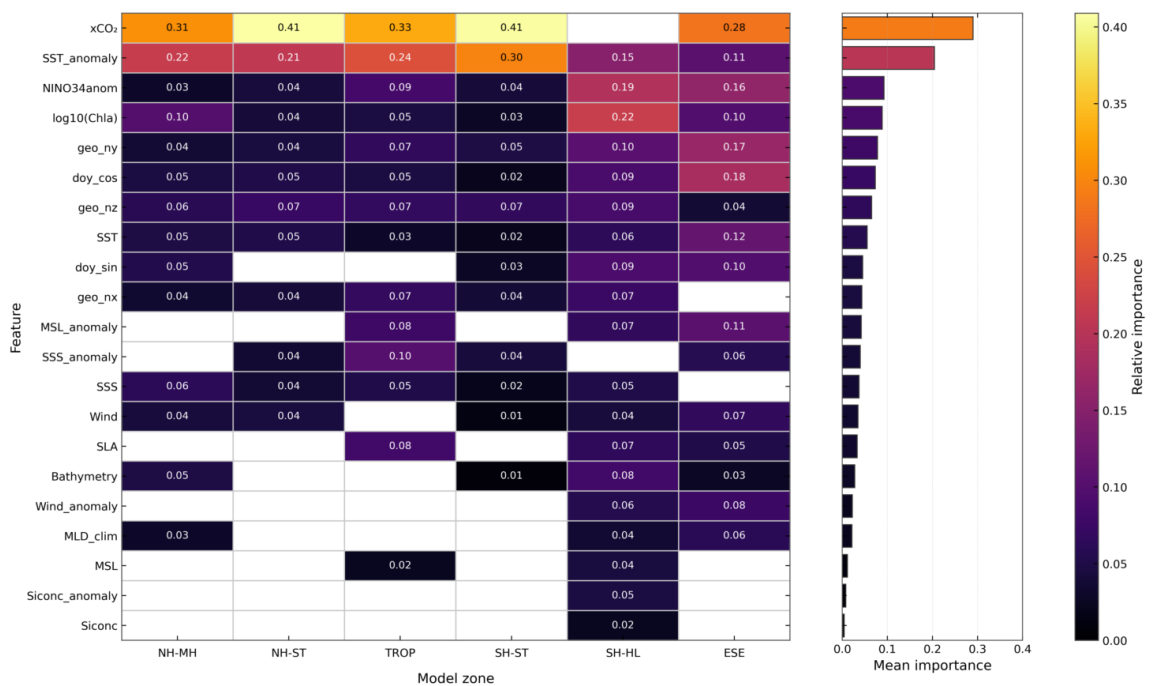


Figure S2. Relative importance of environmental predictors in biome-specific models. The bar chart on the right shows each feature’s mean importance computed over the biomes in which that feature is available.

Figure 2. I think this figure may just be demonstrating overfitting. What is the independent test set here is not clear. If it is the SOCAT held-out years or the NCEI moorings, these are not actually independent, as noted above. Panel b averages across the prior climatology and temperature driven variability, but does not have biogeochemical variability, and then in c we find that this method can fit back the biogeochemical variability. It is not clear that this proves much about skill in this method. The discussion of this figure far overstates the value of this comparison.

Response: We agree that the earlier figure was interpreted too strongly and did not distinguish internal component assessment from independent evaluation.

The original figure has been removed. Model generalization is evaluated separately in Fig. 3 using Temporal, Spatial, and Joint spatiotemporal OOF predictions, and the manuscript explicitly states that these are internal generalization diagnostics rather than fully independent validation of the complete reconstruction.

Changes in the revised manuscript: OOF interpretation, Sect. 4.1 (lines 400–428), Fig. 3.

“4.1 Out-of-fold performance

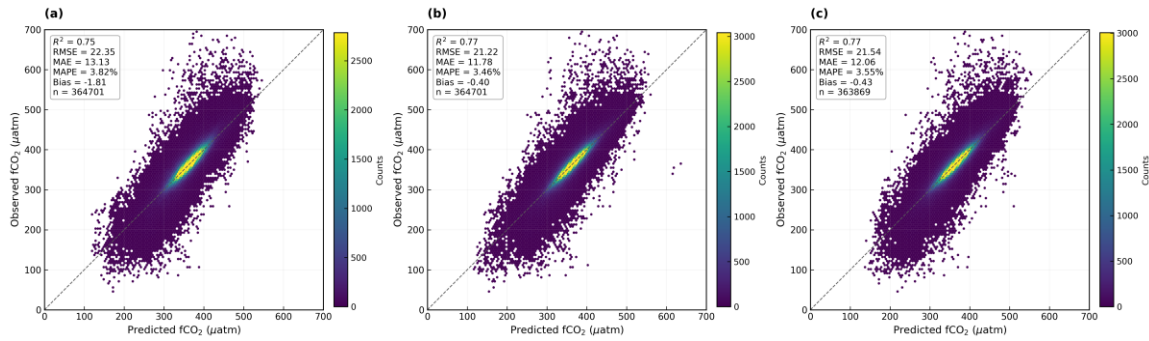


Figure 3. Out-of-fold (OOF) validation of FDU-BTR under different withholding schemes. Observed versus predicted surface-ocean fCO₂ for (a) temporal OOF, (b) spatial OOF, and (c) joint spatiotemporal OOF. Colors indicate sample density, and the dashed line denotes the 1:1 relationship.

The transferability of the learned residual relationships under different forms of withheld observational support was evaluated using Temporal, Spatial, and Joint spatiotemporal OOF schemes (Fig. 3). Because the BTR background is a fixed multi-product monthly climatology constructed in advance, these experiments specifically assess the out-of-fold generalization of the learned residual component rather than providing fully independent validation of the complete FDU-BTR system.

Aggregate predictive performance remains similar across the three withholding schemes. RMSE ranges from 21.22 to 22.35 μatm , MAE from 11.78 to 13.13 μatm , and R^2 from 0.75 to 0.77. Spatial and Joint OOF yield RMSE values of 21.22 and 21.54 μatm , respectively, while Temporal OOF is moderately higher at 22.35 μatm . Predicted and observed fCO₂ values remain broadly aligned with the 1:1 relationship under all three schemes. The relatively small differences among the three schemes indicate that the learned residual relationships remain broadly transferable when temporal, spatial, or combined spatiotemporal support is withheld, rather than depending strongly on nearby training samples.

This aggregate stability, however, masks pronounced regional and seasonal heterogeneity (Fig. S1). NH-MH consistently shows the largest regional RMSE, whereas SH-ST exhibits the lowest errors across the three OOF schemes. Semi-enclosed seas (ESE) show particularly strong sensitivity to the withholding strategy: Temporal OOF produces a bias of approximately $-8.81 \mu\text{atm}$, whereas Spatial OOF remains close to zero and Joint OOF shows a substantially smaller negative bias, indicating greater

difficulty in temporal transfer within this region. Seasonal performance shows a similarly structured contrast. RMSE increases consistently from approximately 18.73 μatm in DJF to 25.62 μatm in SON, while biases under Spatial and Joint OOF remain comparatively small. The seasonal deterioration is therefore expressed mainly as increased prediction scatter rather than a systematic shift in the reconstructed mean state.

These OOF experiments show that broadly stable aggregate performance can coexist with substantial regional and seasonal differences in reconstruction difficulty. The largest departures are concentrated in specific regimes, particularly NH-MH in overall error and ESE in temporal transfer, rather than being distributed uniformly across the global ocean. This distinction motivates the subsequent examination of whether differences in observational support alone can explain the spatial variation in prediction skill. Biome-specific feature combinations and feature importance are presented in Fig. S2.

”

Line 362. This statement is not justified “Figure 3 shows that the controls on residual pCO₂ variability are strongly basin dependent.” Instead the figure shows that a few variability get very low weights in a few regions. The main controls are consistent (SST_{anom}, xCO₂, DIC and pH_{anom}). Consider revising the figure to exclude something like the bottom quartile of features chosen, and this will give a better assessment of what is really driving the predictions.

Response: We agree that model feature importance does not identify causal oceanographic controls. We have removed that interpretation.

The full biome-by-feature display is retained as Fig. S2 so that predictor availability and heterogeneity are visible; its accompanying mean-importance ranking makes the dominant predictive features easier to identify without concealing the low-contribution features. We have not claimed that the bottom quartile was removed from Fig. S2. The text now discusses predictive associations rather than physical attribution.

Changes in the revised manuscript: end of Sect. 4.1 (lines 400–428), Fig. S2.

Also note that with DIC and pH_{anom} and pH having strong importance, this method may simply be restating the predictions of DIC and pH from Copernicus (Chau et al. 2022, 2024). The fact that these predictions are also from ML with SOCAT input data should be discussed.

Response: We agree that strong importance of reconstructed pH and DIC would be difficult to interpret as an independent contribution. We therefore removed pH, DIC, and other reconstructed carbonate-system predictors from the final model and retrained the reconstruction. Their removal increased RMSE by only 0.16 μatm relative to the version including these predictors, suggesting

that much of the predictive information they provide can also be represented by the remaining environmental predictors.

Section 4.4: This appears to be comparing back to the held-out SOCAT data, which, as mentioned above, has already been used in the development of the products used for the climatology. How is this analysis impacted by this?

Response: We agree that held-out SOCAT observations do not provide fully independent validation of a framework whose prescribed background partly inherits SOCAT information.

We therefore changed both the methodology and the interpretation. SOCAT OOF predictions are now used only as internal diagnostics of the learned residual relationship. The climatological prior is explicitly described as a detrended, fixed monthly background rather than an independent field, and the remaining dependence on existing products is acknowledged in the Discussion.

In addition, the former residual-based uncertainty analysis has been replaced by conditional quantile regression with nested spatial-block conformal calibration and external observational evaluation.

Changes in the revised manuscript: conditional interpretation of OOF, Sect. 4.1 (lines 400–428); uncertainty method, Sect. 3.4 (lines 345–380); external interval evaluation, Sect. 6.1 (lines 583–618); prior dependence, Sect. 8 (lines 754–809).

“6.1 Prediction uncertainty

FDU-BTR predictive uncertainty exhibits pronounced spatial differences across the global ocean (Fig. 8). The mean full width of PI90 over 1982–2024 is relatively large in high-latitude regions and parts of the tropics, but generally smaller in the subtropical open ocean. The 67% prediction interval (PI67) shows a spatial pattern broadly consistent with the 90% prediction interval (PI90), but is narrower overall (Fig. 8a, b). Across valid mapped cells, the mean full widths are 47.2 μatm for PI90 and 25.8 μatm for PI67. This indicates systematic regional differences in the degree of predictive constraint.

To examine whether the prediction intervals reflect actual prediction errors, we further evaluated them using five external observational datasets: LDEO, SPOTS, GLODAP, HOT, and BATS (Fig. 8c). All matched samples were divided into five groups according to increasing PI90 width, with Q1 representing the narrowest intervals and Q5 the widest. PI90 width is positively associated with absolute prediction error, with a sample-level Spearman correlation coefficient of 0.39 and a 95% bootstrap confidence interval of 0.35–0.44. The dataset-balanced MAE rises overall from 14.3 μatm in Q1 to 39.1 μatm in Q5. Although Q3 has a slightly lower MAE than Q2, errors are substantially larger in Q4 and Q5, indicating that interval width helps distinguish relative error risk rather than ranking every group

strictly monotonically.

The empirical PI90 coverage (C90) across the five groups ranges from 75.6% to 82.8%, below the nominal 90% level. Thus, PI90 provides useful information about relative error risk but does not achieve nominal coverage for these external observations.

Predictive uncertainty shows an even clearer contrast between shallow and deeper waters (Fig. 8d). Using 200 m water depth as the threshold, the PI90 half-width in shallow-water regions is mostly approximately 39–56 μatm , whereas values in deeper-water regions are concentrated at approximately 18–21 μatm . PI67 shows a consistent depth-dependent contrast. Uncertainty in shallow-water regions also exhibits stronger month-to-month variability, whereas uncertainty in deeper-water regions is generally lower and more stable. Here, the PI half-width is defined as half the difference between the upper and lower prediction-interval bounds.

Overall, FDU-BTR predictive uncertainty shows two clear characteristics: first, larger actual prediction errors are generally associated with wider prediction intervals, indicating that interval width provides a useful measure of relative prediction risk; second, uncertainty is substantially higher in shallow waters than in deeper waters, indicating that accurate reconstruction in shallow-water and shelf environments remains particularly challenging. The pronounced shallow-water contrast also calls for greater caution when interpreting local variability in heterogeneous shallow-water environments.

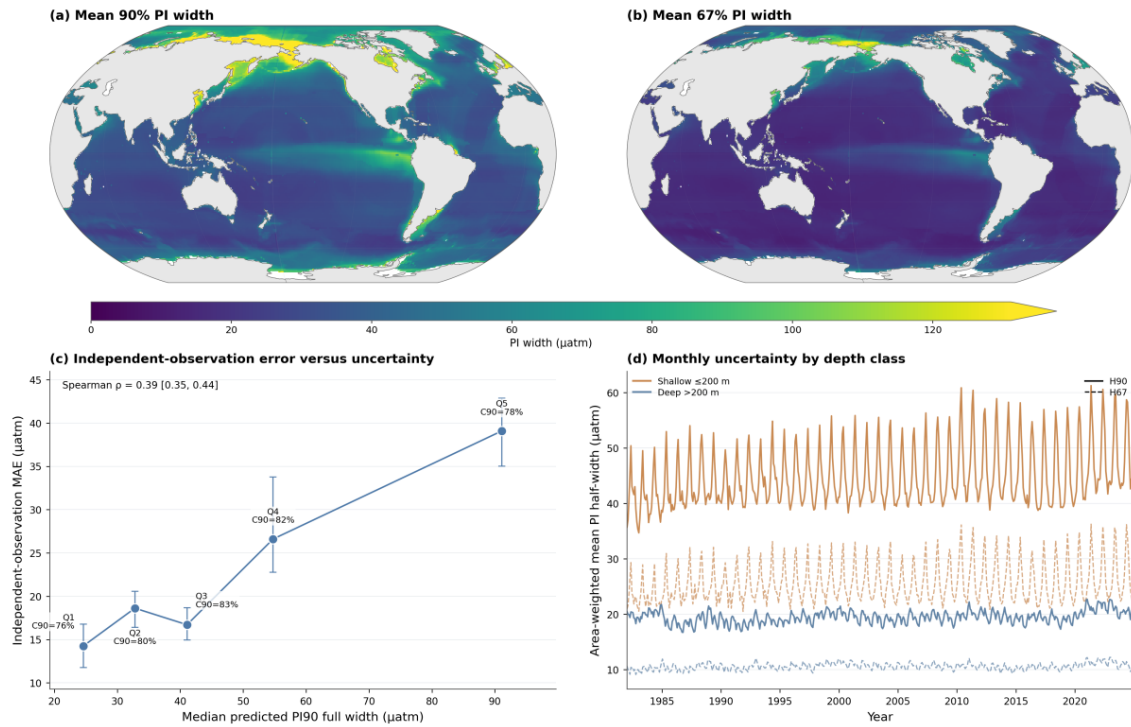


Figure 8. Spatial distribution and evaluation of prediction uncertainty in the FDU-BTR fCO_2 product (1982–2024). (a, b) Mean spatial distributions of the 90% and 67% prediction-interval widths, respectively; wider intervals indicate greater estimated uncertainty. (c) Dataset-balanced MAE versus median PI90 width for five groups of external observations, from the narrowest intervals (Q1) to the widest (Q5). Dataset-balanced MAE is the

unweighted mean of the dataset-specific MAEs within each group. Labels show the empirical PI90 coverage (C90) and sample size; the reported Spearman correlation is calculated at the individual-sample level between PI90 width and absolute prediction error. (d) Monthly area-weighted mean prediction-interval half-widths in shallow and deeper waters.

”

Section 5.1 should be removed as these data are in SOCAT, as mentioned above.

Response: We agree. The NCEI buoy validation has been removed from the revised manuscript.

The current Sect. 5.1 (lines 488–518) concerns large-scale spatial and temporal patterns; the separate external observation analysis is in Sect. 5.2 (lines 519–538).

Figure 7. Why is evaluation at BATS so different from Bennington et al. 2022a,b and Gloege et al 2022 and McKinley et al. 2026, where products are correlated at 0.9–0.95? HOT is consistent with prior work. The authors need to explain.

Response: We thank the reviewer for identifying the original BATS discrepancy. The revised preprocessing retains quality-controlled BATS observations shallower than 10 m. In the updated station time-series analysis (Fig. 7c), the monthly Pearson correlation is $R = 0.96$ for 346 matched months. The common-product subset in Table S1 uses a different set of matched months and therefore has different BATS metrics. We treat this as a site-specific temporal result rather than evidence of global skill.

Section 6.2. Given that the climatologies already give most of the signal based on the full SOCAT, what do the thinning experiments tell us about the real impact? The thinning only impacts this method, not what trains the original methods averaged for the climatology. I don’t trust the resulting relationships because they still depend on full-coverage SOCAT for other products that make the climatology. Please justify these experiments in this context.

Response: We agree that thinning SOCAT residual-training observations while retaining a climatological prior from existing SOCAT-constrained products would confound the intended coverage experiment.

We redesigned the thinning experiment as a separate, single global CatBoost analysis that predicts SOCAT $f\text{CO}_2$ directly. It uses neither the six-product background nor the explicit BTR thermal term or residual target. Hyperparameters selected on the FULL sample are held fixed across the thinning realizations.

Fifty temporal–spatial pairs are constructed with matched retained sample counts and evaluated on the same frozen external sample. This design removes confounding by the BTR prior from the

comparison of the two thinning strategies; it does not establish that coverage effects are independent of all modeling choices.

Changes in the revised manuscript: Sect. 6.2 (lines 619–669), Fig. 9.

“6.2 Observational coverage constraints

The coverage-stratified Spatial OOF comparison showed higher prediction errors in the lower-coverage groups within the selected spatial blocks. This contrast suggests that regions with weaker historical observational support tend to present greater reconstruction difficulty, although observational density is unlikely to be the only factor involved. The prediction-interval analysis further revealed substantial spatial variation in predictive constraint. Together, these findings raise a more direct question: when the amount of training data is held comparable, does the temporal or spatial organization of observational support itself affect extrapolation skill? To address this question, we conducted controlled temporal–spatial thinning experiments in which training-sample size was matched between the two thinning strategies.

The experiments used unique SOCAT cell-month observations from 1985–2024. A single global CatBoost model was trained directly on surface-ocean fCO₂ using a fixed set of environmental and spatiotemporal predictors. To isolate the effect of observational coverage, model hyperparameters were optimized once using the FULL dataset and then held fixed across all thinning realizations. We generated 50 matched temporal–spatial thinning pairs. Temporal thinning removed complete calendar years, whereas spatial thinning removed complete spatial blocks throughout the record. For each temporal realization, a spatial realization with the same number of retained cell-month observations was constructed, allowing the effects of temporal and spatial coverage loss to be compared at matched training-sample sizes.

The distribution of observational support was characterized using Shannon effective coverage. For the temporal or spatial dimension d , Shannon entropy was calculated as

$$H_d = - \sum_{i=1}^{N_d} p_{d,i} \ln p_{d,i}, \quad (24)$$

and converted to effective coverage as

$$C_d^{\text{eff}} = \frac{\exp(H_d)}{N_d}, \quad (25)$$

where $p_{d,i}$ is the fraction of retained observations in bin i , and N_d is the total number of bins. Thus, C_t^{eff} and C_s^{eff} describe the effective temporal and spatial extent of observational support, respectively. The total retained training coverage was quantified separately as

$$C_{\text{OCC}} = \frac{N_{\text{obs}}}{N_t N_s}, \quad (26)$$

where N_{obs} is the number of retained unique grid-cell–month observations, N_t is the number of monthly time steps, and N_s is the number of ocean grid cells within the analysis domain. Thus, $N_t N_s$ represents

the total number of possible grid-cell–month combinations over the study period.

All realizations were evaluated against the same frozen external dataset comprising LDEO, GLODAP, HOT, BATS and SPOTS observations. Prediction skill was evaluated using pooled RMSE and a dataset-balanced macro-RMSE that assigned equal weight to the five external datasets.

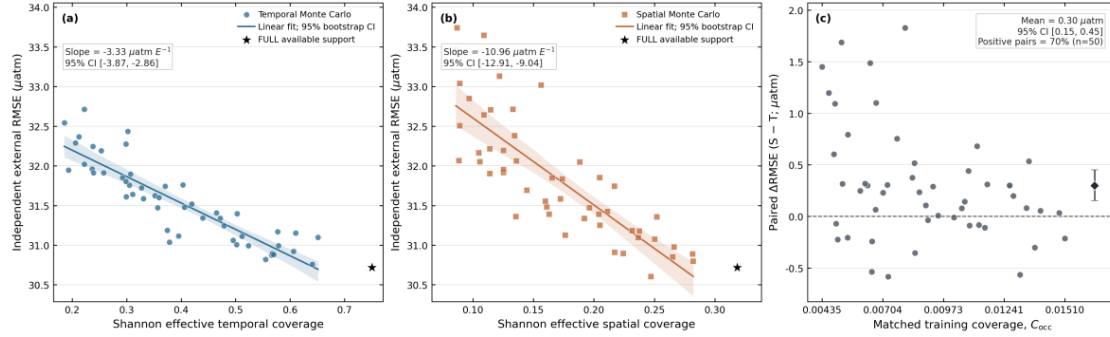


Figure 9. Effects of temporal and spatial observational thinning on external $f\text{CO}_2$ prediction skill. (a) External RMSE versus Shannon effective temporal coverage. (b) External RMSE versus Shannon effective spatial coverage. Solid lines and shaded bands show linear fits and 95% bootstrap confidence intervals; black stars denote the FULL model. (c) Paired RMSE differences between row-matched spatial and temporal thinning realizations as a function of matched training coverage C_{occ} . Gray points show individual pairs, and the black line and gray band indicate the paired mean and its 95% bootstrap confidence interval.

Both temporal and spatial thinning show increasing external RMSE as effective coverage decreases (Fig. 9a, b). Spatial thinning also produces substantially greater variability in prediction error under low-coverage conditions. Realizations with similar effective spatial coverage can yield markedly different RMSE values, indicating that effective coverage alone does not uniquely determine predictive performance on external observations. This interpretation is consistent with Heimdal et al. (2024), who showed that adding a modest number of observations during Southern Hemisphere winter and along meridional gradients in the Southern Ocean substantially reduced reconstruction bias and RMSE, demonstrating that strategically placed observations can provide disproportionate gains in reconstruction skill.

The matched experiments provide a direct comparison between the two thinning strategies at equivalent training-sample sizes (Fig. 9c). Spatially organized observational loss produces a larger average pooled RMSE penalty than temporally organized loss, with a mean paired RMSE difference of $0.30 \mu\text{atm}$ (95% CI: $0.15\text{--}0.45 \mu\text{atm}$). This difference is largest under sparse training coverage and decreases as observational support increases. By contrast, the corresponding difference in dataset-balanced macro-RMSE is much smaller, indicating that the effect of spatial thinning is not uniform across the external observational datasets.

These thinning experiments therefore show that prediction skill depends on both the amount and the spatiotemporal distribution of observational support. At matched sample sizes, temporal and spatial reductions in coverage do not produce equivalent prediction errors, with spatially organized loss imposing the larger average penalty, particularly under sparse coverage.”

Paragraph starting at line 654. For the reasons noted above, most of these statements are not valid. Please revise.

Response: We agree that the original Discussion contained claims stronger than the revised evidence supports.

The Discussion has been extensively rewritten. The revised text now distinguishes the role of the climatological prior, the observational-coverage experiment, prediction-interval calibration, coastal limitations, and the degree of agreement in the global carbon sink. We no longer state that spatial sparsity is universally the dominant limitation or that algorithm choice is secondary.

The discussion of prior dependence, observational coverage, interval undercoverage, coastal limits, and carbon-sink interpretation is now contained in Sect. 8 (lines 754–809).

I find it odd to have data availability and code availability sections prior to Conclusions.

Response: We agree that availability statements placed before the Conclusions interrupted the paper's narrative.

We have reordered the manuscript so that the Conclusions now appear first as Sect. 9, followed by Data availability as Sect. 10 and Code availability as Sect. 11.

The Conclusions appear in Sect. 9 (lines 810–829), followed by Data availability in Sect. 10 (lines 830–848) and Code availability in Sect. 11 (lines 849–851).