Response to Reviewer #2

The reviewer's comments and our responses are shown in black, and excerpts from the revised manuscript are shown in blue. Line numbers cited in our responses are shown in red and refer to the PDF of the revised manuscript without changes marked.

General comments:

This study presents a daily global 4 km dataset of POC, PON and POC:PON for 1998-2023. Such a record could be useful for investigating marine POM dynamics, carbon export, nutrient cycling and long-term ocean change, particularly because comparable continuous global products are limited.

My main concern is whether the record is sufficiently consistent over time for the proposed trend applications. OC-CCI sensor composition and data coverage change throughout the record, with SeaWiFS as the only source during the early years. Figure S4 shows clear changes in coverage, while the retrieved and reconstructed global time series differ most strongly during 1998-2003. The authors need to examine whether these changes affect the product and whether the effects differ among regions and sensor periods. The reconstruction should also be tested using more realistic gap patterns and data availability from different sensor periods. The gap-simulation experiment removes pixels at random on only two dates, whereas gaps in ocean-colour data vary with location, season and sensor. The data files should identify retrieved and reconstructed values and include uncertainty information.

Response: We sincerely thank you for these thoughtful and constructive suggestions. They helped us strengthen the evaluation of GDR-POM and clarify how readers can interpret and use the dataset. We addressed your three concerns as follows.

1. Temporal consistency across regions and sensor periods. To address your concern that changes in OC-CCI sensor composition and coverage could influence the reported trends, we replaced Fig. S4 from the original submission with Fig. S5 in the revised Supplement, which shows daily valid-pixel coverage before and after reconstruction for the global ocean and each of the four Longhurst biomes over 1998–2023. Global coverage increased around 2003 in both the retrieved and reconstructed records. In Sect. 4.1, we used the Temporal Gap Detection Method (TGDM) proposed by Van Oostende et al. (2022) to assess the sensitivity of the estimated trends to observational coverage. We compared trends with and without TGDM in both the full reconstructed record and the retrieval-only subset over 1998–2023. To assess sensitivity to the inclusion of the early SeaWiFS-only years, we repeated the analysis over 2003–2023 using the full reconstructed record. To examine whether this sensitivity differed among regions, we compared global and biome-level trends over the same 2003–2023 interval using TGDM masks derived from 1998–2023 and 2003–2023.

   TGDM changed the global POC and PON slopes more over 1998–2023 than over 2003–2023. The response also varied by biome. In polar waters, TGDM reduced the magnitude of the POC and PON increases, although both remained significant. In coastal waters, POC and PON trends were nonsignificant across the comparisons. POC:PON slopes were negative globally and in each biome examined across both periods and all tested coverage treatments, although the statistical significance of regional declines varied (Figs. 9, 10, and S5, S7–S9). The related revised text are as follows.

"4.1 Sensitivity of POC, PON, and POC:PON trends to spatiotemporal observational coverage

4.1.1 Global trends

Changes in the sensor composition of OC-CCI can alter the spatial and temporal distribution of the observations underlying GDR-POM. Such sampling effects may persist after inter-sensor bias correction,

particularly in coastal and high-latitude waters (Van Oostende et al., 2022; Van Oostende et al., 2023). Daily valid-pixel coverage over the global ocean was substantially higher after reconstruction, although seasonal variability persisted. Coverage increased around 2003 and generally remained above pre-2003 levels in both the original and reconstructed datasets (Fig. S5a). The original and TGDM-homogenised series were therefore compared to assess the sensitivity of global trend estimates to observational coverage. For 1998–2023, we analysed both the full reconstructed record and the retrieval-only subset to assess whether sensitivity to observational coverage persisted after gap filling. We then repeated the comparison over 2003–2023 using the full reconstructed record to assess whether the trend estimates depended on the inclusion of the early SeaWiFS-only years.

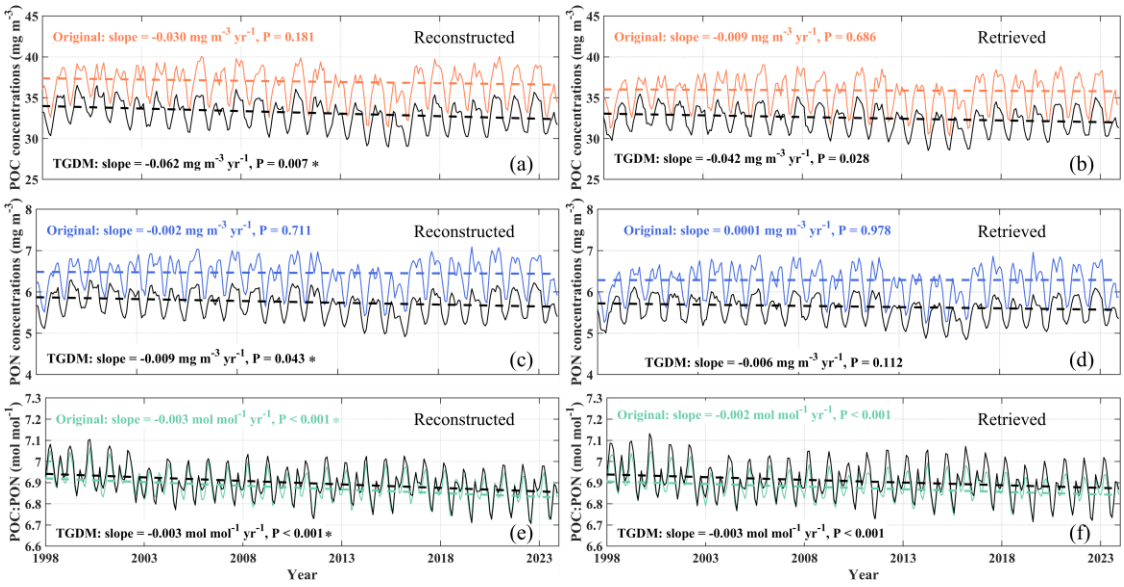


Figure 9. Spatially weighted monthly medians of global oceanic POC, PON, and POC:PON before and after TGDM homogenisation from January 1998 to December 2023. Monthly time series are shown for POC (a, b), PON (c, d), and POC:PON (e, f) in the full reconstructed record (left) and the retrieval-only subset (right). Coloured and black lines represent the original and TGDM-homogenised series, respectively. Dashed lines indicate trends estimated using the Sen slope estimator after removing the mean seasonal cycle. The asterisk (*) indicates that the derived trend is larger than the corresponding uncertainty.

Over 1998–2023, applying TGDM made the global oceanic POC and PON concentration slopes more negative (Fig. 9). The original POC slopes were −0.030 and −0.009 mg m^{-3} yr^{-1} in the full reconstructed record and the retrieval-only subset, respectively, and neither trend was significant. With TGDM, both records showed significant declines, with slopes of −0.062 mg m^{-3} yr^{-1} (P = 0.007) and −0.042 mg m^{-3} yr^{-1} (P = 0.028). PON slopes also became more negative, although the decline was significant only in the full reconstructed record. These similar responses show that gap filling does not eliminate the influence of observational coverage on aggregated concentration trends. The reconstructed record showed closer agreement between trends estimated with and without TGDM over 2003–2023 than over 1998–2023 (Fig. S7). For 2003–2023, the estimated POC slope was −0.039 mg m^{-3} yr^{-1} without TGDM. The two TGDM masks derived from the 1998–2023 and 2003–2023 records yielded slightly more negative and nearly identical slopes of −0.044 and −0.043 mg m^{-3} yr^{-1}, respectively. The corresponding PON slopes were −0.006 mg m^{-3} yr^{-1} without TGDM and −0.007 mg m^{-3} yr^{-1} with each of the two TGDM masks. None was statistically significant. The masks derived from 1998–2023 and 2003–2023 produced different mean concentration levels but nearly identical slopes, suggesting that the mask derivation period mainly affected concentration levels within this interval. The greater differences

between trends with and without TGDM over 1998–2023 highlight the importance of considering the early SeaWiFS-only period when interpreting full-record concentration trends.

The decline in POC:PON was more consistent across the sensitivity tests than the trends in POC and PON. Over 1998–2023, POC:PON declined significantly in both the full reconstructed record and the retrieval-only subset, with and without TGDM (all p <0.001; Fig. 9e and f). The slope for the full reconstructed record was −0.003 mol $mol^{-1} yr^{-1}$ with and without TGDM. For the retrieval-only subset, the slope was −0.002 mol $mol^{-1} yr^{-1}$ without TGDM and −0.003 mol $mol^{-1} yr^{-1}$ with TGDM. Over 2003–2023, slopes were approximately −0.002 mol $mol^{-1} yr^{-1}$ without TGDM and with either TGDM mask derived from the 1998–2023 or 2003–2023 record (Fig. S7). The persistence of the decline across these tests suggests that the direction of POC:PON was relatively insensitive to the tested differences in observational coverage and analysis period. Sampling effects and correlated retrieval errors may partly cancel in the ratio when they affect POC and PON proportionally, although the extent of this cancellation remains to be quantified.

4.1.2 Trends across Longhurst biomes

Daily valid-pixel coverage increased after reconstruction in all four Longhurst biomes, but remained lowest in the polar biome, where coverage showed pronounced seasonal variability (Fig. S5b–e). To assess how observational coverage influences regional trends, the original and TGDM-homogenised time series were compared across four Longhurst biomes using the TGDM mask derived from 1998–2023（Fig. 10a–c and Figs. S8 in the Supplement). In the polar biome, TGDM reduced the POC and PON slopes from 0.487 to 0.235 mg m^{-3} yr^{-1} and from 0.093 to 0.056 mg m^{-3} yr^{-1}, respectively, while all increases remained significant (P < 0.001, Fig. 10a and b). Thus, the direction of polar concentration change was consistent, although the estimated trend magnitudes were influenced by observational coverage. In coastal waters, neither variable showed a significant trend under either treatment. However, TGDM reduced seasonal peaks while seasonal minima changed little (Figs. S8b and f in the Supplement), demonstrating that coverage changes can affect the apparent seasonal amplitude even where long-term trends remain nonsignificant.

In the trade-wind biome, POC slopes were −0.059 and −0.056 mg m^{-3} yr^{-1} before and after TGDM homogenisation, respectively (Fig. 10a), despite P values falling on opposite sides of 0.05 (0.052 and 0.038, Fig. S8c in the Supplement). PON slopes were approximately −0.008 mg m^{-3} yr^{-1} and nonsignificant in both series. In the westerlies, TGDM lowered both POC and PON concentration levels and changed the POC trend from a nonsignificant slope of −0.004 to a significant decline of −0.045 mg m^{-3} yr^{-1} (P = 0.002). The PON slope also shifted towards a decline, although neither trend was significant (Fig. 10b).

POC:PON decreased significantly across all four biomes under both coverage treatments (Fig.10c). TGDM made the slopes more negative in the polar, coastal, and trade-wind biomes, from approximately −0.003 to −0.004, −0.001 to −0.002, and −0.002 to −0.003 mol $mol^{-1} yr^{-1}$, respectively. In the westerlies, the slope became slightly less negative, changing from −0.006 to −0.005 mol $mol^{-1} yr^{-1}$.

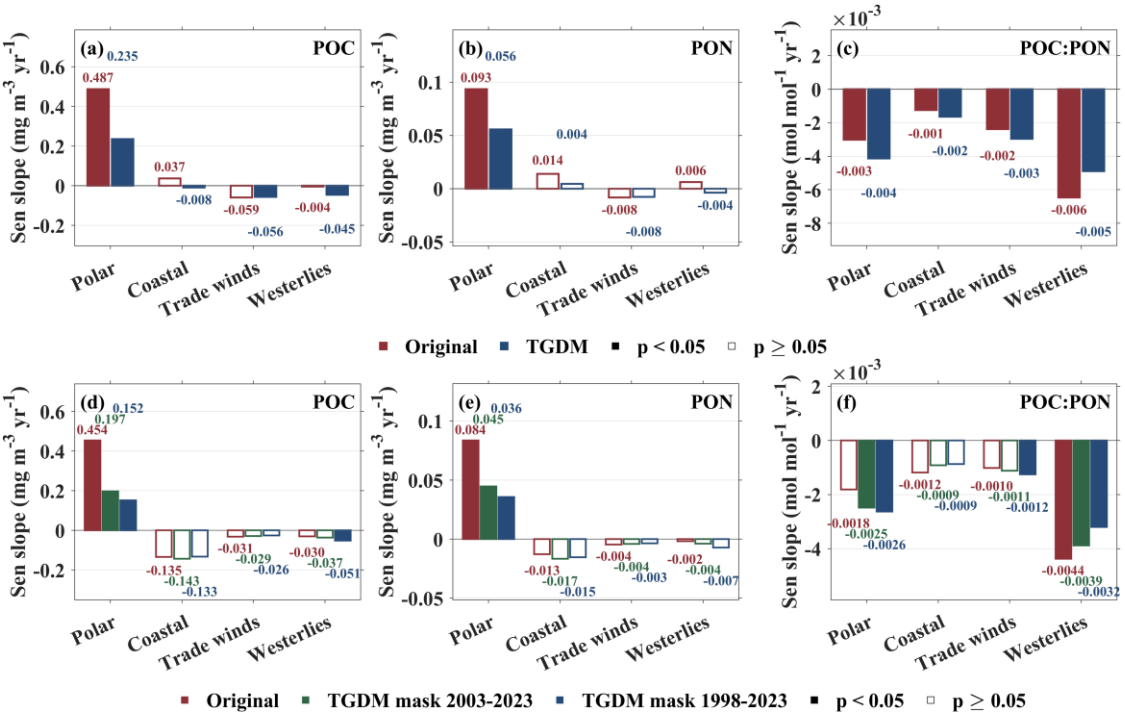


Figure 10. Sen slopes and statistical significance of trends in POC, PON, and POC:PON across the Polar, Coastal, Trade winds, and Westerlies biomes. Panels (a–c) compare the original and TGDM-homogenised estimates over 1998–2023 using the 1998–2023 mask. Panels (d–f) compare trends over 2003–2023 in the original estimates and in the estimates homogenised using TGDM masks derived from 2003–2023 and 1998–2023. Filled bars indicate statistically significant trends ($p < 0.05$), and open bars indicate nonsignificant trends ($p \geq 0.05$).

To assess the sensitivity of regional trends to the reference period used to construct the TGDM mask, we held the trend estimation period fixed at 2003–2023 and compared the original time series with TGDM-homogenised series based on the 2003–2023 and 1998–2023 masks (Fig. 10d–f and Figs. S9 in the Supplement). In the polar biome, both masks reduced the magnitude of the POC and PON increases, while all trends remained statistically significant ($P < 0.05$). The POC slope decreased from 0.454 mg m^{-3} yr^{-1} in the original series to 0.197 and 0.152 mg m^{-3} yr^{-1} using the 2003–2023 and 1998–2023 masks, respectively (Fig. 10d). The corresponding PON slopes decreased from 0.084 to 0.045 and 0.036 mg m^{-3} yr^{-1} (Fig. 10e). In the coastal, trade-wind, and westerlies biomes, the estimated POC and PON slopes were negative in all three series. Notably, the coastal POC and PON slopes were positive in the original 1998–2023 series but negative in the original 2003–2023 series, with neither set of trends reaching statistical significance (Fig. 10a, b, d, e).

Over 2003–2023, POC:PON slopes were negative across all four biomes in all three series, but statistical significance varied with biome and mask choice (Fig. 10f). In the polar biome, the decline was nonsignificant in the original series and significant after TGDM homogenisation with either mask. In the coastal biome, the decline was nonsignificant in all three series, whereas in the trade-wind biome it was significant only with the 1998–2023 mask. In the westerlies, the decline was significant in all three series, although its magnitude decreased from 0.0044 mol $mol^{-1}yr^{-1}$ in the original series to 0.0039 and 0.0032 mol $mol^{-1}yr^{-1}$ with the 2003–2023 and 1998–2023 masks, respectively.

Together, these comparisons show that observational coverage affects estimated concentration levels, seasonal amplitudes, and trend magnitudes. The polar POC and PON increases remained significant

across both analysis periods and all coverage treatments, although TGDM consistently reduced their magnitudes. For POC:PON, the direction of change was consistent across all four biomes, both analysis periods, and all coverage treatments. Statistical significance was less consistent over 2003–2023, with the westerlies being the only biome where the decline remained significant across both periods and all treatments. This consistency in the direction of change supports the use of the GDR POC:PON product to examine regional changes in POM stoichiometry, while the strength of evidence for individual biome trends depends on the analysis period and coverage treatment." (lines 487 to 577) and "

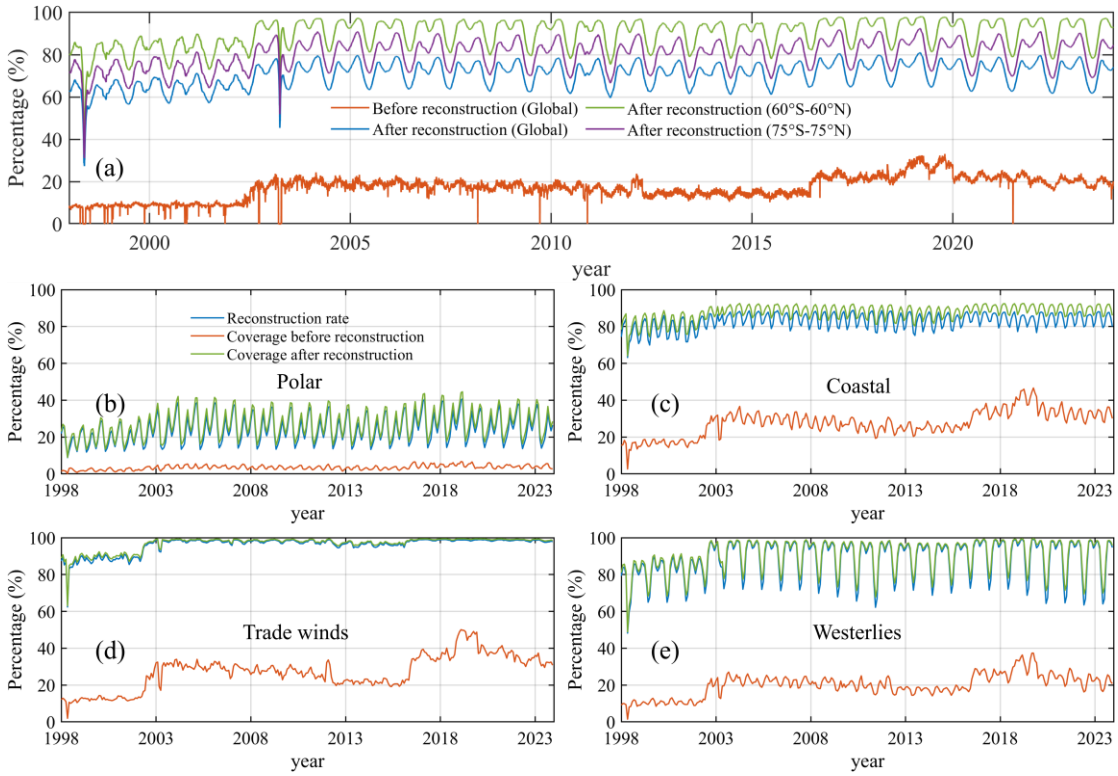


Figure S5. Daily valid-pixel coverage before and after reconstruction of the POC and PON datasets during 1998–2023. (a) Global coverage before reconstruction and coverage after reconstruction for the global ocean, 60° S–60° N, and 75° S–75° N. Panels (b–e) show coverage before and after reconstruction and the reconstruction rate for polar, coastal, trade-wind, and westerly regions, respectively. The legend in panel (b) applies to panels (b–e)." in the Supplement.

2. Reconstruction under realistic gap patterns. To address your concern about the representativeness of random pixel removal on two dates, we expanded the assessment to 16 experiments each for POC and PON. We applied eight representative 31-day observation masks from the SeaWiFS-only and multi-sensor periods to two season-matched target windows each. These masks preserved the spatial clustering and temporal persistence of observational gaps. At masked locations with finite reference values, we compared the recovered center -day values with the corresponding pre-masking values in the previously reconstructed POC and PON fields. The related revised text are as follows.

"**2.4.2 Reconstruction evaluation and uncertainty estimation**

Artificial masking experiments were conducted to evaluate DCT-PLS reconstruction performance and develop empirical models of pixel-level reconstruction uncertainty. The experiments comprised three steps:

1. *Masking selection*. Eight representative 31-day masks were selected from historical OC-CCI

observation patterns over 1998–2023. Five masks represented typical conditions in different periods, and three additional masks captured low observation coverage, persistent gaps, and contrasting spatial patterns of missing observations.

2. *Reconstruction experiments*. Each mask was paired with two target dates from different years in the same calendar month to limit seasonal differences, yielding 16 experiments for each variable. The daily masks were applied to 31-day reference cubes assembled from previously reconstructed POC and PON fields, with the target date at the central slice. The resulting inputs retained any pre-existing gaps and were reconstructed using the same DCT–PLS settings as the final products. Locations with no valid input throughout the masked 31-day window were left missing in the output.
3. *Performance and uncertainty assessment*. Evaluation was restricted to target-day pixels with valid reference values that were artificially masked and successfully reconstructed. The reconstructed values were compared with their values before masking to assess recovery of the reference fields.

The resulting errors were used to fit empirical models of the expected bias and conditional error standard deviation as functions of the distance to the nearest valid input value and the reconstructed concentration class. For each variable, 12 experiments were used for model fitting and four were held out for evaluation, with no overlap in target dates. Appendix B describes the distance calculation, model fitting, and evaluation procedure." (lines 235 to 252)

"3.2.1 Reconstruction performance under realistic missing-data patterns

DCT-PLS reconstruction performance was evaluated through 16 artificial masking experiments. Eight examples covering all eight selected masks are shown in Fig. 4, with the corresponding target dates listed in Table S2 in the Supplement. For POC (Fig. 4a–h), $\log_{10}$-scale R^2 ranged from 0.97 to 0.98, and $\log_{10}$-scale RMSE remained 0.04 across the displayed examples. Linear-scale RMSE ranged from 9.23 to 11.26 mg m^{-3}, and MAPE ranged from 4.47 % to 5.77 %. Regression slopes in $\log_{10}$ space ranged from 0.96 to 0.98, indicating close agreement between the reconstructed and reference POC concentrations.

PON exhibited comparable reconstruction performance in the eight examples shown in Fig. 4i–p. Log_{10}-scale R^2 ranged from 0.97 to 0.98, $\log_{10}$-scale RMSE was approximately 0.04, and regression slopes ranged from 0.96 to 0.98. Linear-scale RMSE ranged from 1.57 to 1.88 mg m^{-3}, while MAPE ranged from 4.54 % to 5.82 %, similar to the MAPE range for POC. Together, these examples demonstrate close agreement with the reference fields and consistent reconstruction of artificially masked POC and PON values across the historical missing-data patterns and target dates examined.

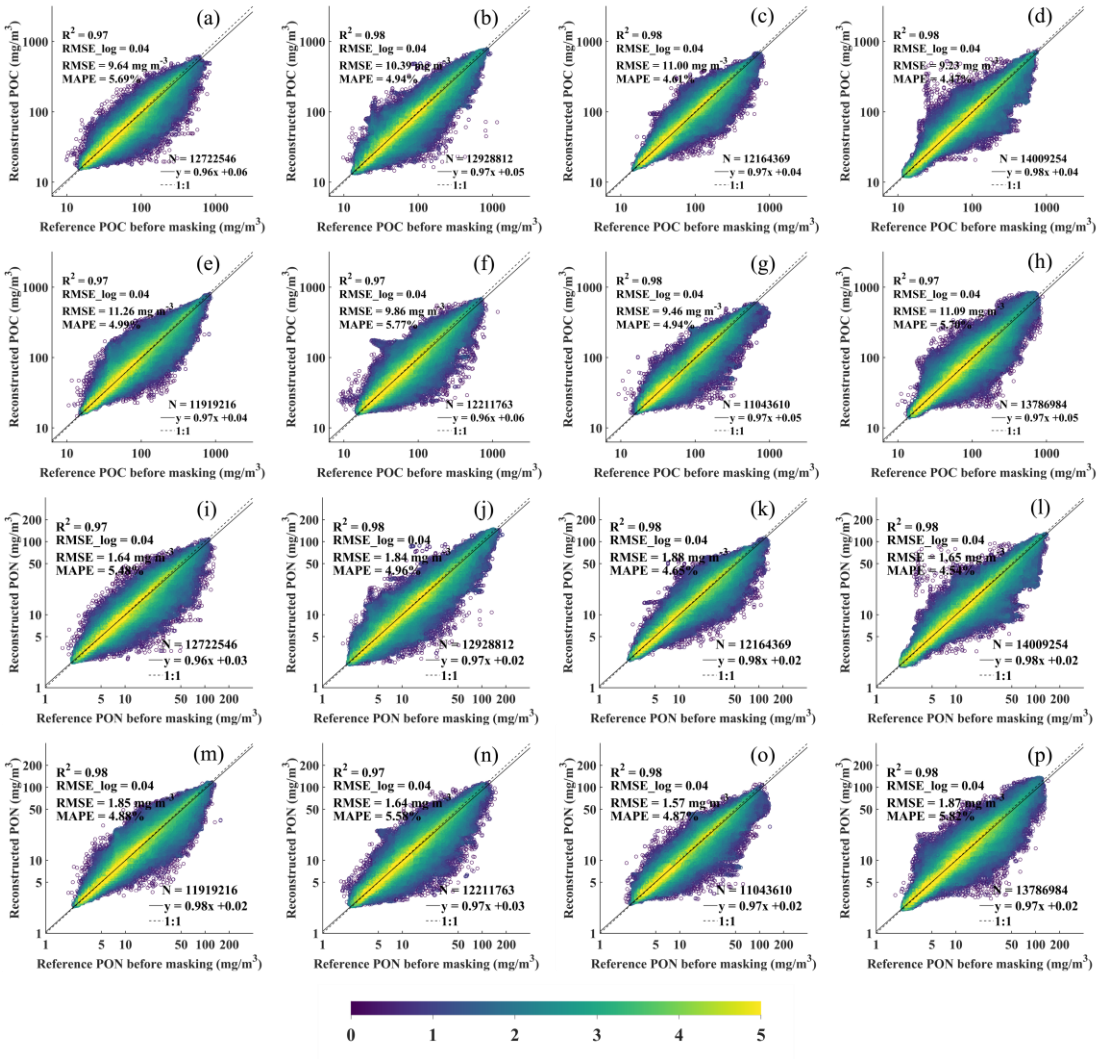


Figure 4. Comparison between reference concentrations before masking and DCT-PLS-reconstructed concentrations for POC (a–h) and PON (i–p) across eight artificial masking experiments. Colours indicate the base-10 logarithm of the number of data points in each two-dimensional bin; colour-bar values of 0, 1, 2, 3, 4, and 5 correspond to 1, 10, 100, 1,000, 10,000, and 100,000 points per bin, respectively." (lines 385 to 400)

3. Identification of retrieved and reconstructed values and their uncertainties. To address your request for pixel-level information, we added a SourceFlag to each NetCDF file to identify pixels with no valid value (0), GPR retrieved values (1), and values reconstructed by DCT-PLS (2). For retrieved pixels, we provide GPR predictive standard deviations and out-of-domain flags where available. To estimate the uncertainty introduced by gap filling at individual reconstructed pixels, we used the errors from the 16 mask experiments described above to model the conditional bias and standard deviation of gap-filling errors. For each of POC and PON, we fitted the model using 12 experiments and evaluated it on four remaining experiments. We then applied the fitted models to provide pixel-level estimates of gap-filling bias and error standard deviation in the released NetCDF files. The related revised text are as follows.

"GPR predictive standard deviations and applicability domain indicators were used to characterise retrieval quality at the pixel level (Appendix A)." (lines 185 to 186)

"The resulting errors were used to fit empirical models of the expected bias and conditional error standard deviation as functions of the distance to the nearest valid input value and the reconstructed concentration

class. For each variable, 12 experiments were used for model fitting and four were held out for evaluation, with no overlap in target dates. Appendix B describes the distance calculation, model fitting, and evaluation procedure.” (lines 249 to 252)

“Appendix A: Predictive uncertainty and applicability domain of GPR retrieved pixels

For each satellite retrieval of POC and PON, predictive uncertainty was quantified as the standard deviation of the corresponding GPR predictive distribution on the $\log_{10}$ concentration scale. For a satellite pixel with a predictor vector X_*, the predictive variance of the response, including observation noise (Rasmussen and Williams, 2006), was calculated as:

$$\sigma^2_{\text{pred}}(X_*) = k(X_*, X_*) + \sigma_n^2 - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*, \quad \text{(A1)}$$

where k is the exponential covariance kernel, $\mathbf{K}$ is the kernel matrix evaluated at the training inputs, and $\mathbf{k}_*$ contains the kernel covariances between X_* and the training inputs. The estimated observation noise variance is denoted by σ_n^2, and $\mathbf{I}$ is the identity matrix. Predictive standard deviations were obtained by taking the square root of Eq.(A1) and were retained on the $\log_{10}$ concentration scale.

On the $\log_{10}$ scale, the POC:PON ratio is the difference between the two log concentrations plus a constant for the conversion from mass to molar units. This constant does not affect the variance. The predictive variance of the $\log_{10}$-transformed ratio was therefore estimated as:

$$\sigma^2_{\log_{10}(\text{POC:PON})} = \sigma^2_{\log_{10}(\text{POC})} + \sigma^2_{\log_{10}(\text{PON})} - 2\rho\sigma_{\log_{10}(\text{POC})}\sigma_{\log_{10}(\text{PON})}, \quad \text{(A2)}$$

where $\sigma_{\log_{10}}(\text{POC})$ and $\sigma_{\log_{10}}(\text{PON})$ are the predictive standard deviations of the $\log_{10}$-transformed concentrations from their respective GPR models. The correlation coefficient ρ was estimated from paired standardized prediction residuals on the $\log_{10}$ scale for samples held out from both retrieval model fits. A separate coefficient was estimated for each of three absolute latitude bands, pooling both hemispheres, and applied uniformly across longitude and time within that band. The predictive standard deviation of the ratio on the $\log_{10}$ scale was obtained as the square root of Eq. (A2).

Distances in predictor space can be used to assess whether prediction conditions are represented in the training data (Meyer and Pebesma, 2021). The applicability domain (AD) was assessed separately for each retrieval model using nearest-neighbour distances in standardized predictor space. The 12 predictors were standardized using the means and standard deviations of the corresponding training dataset. For each satellite pixel, the AD distance was defined as the minimum Euclidean distance to any training sample:

$$d(X_*) = \min_{i=1,\ldots,N}\sqrt{\sum_{j=1}^{12}(z_{*,j} - z_{i,j})^2}, \quad \text{(A3)}$$

where N is the number of training samples, and $z_{*,j}$ and $z_{i,j}$ are the standardized values of predictor j for the satellite pixel and training sample i, respectively. For each model, the AD threshold was set to the 95th percentile of the nearest-neighbor distances calculated for its training samples, with each sample excluded from its own neighbor search. The thresholds were 1.25 for POC and 1.24 for PON. Pixels with distances above the corresponding threshold were flagged as outside the applicability domain (OOD). A POC:PON ratio was flagged as OOD if either constituent retrieval was OOD.

Predictive standard deviations and AD flags were retained as pixel-level quality information for daily satellite retrievals over 1998–2023. The standard deviations quantify model-based predictive uncertainty, and the AD flags indicate where predictor conditions are poorly represented in the training data.

Appendix B: Uncertainty assessment of DCT-PLS reconstructed pixels

The masking experiments described in Sect. 2.4.2 were used to develop empirical models of reconstruction bias and error standard deviation as functions of distance to the nearest available input, with separate models fitted for different reconstructed concentration classes.

Reconstruction errors were calculated relative to the values before artificial masking. These reference fields contained both original retrievals and previously gap-filled values; therefore, the estimated errors characterize recovery of the mixed reference fields rather than deviations from true concentrations. The signed error was defined as $e = \log_{10}(X_{rec}) - \log_{10}(X_{ref})$, where X_{rec} and X_{ref} denote reconstructed and reference concentrations, respectively. Positive errors indicate overestimation.

For each evaluation pixel, the distance d was calculated as the three-dimensional Euclidean distance to the nearest finite input value remaining after masking, using spatial-grid and daily-time indices within the corresponding reconstruction block, including its overlap, over the 31-day window. Samples without available input support were excluded. Samples were classified by reconstructed concentration as low (< 50 mg m^{-3}), medium (50–80 mg m^{-3}, inclusive), or high (> 80 mg m^{-3}) for POC, and low (< 5 mg m^{-3}), medium (5–10 mg m^{-3}, inclusive), or high (> 10mg m^{-3}) for PON. Samples from all eight masks were pooled within each class. Within each concentration class, distance-dependent functions were specified for reconstruction bias and error standard deviation. The expected reconstruction bias was modelled as:

$$\mu_c(d) = \bar{e}_c + s_{e,c}\,(b_{0,c}+b_{1,c}u+b_{2,c}u^2+b_{3,c}u^3),\ u = \frac{\ln(1+d)-m_{l,c}}{s_{l,c}}, \quad \text{(B1)}$$

where c denotes the concentration class; $\bar{e}_c$ and $s_{e,c}$ are the mean and standard deviation of the training errors, and $m_{l,c}$ and $s_{l,c}$ are the mean and standard deviation of ln(1+d) in the training samples. The coefficients $b_{0,c}$–$b_{3,c}$ describe the relationship between bias and standardized log-distance.

The standard deviation describes error variability around the expected bias and was modelled as

$$\sigma_c(d) = \sigma_{0,c}+A_c[1-\exp(-d/L_c)], \quad \text{(B2)}$$

where $\sigma_{0,c}$ is the error standard deviation at zero distance, A_c is the difference between the asymptotic and baseline standard deviations, and L_c is the characteristic distance governing the approach to the asymptote. The constraints $\sigma_{0,c} > 0$, $A_c \geq 0$, and $L_c > 0$ ensure that $\sigma_c(d)$ remains positive and non-decreasing, approaching $\sigma_{0,c} + A_c$ as distance increases.

Model parameters were estimated jointly by minimizing the Gaussian negative log-likelihood. Fitting was repeated 1,000 times per concentration class using up to 10,000 training samples per repetition. Final estimates were obtained by taking the median of each parameter across successful fits. Performance was evaluated on held-out experiments using standardized residuals and interval coverage." (lines 623 to 682)

Comments in manuscript order:

1. Lines 18-21: Please explain what is meant by "consistent". It is unclear whether this refers to the method, to consistency in space and time, or to something else. The manuscript later states that some areas remain unreconstructed when too few satellite observations are available. Please define "global gap-free" when the dataset is first introduced.

Response: Thank you for pointing out this ambiguity. In the original manuscript, "consistent" referred to the use of common ocean-colour satellite and reanalysis data sources and a common processing framework to generate the POC, PON, and POC:PON products. To avoid implying uniform satellite coverage or the absence of temporal biases, we removed the term "consistent". We also removed "gap-free" because some areas remain unreconstructed where satellite observations are insufficient. The revised title and manuscript now use "reconstructed" to describe the products and "reconstruction" to

describe the gap-filling process. The related revised text is "Here, we present a global daily reconstructed dataset of POC, PON and POC:PON (GDR-POM) at 4 km resolution". (lines 19 to 20)

2. Lines 30–32: It is unclear what is meant by "better captures trends and magnitudes" or what the gap-free product is being compared with. Greater completeness does not itself demonstrate greater accuracy. The basis of this claim should be stated or the wording revised.

Response: Thank you for this suggestion. We removed the claim that the reconstructed product "better captures trends and magnitudes". The revised abstract quantifies the improvement in spatial coverage directly. Gap filling increased mean daily valid-pixel coverage over the global ocean from 17.03 % to 70.49 %, relative to the satellite-retrieved fields before reconstruction.

In response to this and related comments, we added a sensitivity analysis in Sect. 4.1 to assess how changes in observational coverage affect trend estimates from GDR-POM. The revised abstract now distinguishes the sensitivity of POC and PON concentration trends from the greater consistency in the direction of the POC:PON trend across the tested coverage treatments. This distinction clarifies how each variable can be interpreted in long-term analyses.

The related revised text in the Abstract is: "Applying the Temporal Gap Detection Method (TGDM) made global POC and PON slopes more negative over 1998–2023, with little effect over 2003–2023. POC:PON declined globally and in each biome examined across both periods and all tested coverage treatments, making it a relatively coverage-insensitive indicator of the direction of long-term change in marine POM stoichiometry." (lines 27 to 31)

3. Lines 36-41: Please define POC, PON and POM at first use in the main text.

Response: Thank you for pointing this out. In the first paragraph of the main text, we now spell out particulate organic carbon (POC) and particulate organic nitrogen (PON) and define POC:PON as the ratio of POC to PON. The related revised text is "Particulate organic carbon (POC) and particulate organic nitrogen (PON) are primarily produced by phytoplankton" (line 35), "Variations in the ratios of POC to PON (POC:PON)" (line 42) and "particulate organic matter (POM)" (line 41)

4. Lines 45-53: Because POC and PON are reported in mg m^-3, the derived POC:PON appears to be a mass ratio. Please state this explicitly, as the Redfield ratio quoted here is a molar ratio.

Response: Thank you for pointing out the distinction between mass and molar ratios. To enable direct comparison with the Redfield ratio, we now express POC:PON as a molar carbon-to-nitrogen ratio (mol mol^{-1}). We convert the mass concentrations of POC and PON (mg m^{-3}) to molar concentrations using the molar masses of carbon and nitrogen, respectively. The molar ratio is therefore (POC/12)/(PON/14)=(POC/PON)*(14/12). This conversion has been applied to the archived GDR-POM dataset and to all POC:PON values reported in the manuscript, including those in the revised Figs. 5–7, 9, 10, and S7–9.

5. Lines 129-134: SeaWiFS is the only contributing sensor before April 2002 and was processed using a different atmospheric correction method. Coverage also changes as other sensors are added to the merged OC-CCI dataset. Please examine whether these changes affect GGFD across regions and time, particularly in coastal and high-latitude waters, and whether they influence the reported trends.

Response: Thank you for highlighting the potential effects of the early SeaWiFS-only period and subsequent changes in OC-CCI coverage. As described in our response to your general comment on

temporal consistency, we added global and biome-level coverage statistics to Fig. S5 in the Supplement and compared trends with and without TGDM over 1998–2023 and 2003–2023 in Sect. 4.1. Global valid-pixel coverage increased around 2003 both before and after reconstruction. The polar biome retained the lowest coverage after reconstruction and showed pronounced seasonal variability.

The trend response differed between polar and coastal waters. Over 1998–2023, TGDM reduced the polar POC slope from 0.487 to 0.235 mg m^{-3} yr^{-1} (Fig. 10a) and the PON slope from 0.093 to 0.056 mg m^{-3} yr^{-1}(Fig. 10b), while both increases remained statistically significant. In coastal waters, POC and PON trends were nonsignificant in both periods across all tested coverage treatments. Over 2003–2023, the polar POC:PON decline was significant with either TGDM mask, whereas the original polar series and all three coastal series showed nonsignificant declines (Fig. 10f).

At the global scale, the estimated POC and PON slopes were more sensitive to TGDM homogenisation when the early SeaWiFS-only years (1998-2002) were included (Fig. 9 and Fig. S7 in the Supplement). Please see our general response above for the relevant text.

These comparisons establish how sensitive the estimated trends are to observational coverage and to the inclusion of the early SeaWiFS-only years.

6. Lines 134-136 and Table 1: The temporal coverage of GLORYS12V1 is given as 2001-2022 in the text but as 1 January 1998-31 December 2023 in Table 1.

Response: Thank you for pointing out this discrepancy. As the GLORYS12V1 record is updated regularly, we have revised both passages to specify the period used in this study. The daily GLORYS12V1 data used in our analysis span 1 January 1998 to 31 December 2023. The related revised text is "GLORYS12V1, data for 1 January 1998–31 December 2023 used in this study" (line 136).

7. Lines 149-150: Please describe how the GLORYS12V1 fields were resampled to the OC-CCI grid, including the interpolation or aggregation method.

Response: Thank you for this comment. We have added the following clarification to Section 2.2.3: "To ensure spatial consistency between the ocean colour data and ocean physical variables, the GLORYS12V1 fields were resampled from their native spatial resolution of 1/12° to the 4-km OC-CCI grid using bilinear interpolation. Before interpolation, potential temperature, salinity, and the Eastward (u_0) and Northward (v_0) current components were averaged with equal weights across the upper 10 depth levels (0.49, 1.54, 2.65, 3.82, 5.08, 6.44, 7.93, 9.57, 11.40, and 13.47 m). Mixed-layer depth (Z_m) and sea surface height (SSH) were interpolated directly onto the 4-km OC-CCI grid." (lines 142 to 146)

8. Section 2.3.1: The information provided is not sufficient to reproduce the GPR models. Please report the transformations and scaling applied to the response and predictor variables, the logarithm base, the kernel and noise model, the hyperparameter optimisation, and the software and version used. Please also explain how predictions were converted back to concentration units.

Response: Thank you for pointing this out. We expanded Sect. 2.3.1 to specify the transformations and scaling of the response and predictor variables, the covariance kernel and noise model, the method used to estimate the hyperparameters, the software version, and the back-transformation of predictions to concentration units. The revised text is: "The matchups were then randomly divided into calibration (80 %) and validation (20 %) subsets (Table S1 in the Supplement). Figure S2 shows the temporal distributions of the matchups in the two subsets. The response variables were the base-10 logarithms of POC and PON concentrations in mg m^{-3}, with no further standardisation. Four predictors were $\log_{10}$-

transformed before standardisation: Chla, $b_{bp}(665)$, $R_{rs}(443)$ and $K_d(490)$. For the final models, all predictors were centred and scaled using their respective means and standard deviations from the full calibration subset. The same scaling parameters were applied to the validation and satellite predictors.

GPR was selected following preliminary comparisons with alternative regression models and implemented using fitrgp in the Statistics and Machine Learning Toolbox in MATLAB R2023a Update 2. The models used a constant basis function, an exponential covariance kernel, and additive Gaussian observation noise with constant variance. The kernel parameters and noise standard deviation were estimated by maximising the marginal log-likelihood using a quasi-Newton method. A separate ten-fold cross-validation was performed within the calibration subset using the same final predictor set in every fold. Both the 80/20 split and the cross-validation folds were formed at the individual-observation level. The final models were fitted to the full calibration subset and evaluated on the validation subset. A prediction z on the $\log_{10}$ scale was back-transformed as $c = 10^z$, with c in mg m^{-3}." (lines 172 to 185)

9. Lines 159-175: Please clarify whether feature selection and any model tuning were carried out before or after the 80/20 split, and how these steps were handled during ten-fold cross-validation. Also state whether the split and cross-validation folds were grouped by cruise, location or sampling period.

Response: Thank you for raising this point. Bio-optical predictor selection was completed on the full matchup dataset before the random 80/20 split. This step included correlation screening, ranking with RFE-XGBoost, and selection of the number of bio-optical predictors using AICc. The ten-fold cross-validation used to compare candidate predictor sets kept the feature ranking fixed across folds. After the 80/20 split, we performed a separate ten-fold cross-validation within the calibration subset using the final predictor set throughout. The final GPR models were then fitted to the full calibration subset and evaluated on the validation subset. Both the 80/20 split and the cross-validation folds were formed at the individual-observation level, without grouping by cruise, geographic location, or sampling period. We clarified this sequence in the revised Sect. 2.3.1.

The revised text is: "Bio-optical feature selection was completed using the full matchup dataset before the calibration–validation split.

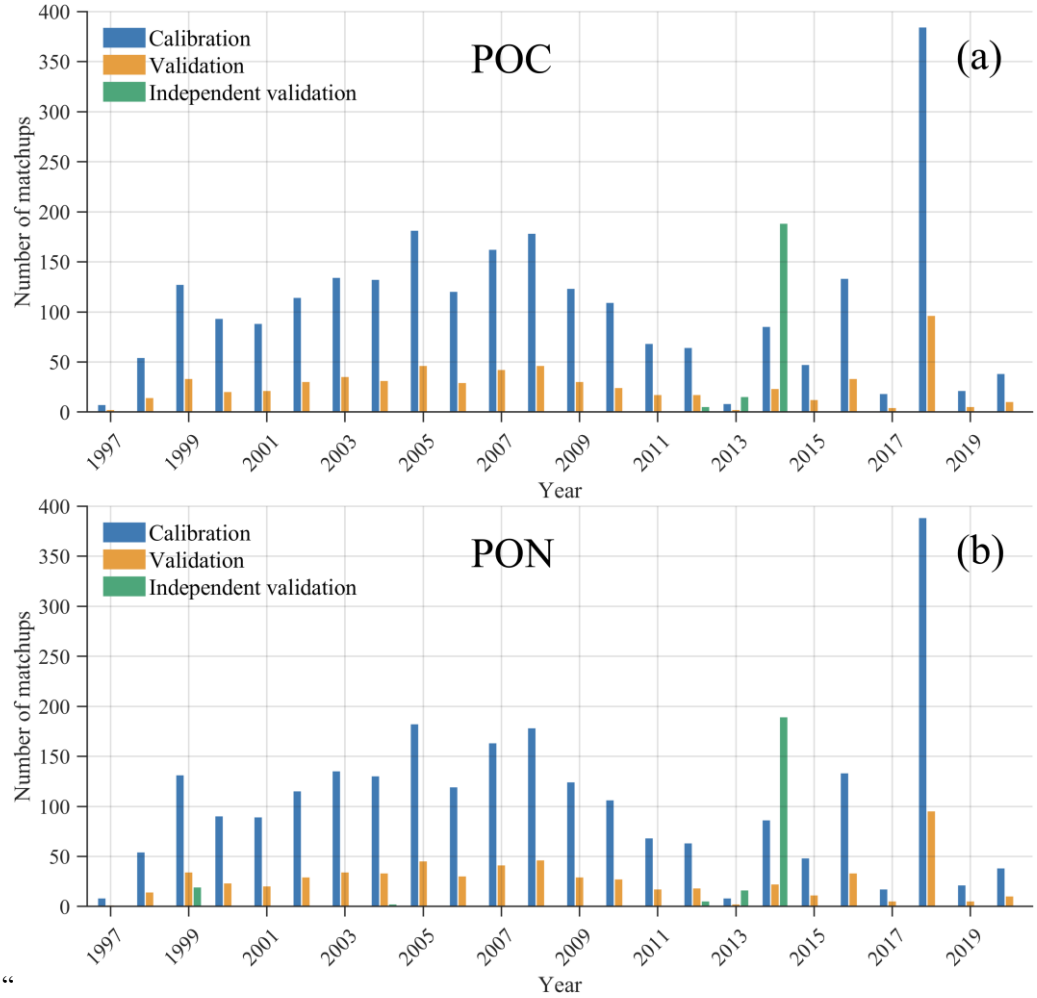
For each bio-optical predictor, the Pearson correlation with the target POC or PON concentration was calculated for both the original and $\log_{10}$-transformed values. The form with the larger absolute correlation coefficient was retained. Predictors with an absolute Pearson correlation coefficient below 0.7 were excluded, and the remaining predictors were ranked by recursive feature elimination using XGBoost (RFE-XGBoost, Zhang et al. (2024)). Candidate GPR models used the same physical predictors and increasing numbers of the highest-ranked bio-optical predictors. These models were evaluated by ten-fold cross-validation using the same feature ranking in every fold. The corrected Akaike information criterion (AICc) was used to select the final number of bio-optical predictors.

The matchups were then randomly divided into calibration (80 %) and validation (20 %) subsets (Table S1 in the Supplement)." (lines 163 to 171) and "A separate ten-fold cross-validation was performed within the calibration subset using the same final predictor set in every fold. Both the 80/20 split and the cross-validation folds were formed at the individual-observation level. The final models were fitted to the full calibration subset and evaluated on the validation subset." (lines 181 to 184)

10. Lines 170-175; Please show how the matchups used for model development and independent validation are distributed across the study period.

Response: Thank you for this suggestion. We have added Fig. S2 to the Supplement to show when the

matchups in the calibration and validation subsets were collected. We have also added a reference to this figure in Sect. 2.3.1: "The matchups were then randomly divided into calibration (80 %) and validation (20 %) subsets (Table S1 in the Supplement). Figure S2 shows the temporal distributions of the matchups in the two subsets." (lines 172 to 173)



"

Figure S2. Annual counts of POC (a) and PON (b) matchups used for model calibration, validation, and independent validation." in the Supplement.

11. Lines 173-175: The cited BCO-DMO dataset ends in July 2010, whereas the listed independent cruises took place in 2012-2014. Is this the correct dataset? If not, what source was used for the independent observations?

Response: Thank you for pointing out this discrepancy. We apologise for the incorrect citation. We confused the sources of the independent cruise observations with the BCO-DMO dataset cited earlier for the in situ matchups. We have verified the data sources for all four cruises and added a dedicated section, "Section 2.6 Independent evaluation", to describe the independent validation experiment. This section provides the download link and cites the relevant data paper or report for each cruise. The related revised text is: "**2.6 Independent evaluation**

To further evaluate retrieval and reconstruction accuracy, an independent in situ dataset that was not involved in feature selection or model development was used to validate both the GPR-retrieved POC and PON concentrations and the final DCT-PLS-reconstructed products. This dataset comprised observations collected during the AE1206, AE1319, NH1418, and AMT24 cruises, with sampling

locations shown as orange circles in Fig. S1a–b in the Supplement. Data from AE1206 and AE1319 were obtained from https://doi.org/10.26008/1912/bco-dmo.538091.2, while NH1418 data were obtained from https://doi.org/10.26008/1912/bco-dmo.829895.1 (Larkin et al., 2021). AMT24 data were obtained from the SeaBASS archive (https://seabass.gsfc.nasa.gov/archive/BIGELOW/BALCH/AMT/archive; Smyth (2014)).” (lines 285 to 292)

12. Section 2.3.1: Was uncertainty estimated for the individual GPR predictions? It is also unclear how users can identify predictions made under conditions that are poorly represented in the training data.

Response: Thank you for this suggestion. We now provide pixel-level GPR predictive standard deviations on the $\log_{10}$ scale and out-of-domain (OOD) flags for satellite retrievals. The NetCDF files contain POC_gpr_ood_flag, PON_gpr_ood_flag, and POC_PON_ratio_ood_flag. A flag value of 1 indicates that the predictor vector exceeded the model-specific distance threshold, and 0 indicates that it fell within the applicability domain. POC:PON was flagged as OOD if either constituent retrieval was OOD. Section 2.3.1 introduces these diagnostics, and Appendix A describes their calculation.

The two indicators are introduced in Section 2.3.1 “GPR predictive standard deviations and applicability domain indicators were used to characterise retrieval quality at the pixel level (Appendix A).” (lines 185 to 186). Appendix A (lines 623 to 654) describes how the predictive standard deviations and applicability domain flags were calculated. The full text of Appendix A is reproduced in our general response above.

13. Lines 195-206: The DCT–PLS settings are not fully described. Which software was used, and how were the smoothing parameter and robust weighting specified? The treatment of the first and last 15 days is also unclear. At the ends of the record, the target day cannot be the central slice of a 31-day window. Which window and slice were used for these dates?

Response: Thank you for identifying these points. We have expanded the Methods section to specify the software, smoothing schedule, and observation weights used for DCT–PLS reconstruction. The related revised text is: “DCT-PLS reconstruction was implemented in MATLAB R2023a with GPU acceleration. The smoothing parameter was decreased from 10^{6} to 10^{-6} in logarithmically spaced steps over 100 iterations. This schedule was prescribed in advance rather than selected by generalized cross-validation. A fixed binary observation mask assigned weights of 1 to valid observations and 0 to missing values. No residual-based robust reweighting was applied.” (lines 222 to 225)

In the original implementation, we used the nearest complete 31-day window containing the target date at each end of the record and extracted the slice corresponding to that date. For the revised product, we reprocessed the first 15 days of January 1998 and the last 15 days of December 2023 using additional input data from December 1997 and January 2024. Every target date now uses a centred 31-day window, with the daily field extracted from slice 16. For example, the window for 1 January 1998 spans 17 December 1997 to 16 January 1998, and the window for 31 December 2023 spans 16 December 2023 to 15 January 2024. The additional inputs were used solely to construct the boundary windows. The released product covers 1 January 1998 to 31 December 2023. The related revised text is: “To centre the 31-day window on every target date, we used additional inputs from 17–31 December 1997 and 1–15 January 2024 for the boundary windows.” (lines 229 to 230)

14. Lines 210-217: The gap-simulation experiment randomly removes 10% of pixels that were originally observed and is limited to two dates in 2020. It therefore does not test reconstruction across large or persistent gaps, or under lower observation densities in parts of the record. Performance should also

be assessed using more realistic gap patterns and periods with different data availability.

Response: Thank you for this suggestion. As described in our response to your general comment on reconstruction under realistic gap patterns, we replaced the two-date random-removal experiment with 16 experiments each for POC and PON, using eight 31-day masks derived from historical OC-CCI observations. Five masks represented typical coverage in different periods, and three captured low observational coverage, persistent gaps, and contrasting spatial patterns of missing observations. Each mask was applied to two target windows from different years, matched by calendar month.

We assessed recovery by comparing successfully reconstructed center-day pixels with their values in the previously reconstructed reference fields before masking. Pixels with no valid input throughout the masked 31-day window remained missing in the output. The revised experimental design and evaluation criteria are described in Sect. 2.4.2 (lines 235 to 252). The corresponding results are presented in Sect. 3.2.1 (lines 385 to 400) of the revised manuscript. Please see our general response above for the relevant text.

15. Lines 218-225 and Table 2: the exact MODIS-Aqua and MULTIOBS products, versions and common comparison period are not given. Please also explain how the daily and weekly data were converted to monthly values and mapped to 1degree, how the common comparison domain was defined, and how many collocated months were required per grid cell.

Response: Thank you very much for your suggestion. We have added the relevant information as requested. The related text is "TCA was applied to compare the relative error characteristics of GDR POC with two global POC products: the MODIS-Aqua Level-3 Monthly standard mapped POC product at a spatial resolution of 4 km, provided by NASA's Ocean Biology Processing Group (http://oceancolor.gsfc.nasa.gov/) and the weekly global ocean 3D biogeochemical MULTIOBS product at a horizontal resolution of 0.25° × 0.25° (MULTIOBS_GLO_BIO_BGC_3D_REP_015_010, https://doi.org/10.48670/moi-00046; Table 2). Monthly GDR POC fields were obtained by averaging the daily values at each grid cell within each calendar month. The MODIS-Aqua monthly composites were used as provided. Weekly MULTIOBS POC values were converted to monthly means by weighting each weekly value according to the number of days the corresponding week overlapped with each calendar month.

All three monthly POC fields from July 2002 to April 2022 were resampled to a regular 1° × 1° grid using bilinear interpolation. The common comparison domain was defined at each grid cell using the temporal intersection of valid observations from all three datasets. Only months with finite POC values in all three datasets were retained, and grid cells with fewer than 12 collocated monthly observations were excluded." (lines 254 to 265)

16. Lines 218-225: TCA assumes that errors in the three products are independent, but this may not hold for this comparison. GGFD is derived from OC-CCI, which includes MODIS-Aqua observations for part of the record, while MODIS-Aqua is also used as a separate comparison product. MULTIOBS also uses satellite ocean-colour data. Please discuss whether the independence assumption is reasonable for these products.

Response: Thank you for highlighting this issue. Because the three products draw on overlapping ocean-colour observations, their errors may be correlated. We revised Sect. 2.5 to explain how the shared inputs could affect the TCA estimates and Sect. 3.4 to clarify the scope of the comparison. We also removed the statement extending the POC findings to PON and POC:PON.

The revised text in Sect. 2.5 is: "The assumption of mutually uncorrelated errors may not fully hold for this product triplet because some observational inputs are shared. GDR POC is derived from OC-CCI ocean-colour data that incorporate MODIS-Aqua observations during part of the record, whereas the comparison MODIS-Aqua POC product is derived from that sensor alone. MULTIOBS also incorporates merged satellite ocean-colour information together with BGC-Argo observations. These shared inputs may introduce cross-correlated errors and affect the TCA estimates despite differences in product-generation methods. Here, TCA was used to compare the relative error characteristics of the three POC products. The results therefore provide a comparative assessment within this product triplet rather than definitive estimates of absolute product accuracy." (lines 277 to 283)

The revised text in Sect. 3.4 is: "Across the grid cells included in the comparison, GDR POC had the lowest median TCA-estimated *fMSE* and the highest median TCA-estimated R^2 of the three products. The distributions overlapped, and MODIS-Aqua had lower estimated *fMSE* than GDR POC in some regions. These results describe the relative error characteristics inferred by TCA for the three products over the common comparison period." (lines 479 to 482)

17. Lines 235-241: Please explain how the regression slopes were converted to % yr^-1 and how uncertainty in the retrieved and reconstructed time series was accounted for. Please indicate which reported trends remain distinguishable from product uncertainty.

Response: Thank you for this comment. We removed the grid-cell trends expressed in % yr^{-1} from Sect. 2.7 (formerly Sect. 2.5). The revised analysis in Sect. 4.1 examines how POC, PON, and POC:PON trend estimates respond to changes in spatiotemporal observational coverage. We now report annual Sen slopes in mg m^{-3} yr^{-1} for POC and PON and in mol mol^{-1} yr^{-1} for POC:PON.

Following the approach described in (Van Oostende et al., 2023), we assessed statistical significance and then compared the magnitude of each significant trend with an error benchmark. We multiplied the absolute annual Sen slope by one year to express the fitted change in the same units as the benchmark. For the full reconstructed record, we used the mean absolute difference between the global spatially weighted medians before and after correcting the gap-filled pixels for their estimated bias. For the retrieval-only record, we calculated the global spatially weighted medians of the lower and upper pixel-level GPR prediction bounds separately at each time point. We then averaged half the difference between these two medians across the time points analysed and used this average as the uncertainty benchmark.

In the full reconstructed record, four significant declines exceeded the estimated effect of gap-filling bias. After TGDM homogenisation, the fitted one-year declines in POC and PON were 0.062 and 0.009 mg m^{-3}, respectively, compared with corresponding benchmarks of 0.027 and 0.004 mg m^{-3}. For POC:PON, the fitted one-year decline was approximately 0.003 mol mol^{-1} both before and after TGDM homogenisation, compared with benchmarks of 0.0006 and 0.0005 mol mol^{-1}, respectively. In the retrieval-only record, the significant declines in POC after TGDM homogenisation and in POC:PON under both coverage treatments were smaller than their respective GPR prediction benchmarks. None of the retrieval-only trends met both screening criteria. In Fig. 9, asterisks (*) mark statistically significant trends for which the fitted one-year decline exceeded the estimated effect of gap-filling bias.

18. Lines 245-250: It is unclear to me how ΔPOC and ΔPON were estimated. Please define these uncertainties and state whether any limits or checks were applied when calculating POC:PON.

Response: Thank you for your suggestion. We have revised Sect. 2.8 to explicitly define ΔPOC and ΔPON as the differences between in situ measurements and the corresponding estimated values, with

retrieval and reconstruction results analyzed separately. These residuals were used to assess their individual first-order contributions to POC:PON errors. In addition, no limits or checks were applied when calculating POC:PON. The related revised text is "First-order error propagation was used to assess how errors in POC and PON affect the derived POC:PON ratio (Lee et al., 2010). Retrieval and reconstruction results were analysed separately using their respective matchups with in situ measurements. For each matchup, signed validation residuals were defined as $\Delta POC = POC_{measured} - POC_{estimated}$ and $\Delta PON = PON_{measured} - PON_{estimated}$. These residuals were used as perturbations in the first-order calculation. Each contribution was calculated using Eq. (8) at the measured concentrations, with the residual of the other concentration set to zero.

$$\Delta CNR = \sqrt{\frac{\Delta POC^2}{PON^2} + \frac{POC^2 \Delta PON^2}{PON^4}}, \quad (8)$$

Each absolute contribution was expressed as a percentage of the measured molar POC:PON ratio. The relative POC contribution was calculated by dividing the absolute POC contribution by the sum of the absolute POC and PON contributions and expressing the result as a percentage." (lines 317 to 325)

19. Lines 259-272 and 385-399 Figures 2, 4 and 7: It is unclear whether R^2, RMSE and MAPE were calculated using the original or log-transformed concentrations. Please give the calculation scale and metric equations. The unit of the POC:PON RMSE at line 388 should be corrected, as the ratio is dimensionless.

Response: Thank you for your suggestion. We have added the equations for R^2, RMSE, and MAPE and explicitly stated whether each metric was calculated using linear-scale or $\log_{10}$-transformed concentrations. The revised text is provided below: "Performance was quantified using the squared Pearson correlation coefficient (R^2), root mean square error on the $\log_{10}$ scale (RMSE_log), root mean square error (RMSE), and mean absolute percentage error (MAPE). For POC and PON, R^2 and RMSE_log were calculated from observed and predicted concentrations on the $\log_{10}$ scale. RMSE and MAPE were calculated on the original concentration scale after back-transforming the predictions. These metrics were defined as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})\right]^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2 \sum_{i=1}^{N}(Y_i-\bar{Y})^2}, \quad (1)$$

$$RMSE_{log} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-X_i)^2}, \quad (2)$$

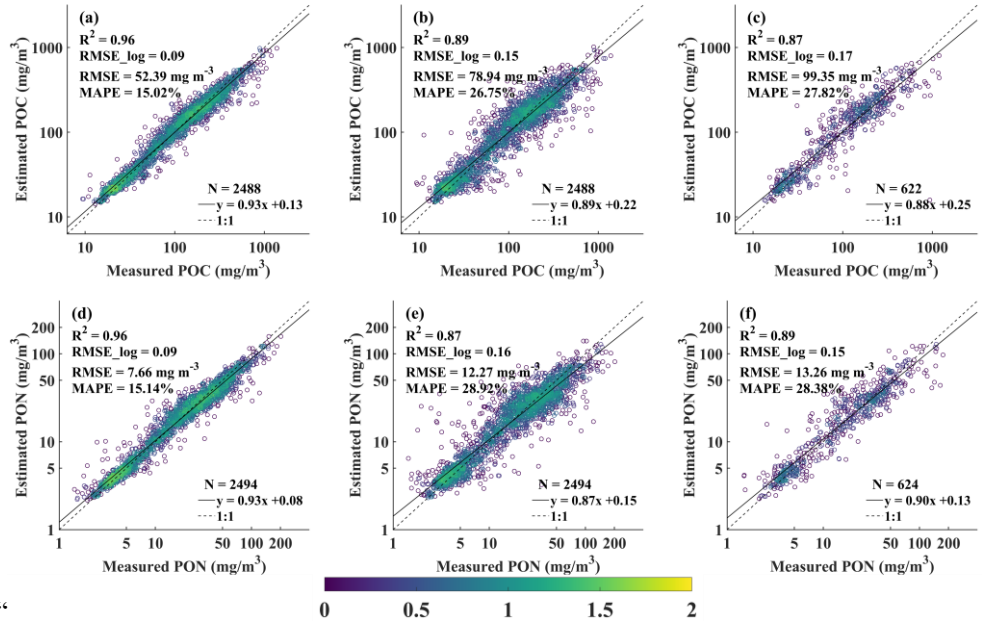
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-x_i)^2}, \quad (3)$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i-x_i}{x_i}\right|, \quad (4)$$

where N is the number of valid observations, and x_i and y_i denote the observed and predicted values, respectively, on the original scale. The $\log_{10}$-transformed concentrations are denoted by $X_i = \log_{10}(x_i)$ and $Y_i = \log_{10}(y_i)$, with respective means $\bar{X}$ and $\bar{Y}$. For POC:PON, R^2, RMSE, and MAPE were calculated using the untransformed molar ratios. The same metrics were used to evaluate gap-filling performance." (lines 189 to 200). We have corrected the POC:PON RMSE notation to mol mol^{-1}, consistent with the linear-scale molar ratios used in its calculation. The related revised text is "The retrieved POC:PON ratio exhibited a lower linear-scale R^2 of 0.43, but the lowest MAPE of 11.69 %, with a linear-scale RMSE of 1.24 mol mol^{-1}." (lines 446 to 447)

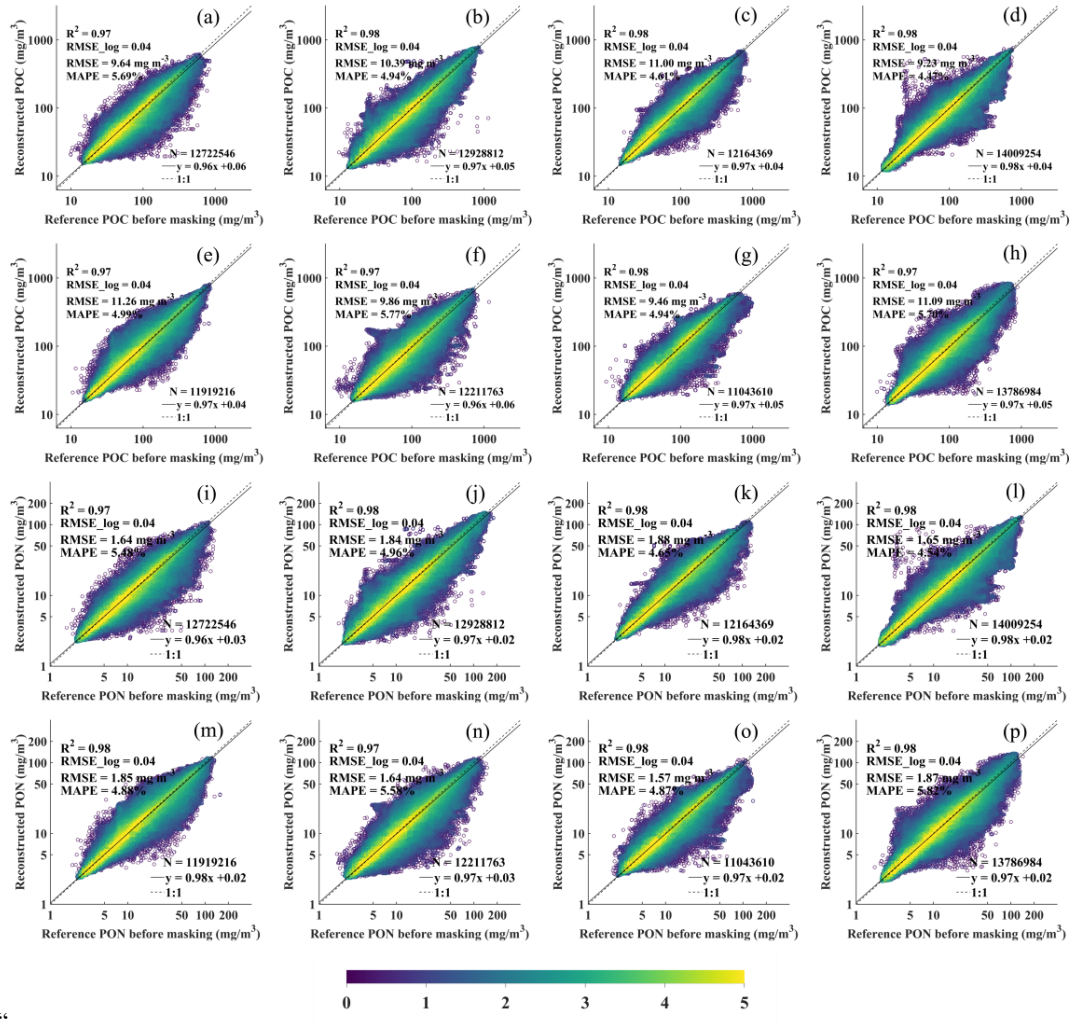
The revised Figs. 2, 4, and 7 are shown below with the calculation scales clarified and the metric labels

and units corrected.



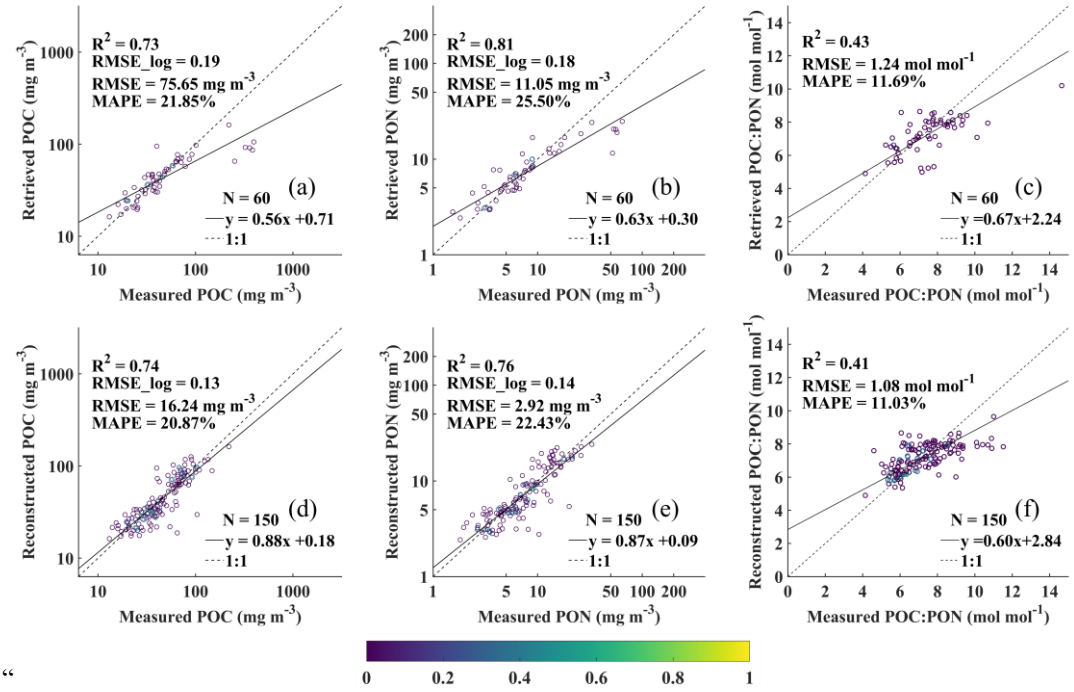
“

Figure 2. Scatterplots of GPR-retrieved POC (a–c) and PON (d–f) against in situ observations: calibration (a, d), 10-fold cross-validation (b, e), and validation (c, f). Colours indicate the base-10 logarithm of the number of data points in each two-dimensional bin; colour-bar values of 0, 1, and 2 correspond to 1, 10, and 100 points per bin, respectively.” (lines 340 to 343)



“

Figure 4. Comparison between reference concentrations before masking and DCT-PLS-reconstructed concentrations for POC (a–h) and PON (i–p) across eight artificial masking experiments. Colours indicate the base-10 logarithm of the number of data points in each two-dimensional bin; colour-bar values of 0, 1, 2, 3, 4, and 5 correspond to 1, 10, 100, 1,000, 10,000, and 100,000 points per bin, respectively.” (lines 396 to 400)



"

Figure 7. Scatterplots comparing satellite-retrieved (a-c) and reconstructed (d–f) POC (a,d), PON (b, e), and POC:PON (c, f) with in situ measurements from an independent validation dataset. Colours indicate the base-10 logarithm of the number of data points in each two-dimensional bin; colour-bar values of 0 and 1 correspond to 1 and 10 points per bin, respectively." (lines 450 to 453)

20. Figures 2, 4 and 7: The colour bar is not explained in Figure 2. In Figures 4 and 7, it is described as point density, but the numerical values are not defined. Please explain what the colour scales represent.

Response: Thank you for this suggestion. We have clarified the colour-scale definition in the captions of Figs. 2, 4, and 7. The colours show the base-10 logarithm of the number of data points in each two-dimensional bin, so a colour-bar value of v corresponds to 10^v points per bin. For example, values of 0 and 2 in Fig. 2 correspond to 1 and 100 points per bin, values of 0 and 5 in Fig. 4 correspond to 1 and 100,000 points per bin, and values of 0 and 1 in Fig. 7 correspond to 1 and 10 points per bin. The revised figures and captions are shown in our response to Comment 19.

21. Table 3: Please state which correlation coefficient is reported and use consistent notation for u0 and v0 across Tables 1 and 3.

Response: Thank you for this suggestion. We have revised the caption of Table 3 to specify that the reported coefficients are Pearson correlation coefficients. We have also standardized the notation for u_0 and v_0 across Tables 1 and 3.

The revised text is "Consistent with the strong positive Pearson correlation between $\log_{10}$(POC) and $\log_{10}$(Chla) (r = 0.90; Table 3)" (lines 353 to 354) and the caption of Table 3 "**Table 3. Pearson correlations of the predictors with POC and PON concentrations.**" (line 383)

22. Lines 343-361: On average, retrieved pixels account for only 24.21% of the pixels available after reconstruction. It would therefore be useful for the archived files to distinguish retrieved from

reconstructed values and to provide some indication of uncertainty or reconstruction quality.

Response: Thank you for this suggestion. As described in our response to your general comment on pixel provenance and uncertainty, each GDR-POM NetCDF file now includes a SourceFlag shared by POC, PON, and POC:PON. The flag identifies missing values (0), satellite-retrieved values (1), and gap-filled values (2). Where available, the files also provide GPR predictive standard deviations and out-of-domain flags for retrieved pixels, and conditional gap-filling bias and error standard deviations for reconstructed pixels.

23. Figure 5: Please state that panels (a–c) show the retrieved fields and panels (d–f) the reconstructed fields.

Response: Thank you for your suggestion. We have revised the caption of Fig. 5 to explicitly state that panels (a–c) show the satellite-retrieved fields and panels (d–f) show the reconstructed fields. The revised caption is as follows: "Figure 5. Global oceanic distributions of POC concentration (a, d), PON concentration (b, e), and POC:PON ratios (c, f) on 1 July 2020. Panels (a–c) show the satellite-retrieved fields, while panels (d–f) show the reconstructed fields." (lines 410 to 411)

24. Figure 6: The red rectangles referred to in the text are not explained in the caption. In addition, the colour scale covers ±10%, whereas the text reports values from about −70% to +125%. Please state whether values outside the displayed range were clipped and explain the choice of colour scale.

Response: Thank you for highlighting these ambiguities. We revised the caption of Fig. 6 to explain that the red rectangles in panel (a) mark areas where data gaps persisted after reconstruction. In panels (d–f), we set the colour limits to −10 % and +10 % because most values fell within that interval and the narrower scale made their spatial variation easier to see. Values outside these limits were shown using the corresponding endpoint colours. The colour limits applied only to the display, and the range reported in the text (approximately −70 % to +125 %) was calculated from the original, unclipped relative differences. The caption was revised as "Figure 6. Monthly mean spatial distributions of reconstructed POC, PON, and POC:PON (a-c) and the corresponding relative differences between reconstructed and satellite-derived fields (d-f) for July 2020. Red rectangles in panel (a) outline areas where data gaps persisted after reconstruction." (lines 423 to 425)

25. Lines 407-430 and Figure 8: Some of the wording in this section, including "significantly lower", "consistently superior" and "clear advantage", seems stronger than warranted by the results shown. No statistical test is reported to support "significantly lower", the GGFD and MULTIOBS distributions overlap, and MODIS-Aqua has lower fMSE in some regions. These statements should either be supported statistically or revised to acknowledge the overlap and regional differences.

Response: Thank you for pointing this out. We have removed expressions such as "significantly lower", "consistently superior", and "clear advantage". The revised text reports the median values of TCA-estimated R^2 and *fMSE* across grid cells and acknowledges that the distributions of the three products overlapped. It also identifies regions where MODIS-Aqua had lower estimated *fMSE* than GDR POC. The revised text is: "The monthly POC fields from GDR, MODIS-Aqua, and MULTIOBS were compared using TCA. The distributions of TCA-estimated R^2 and *fMSE* are shown in Fig. 8a-d. GDR has a median R^2 of 0.86 and a median *fMSE* of 0.14, compared with 0.82 and 0.18 for MULTIOBS and 0.69 and 0.31 for MODIS-Aqua, respectively. The *fMSE* cumulative distribution function (CDF) of GDR POC was generally shifted toward lower values than those of the other two products, with a

corresponding shift toward higher R^2. However, the CDFs overlapped over part of the respective ranges, indicating that the relative performance of the three products was not uniform across all grid cells.

The spatial distributions of R^2 and *fMSE* revealed regional differences among the products (Fig. 8e-j). GDR exhibited lower TCA-estimated *fMSE* across large areas of the global ocean, with most values below 0.5 (Fig. 8e). Within the TCA framework, these values correspond to an estimated signal variance greater than the estimated error variance. MULTIOBS displayed a broadly similar spatial pattern, including elevated *fMSE* in subtropical gyres (Fig. 8g). MODIS-Aqua showed lower *fMSE* in subtropical gyres, particularly in the southern Pacific, but higher values in mid- to high-latitude regions (Fig. 8f). The distributions of R^2 reflected these patterns: GDR and MULTIOBS showed broadly similar spatial structures, with generally higher values for GDR (Fig. 8h and j), whereas MODIS-Aqua exhibited lower values across much of the mid- to high-latitude ocean (Fig. 8i)." (lines 465 to 478)

26. Figure 9: The caption should identify which lines represent the retrieved and reconstructed series.

Response: Thank you for this suggestion. We have revised Fig. 9 to include the TGDM sensitivity analysis and updated the caption to identify the two records by panel column. The left panels show the full reconstructed record, and the right panels show the retrieval-only subset. Within each panel, the coloured and black lines distinguish the original and TGDM-homogenised series.

The revised caption is: "Figure 9. Spatially weighted monthly medians of global oceanic POC, PON, and POC:PON before and after TGDM homogenisation from January 1998 to December 2023. Monthly time series are shown for POC (a, b), PON (c, d), and POC:PON (e, f) in the full reconstructed record (left) and the retrieval-only subset (right). Coloured and black lines represent the original and TGDM-homogenised series, respectively. Dashed lines indicate trends estimated using the Sen slope estimator after removing the mean seasonal cycle." (lines 500 to 504)

27. Lines 448-460; Figures 9 and S4: The difference between the retrieved and reconstructed series changes markedly around 2002-2003, particularly for POC:PON. This coincides with the transition from the SeaWiFS-only period to a multi-sensor record and with the large increase in valid observations shown in Figure S4. As Chl-a is the strongest predictor in both retrieval models, changes in OC-CCI coverage may propagate into GGFD, and gap filling does not necessarily remove such temporal inconsistencies. The authors should examine whether this pattern is related to sensor composition, observation density, upstream processing, reconstruction, or a genuine change. The existing validation supports the accuracy of individual reconstructed values under the tested conditions, but does not establish whether the temporal trends are preserved by the reconstruction or affected by changes in sensor composition and coverage in the underlying OC-CCI record. This should be assessed before the reconstructed trends are described as "more accurate".

Response: Thank you for raising this point. We agree that validation of individual reconstructed values alone cannot establish whether reconstruction preserves long-term trends. As detailed in our response to the earlier comment on temporal consistency, we revised Sect. 4.1.1 to compare trends in the full reconstructed record and the retrieval-only subset over 1998–2023, with and without TGDM homogenisation (Fig. 9, Please see our general response in the beginning for Fig. 9). We repeated the comparison over 2003–2023 using the full reconstructed record (Fig. S7).

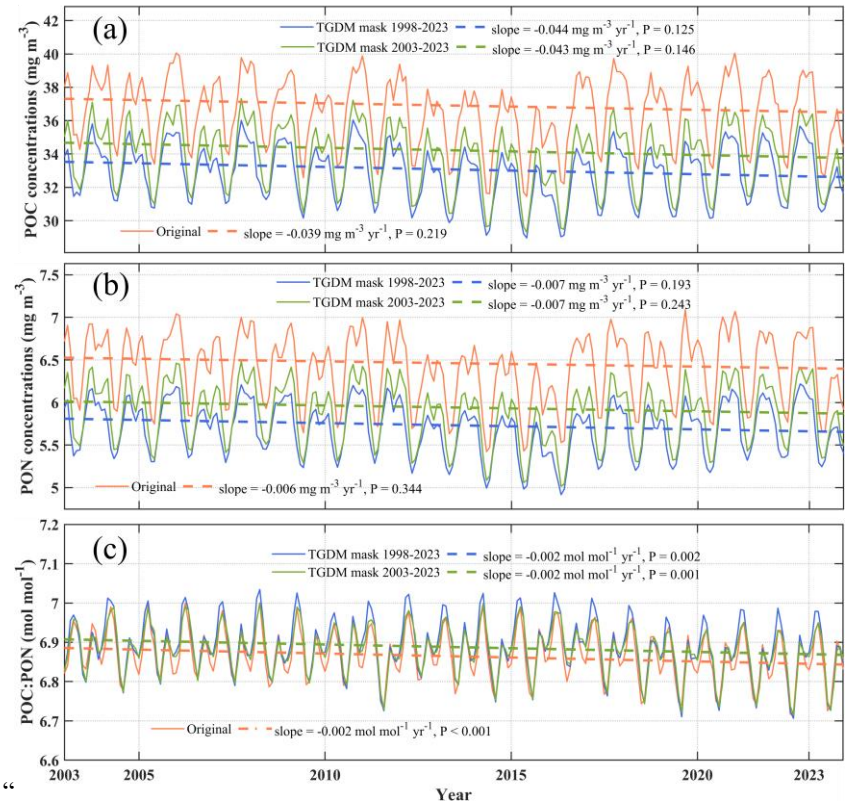


Figure S7. Spatially weighted monthly medians of global oceanic (a) POC, (b) PON, and (c) POC:PON in the full reconstructed record from January 2003 to December 2023. Orange lines show the original series, and blue and green lines show the series homogenised using TGDM masks derived from 1998–2023 and 2003–2023, respectively." in the Supplement.

Global POC and PON slopes were more sensitive to TGDM over 1998–2023 than over 2003–2023, indicating that sensitivity to observational coverage persisted after gap filling. POC:PON showed a more consistent decline across the records, periods, and coverage treatments examined. These comparisons address the sensitivity of the estimated long-term trends to coverage and the inclusion of the early SeaWiFS-only years. Disentangling the effects of sensor composition, observation density, upstream processing, reconstruction, and genuine ocean variability on the full reconstructed and retrieval-only time series around 2002–2003 was beyond the scope of this study.

We removed the description of the reconstructed trends as "more accurate" and revised the manuscript to report the gain in valid-pixel coverage and the trend conclusions supported by the sensitivity analyses.

28. Lines 463-475 and Figure 10: The caption should identify the variable and data type shown in each panel. It should also explain how cells without sufficient data are displayed and distinguish them from cells where a trend was tested but was not significant. The colour scale is limited to ±20% yearr ^-, and many pixels appear to reach these limits. Please state whether values were clipped, report how frequently such large trends occur, and discuss their plausibility overall and in relation to product uncertainty, and, where available, comparable published trends.

Response: Thank you for these detailed comments. As part of the revision of Sect. 4.1, we removed the grid-cell trend maps in the original Fig. 10 and the associated discussion. Section 4.1 now focuses on the sensitivity of POC, PON, and POC:PON trend estimates to changes in spatiotemporal observational coverage. The revised Fig. 10 presents annual Sen slopes for the four Longhurst biomes over 1998–2023

and 2003–2023. The caption identifies the variables, analysis periods, and coverage treatments and explains that filled and open bars indicate statistically significant and nonsignificant trends, respectively.

Data and Suplement

1. The file I inspected contains POC, PON and POC:PON with units and fill values, but does not include a product version or processing history. Providing this metadata and making the GPR and DCT–PLS code available in a persistent repository would make the dataset easier to assess and reuse. If the code cannot be shared, further implementation details are needed to reproduce the workflow.

Response: Thank you for this suggestion. We have added product_version and history as global attributes in the final GDR-POM NetCDF files. The following entries are from one of the files:

"product_version = '1.0.0'

history = '2026-11-09T03:34:54Z merged same-day MAT products and source uncertainty files; no resampling'".

We also archived the fitted GPR models for POC and PON and a MATLAB implementation of DCT–PLS gap-filling on Zenodo (https://doi.org/10.5281/zenodo.24133931). The accompanying README files describe the predictor order, input transformations, and instructions for running the models and reconstruction code.

We expanded Sect. 2.3.1 to document the development of the GPR retrieval models and Sect. 2.4.1 to describe the DCT–PLS reconstruction procedure in more detail. The corresponding revisions are provided in our responses to Comments 8 and 13 above.

2. Figure S1: In panels (a–b), the coloured map background has no scale or explanation and the black and orange symbols are not defined in the caption. In panels (c–d), the colours and outlines of the histogram series are not identified.

Response: Thank you for pointing this out. We have revised the caption of Fig. S1 to explain the coloured map backgrounds and define the black and orange symbols in panels (a–b) The caption also identifies the calibration, validation, and independent validation samples by their histogram colours in panels (c–d). The revised caption is: "Figure S1. Geographical distributions of in situ–OC-CCI matchups for POC (a) and PON (b), and histograms of the corresponding in situ concentrations (c-d). The colored backgrounds in panels (a–b) represent mean reconstructed POC and PON concentrations in 2020, respectively. Black open circles indicate calibration and validation matchups, while orange open circles indicate independent validation samples. In panels (c–d), bars with black outlines represent calibration matchups, and green and orange bars represent validation and independent validation matchups, respectively." in the Supplement.

3. Table S1: Units should be given for all concentration statistics. The maximum PON concentration is reported as 175.56 for the complete dataset but 202.26 for the training subset, which is not possible if the latter is a subset of the former. Please check these values.

Response: Thank you for identifying this inconsistency. We have added units to all concentration statistics in Table S1. We also checked the PON values and corrected the maximum PON concentration in the whole matchup dataset from 175.56 to 202.26 mg m^{-3}. The maximum was 202.26 mg m^{-3} in the calibration subset and 175.56 mg m^{-3} in the validation subset. The revised Table S1 is as:

"**Table S1. Statistical Summary of POC and PON Concentrations in the Matchup Dataset.**

Param	Dataset	N	Min	Max	Mean	Median	Std.

eters			(mg m^{-3})	(mg m^{-3})	(mg m^{-3})	(mg m^{-3})	(mg m^{-3})
POC	Full dataset	3110	9.36	1377.96	156.69	113.13	161.36
	Calibration	2488	9.36	1377.96	155.97	113.98	157.21
	Validation	622	13.42	1191.86	159.60	108.56	177.10
PON	Full dataset	3118	1.44	202.26	25.25	18.59	24.58
	Calibration	2494	1.44	202.26	25.25	18.94	24.15
	Validation	624	1.50	175.56	25.25	16.89	26.26”

4. Figure S2 and Table S2: Figure S2 shows nine red rectangular areas, whereas the caption refers to a single “red diamond-shaped area”. Table S2 should state the year of the 1 January and 1 July results and explain the backslashes used for the Arctic results.

Response: Thank you for pointing out these inconsistencies. As part of our response to your suggestion to use more realistic gap patterns, we replaced the original two-date random-removal experiment with reconstruction experiments based on 31-day masks derived from historical OC-CCI observations. We removed the figure and table labelled Fig. S2 and Table S2 in the original submission. The revised Fig. 4 presents results from the updated reconstruction assessment.

5. Figures S4: The grey and orange series in Figure S4 are not identified.

Response: Thank you for pointing this out. To make the effect of reconstruction on valid-pixel coverage clearer, we replaced Fig. S4 from the original submission with Fig. S5 in the revised manuscript. The new figure shows daily valid-pixel coverage before and after reconstruction of the POC and PON datasets over 1998–2023. The revised figure and caption are shown below.

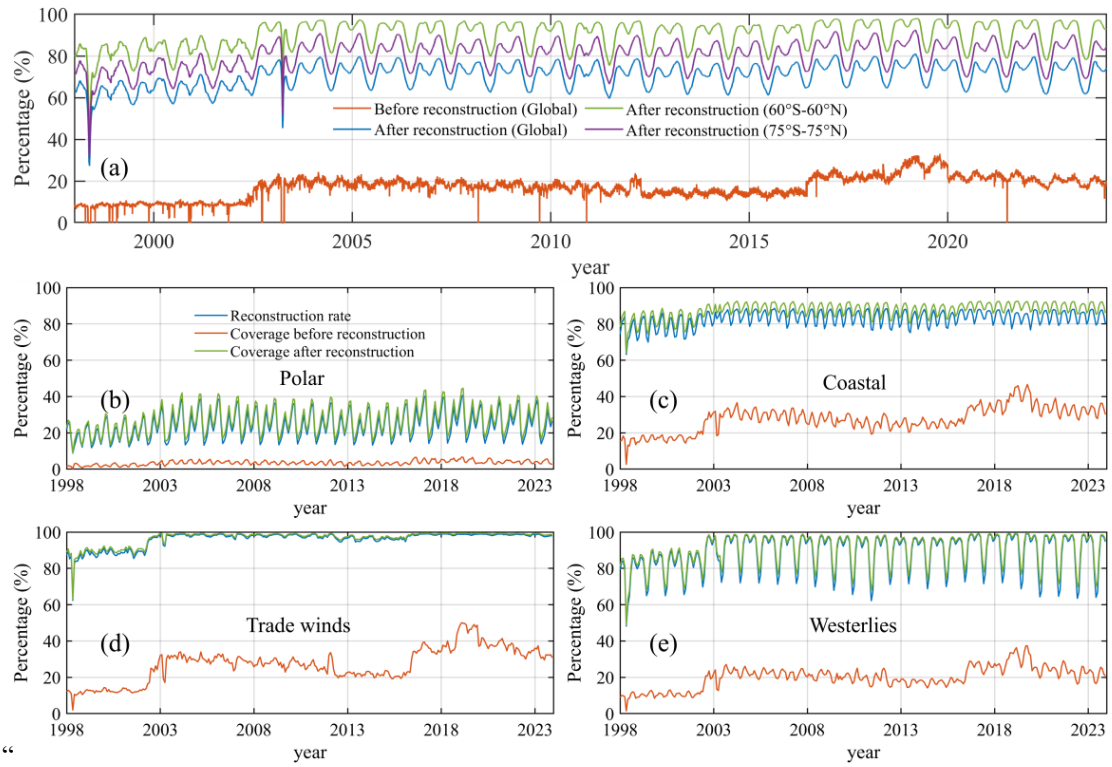


Figure S5. Daily valid-pixel coverage before and after reconstruction of the POC and PON datasets during 1998–2023. (a) Global coverage before reconstruction and coverage after reconstruction for the global ocean, 60° S–60° N, and 75° S–75° N. Panels (b–e) show coverage before and after reconstruction and the reconstruction rate for polar, coastal, trade-wind, and westerly regions, respectively. The legend in panel (b) applies to panels (b–e).”

Reference

Larkin, A. A., Garcia, C. A., Garcia, N., Brock, M. L., Lee, J. A., Ustick, L. J., Barbero, L., Carter, B. R., Sonnerup, R. E., Talley, L. D., Tarran, G. A., Volkov, D. L., and Martiny, A. C.: High spatial resolution global ocean metagenomes from Bio-GO-SHIP repeat hydrography transects, Scientific Data, 8, 10.1038/s41597-021-00889-9, 2021.

Lee, Z., Arnone, R., Hu, C. M., Werdell, P. J., and Lubac, B.: Quantification of uncertainties in remotely derived optical properties of coastal and oceanic waters, Conference on Ocean Sensing and Monitoring II, Orlando, FL, Apr 05-06, WOS:000285624000001, https://doi.org/10.1117/12.849455, 2010.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, 12, 1620-1633, https://doi.org/10.1111/2041-210X.13650, 2021.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA2006.

Smyth, T. J.: AMT24 Cruise Report, Plymouth Marine Laboratory, 2014.

van Oostende, M., Hieronymi, M., Krasemann, H., and Baschek, B.: Global ocean colour trends in biogeochemical provinces, Volume 10 - 2023, 10.3389/fmars.2023.1052166, 2023.

van Oostende, M., Hieronymi, M., Krasemann, H., Baschek, B., and Röttgers, R.: Correction of inter-mission inconsistencies in merged ocean colour satellite data, Volume 3 - 2022, 10.3389/frsen.2022.882418, 2022.

Zhang, Y., Liu, H. Z., Wang, F., Zhu, P., Zhang, Z. X., Wang, Y. R., Wang, Y. Q., Wu, G. F., and Li, Q.

Q.: Toward Applicable Retrieval Models of Oceanic Particulate Organic Nitrogen Concentrations for Multiple Ocean Color Satellite Missions, IEEE Trans. Geosci. Remote Sensing, 62, https://doi.org/10.1109/tgrs.2024.3447699, 2024.