https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

OMNIE-Nutrients: A Global Gridded Ocean Nutrient Dataset (1965-2022) Derived from a Neural Network Emulator and In-Situ Observations

Said Ouala[1,2] and Zouhair Lachkar[3]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France
[2]ODYSSEY Team-Project, INRIA, Brest, France
[3]Mubadala Arabian Center for Climate and Environmental Sciences, NYUAD Research Institute, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Correspondence: Said Ouala (said.ouala@imt-atlantique.fr)

Abstract. Climate-driven changes in upper-ocean stratification are increasingly restricting the supply of marine macronutrients (nitrate, phosphate, and silicate) to the surface, profoundly altering ocean productivity and the biological carbon pump. Quantifying the regional patterns and temporal evolution of these nutrient inventories relies on spatio-temporal interpolation tools to fill massive gaps in historical observational coverage. However, this task is highly challenging due to severe data sparsity, and classical interpolation methods often apply heavy smoothing that obscures fine-scale biogeochemical variability. In this work, we develop OMNIE-Nutrients, a novel global gridded macronutrient product generated by integrating direct in situ nutrient observations with machine-learning-emulated estimates derived from temperature and salinity profiles. The final gridded fields are produced through the optimal interpolation of both observed and emulated data onto a 1° horizontal grid across 65 depth levels. This approach yields two distinct products: a yearly dataset spanning the historical record (1965–2022) to resolve decadal variability, and a monthly dataset tailored to the Argo era (2005–2022) to resolve seasonal dynamics. The neural network emulator demonstrates high accuracy on independent test regions (correlations > 0.96, with RMSE below 1.8 μmol/kg for nitrate, 0.2 μmol/kg for phosphate, and 1.0 μmol/kg for silicate). While the resulting gridded product shows excellent large-scale agreement with the WOA23 baseline climatology, it significantly improves upon it by expanding the effective observational network by over an order of magnitude. Consequently, OMNIE-Nutrients successfully suppresses the severe interpolation artifacts and unrealistic deep-ocean seasonal noise that heavily limit conventional climatologies due to data sparsity. Furthermore, by expanding the effective observational network, our methodology significantly reduces mapping uncertainties in historically data-sparse regions. This study highlights the potential of fusing machine learning with classical objective analysis to reconstruct high-resolution, long-term marine biogeochemical fields, thereby enabling analyses of ocean biogeochemical variability over timescales ranging from seasonal to multidecadal and across broad regions of the global ocean. The OMNIE-Nutrients datasets are freely available at (Ouala and Lachkar (2026b); https://doi.org/10.5281/zenodo.22205370).

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

1 Introduction

Marine macronutrients, specifically nitrate, phosphate, and silicate, act as the primary fuel for ocean productivity and dictate the efficiency of the biological carbon pump (Falkowski et al., 1998; Moore et al., 2013). As global temperatures rise, enhanced upper-ocean stratification is increasingly restricting the physical supply of these essential nutrients to sunlit surface waters (Li et al., 2020; Kwiatkowski et al., 2020). Such physical disruptions carry severe consequences for marine food webs (Lotze et al., 2019) and the ocean's long-term capacity to sequester atmospheric carbon (Boyd et al., 2019). Consequently, tracking the global distribution and temporal evolution of nutrient inventories is a pressing priority for climate science. However, accurately quantifying these large-scale biogeochemical shifts is fundamentally bottlenecked by the historical scarcity of direct *in situ* measurements.

While the recent deployment of the Biogeochemical-Argo (BGC-Argo) fleet marks a revolutionary step forward in ocean monitoring (Claustre et al., 2020), direct nutrient sampling remains far too sparse to fully capture global spatio-temporal dynamics. The current and legacy observational record is heavily dominated by shipboard surveys, which leave massive geographical voids, particularly in the remote Pacific, the Indian Ocean, and polar regions (Garcia et al., 2024). Furthermore, harsh maritime winter conditions have historically created severe seasonal sampling biases. To bridge these gaps, researchers typically rely on objective analysis to create gridded climatologies. Yet, conventional mapping techniques often apply heavy spatial smoothing, which can obscure critical biogeochemical variability in unobserved regions, particularly at seasonal and finer spatiotemporal scales.

Over the past few years, data-driven methodologies have emerged as powerful alternatives to standard tools used in geosciences in general, achieving breakthrough performance in forecasting applications (Pathak et al., 2022; Lam et al., 2023; Dueben et al., 2022; Chen et al., 2023; Bi et al., 2023; Nguyen et al., 2023; Wang et al., 2024; Aouni et al., 2024; Durand et al., 2026), reduced order modeling (Brunton et al., 2016; Ouala et al., 2020, 2023a, b), climate model components emulation for climate simulation applications (Rasp et al., 2018; Reichstein et al., 2019; Beusch et al., 2020; Watson-Parris et al., 2022; Nguyen et al., 2023; Ouala et al., 2024; Kochkov et al., 2024), data-assimilation (Brajard et al., 2019; Ouala et al., 2018, 2023c; Cheng et al., 2023), and state emulation from observations (Sauzède et al., 2017; Bittig et al., 2018; Fourrier et al., 2020; Du et al., 2026). In the context of marine biogeochemistry, pioneering efforts have successfully utilized machine learning to synthesize global gridded fields of dissolved oxygen (Sharp et al., 2022; Ito et al., 2024; Ouala et al., 2026). However, the reconstruction of global macronutrient fields has received comparatively little attention, with works mainly targeting regional reconstructions of the Southern Ocean (Liniger et al., 2025; Zhong et al., 2025).

In this work we aim at addressing this gap, we present a new, long-term gridded macronutrient dataset spanning 1965 to 2022. Our approach fuses quality-controlled *in situ* nutrient profiles with neural-network-derived synthetic estimates from hydrographic data. This methodology allows to significantly expand the effective observational network which in turn, reduces smoothing and interpolation artifacts when using optimal interpolation (OI). The resulting products, collectively termed OMNIE-Nutrients (Optimally Mapped Nutrients from In-situ and Emulated data), comprise OMNIE-PO_4, OMNIE-NO_3, and

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

OMNIE-Si for phosphate, nitrate, and silicate, respectively. The products exhibit robust agreement with established climatologies while reducing mapping uncertainties.

2 Materials and Methods

We construct a gridded dissolved-nutrient concentration product by combining nutrient observations with emulated estimates derived from temperature, salinity, and dissolved oxygen profiles. The methodology follows the approach used to produce the dissolved-oxygen product (OUALA et al., 2025) described in (Ouala et al., 2026), and proceeds in three steps: (i) building quality-controlled training dataset, (ii) training a neural network to emulate nutrient profiles from temperature, salinity, and dissolved oxygen, and (iii) combining observed and emulated data through optimal interpolation (OI) to produce the final gridded product. Similarly to the product in (Ouala et al., 2026), we construct two gridded products at different temporal resolutions: (i) a yearly product spanning 1965–2022, suited to decadal and climate-change-related variability, and (ii) a monthly product spanning 2005–2022 (the Argo era), suited to seasonal and interannual variability. Each step is detailed below.

2.1 Data

The data used in this study were obtained from the *World Ocean Database* (WOD). We use measurements of dissolved oxygen (DO), nitrate ($\mathrm{NO_3^-}$), phosphate ($\mathrm{PO_4^{3-}}$), silicate ($\mathrm{Si(OH)_4}$), temperature (T), and salinity (S), collected from a wide range of observational platforms and contributing institutions. While the WOD23 dataset includes both nitrate ($\mathrm{NO_3^-}$) and nitrate+nitrite ($\mathrm{NO_3^-} + \mathrm{NO_2^-}$, sometimes referred to as NO_x) observations, we follow the World Ocean Atlas 2023 (WOA23) conventions and refer to both observation types simply as "nitrate" (Garcia et al., 2024). This approach is justified because the analytical distinction between these measurements is not always unambiguous, particularly for historical observations, and because nitrite concentrations are generally negligible in the open ocean outside of oxygen minimum zones.

2.1.1 Quality control

Standard WOD quality control (QC) is applied to all profiles, retaining only observations flagged as accepted (flag 0). Oxygen profiles receive additional QC following Ouala et al. (2026). Specifically, profiles with a maximum–minimum difference below $5\ \mu\mathrm{mol/kg}$, or with differences below $0.5\ \mu\mathrm{mol/kg}$ across 18 depth levels, were excluded, as were profiles with surface concentrations below $100\ \mu\mathrm{mol/kg}$. Furthermore, profiles exceeding 115% saturation below 200 m depth, and profiles below 90% of expected surface saturation, were discarded.

For the nutrient variables, we also implement a multi-step QC procedure to ensure physical consistency and remove erroneous measurements. First, we apply a spatial climatology check using WOA23. Specifically, individual measurements at depths greater than 500 m that deviate from the WOA23 climatological mean by more than 50% are flagged as outliers and masked. We also discard any profile retaining fewer than three valid measurements. Finally, we perform a selective spike removal by evaluating vertical gradients. Depth intervals exhibiting vertical gradients exceeding the 99th percentile of the global dataset are identified as anomalous, and a localized window of points around the spike is masked.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



2.1.2 Training and mapping datasets

The quality-controlled measurements are used to build two distinct datasets, one for training the machine learning model, and another for generating the gridded nutrient products. These datasets are detailed below.

2.1.3 ML training dataset.

This dataset consists of co-located T, S, DO, NO_3^-, PO_4^{3-}, and $Si(OH)_4$ profiles spanning 1965–2022. We partition these data into training, validation, and test subsets. The test set comprises 23 independent $1^\circ \times 1^\circ$ regions distributed across all major ocean basins (North and South Pacific, North and South Atlantic, and Indian Ocean). These regions were selected to capture a broad diversity of biogeochemical regimes, including oxygen minimum zones, nutrient-rich regions, and well-oxygenated waters. Test-region locations and emulator performance are shown in Fig. 2. The remaining data are randomly split 80/20 into training and validation sets.

2.1.4 Augmented mapping dataset.

Once the ML model is trained and calibrated, we construct the dataset used for objective mapping by combining direct WOD nutrient observations with emulated samples generated by the neural network. To prevent the inclusion of out-of-distribution artifacts, all emulated samples undergo an additional QC step (Sect. 2.3) ensuring they remain within the observed range of variability.

Figure 1 details the number of observed and emulated profiles retained after this final QC. Overall, the inclusion of emulated profiles substantially improves the spatial sampling density of nutrients by over an order of magnitude across all ocean basins. This significantly mitigates historical and seasonal sampling biases, particularly in the Arctic and Southern Oceans. Furthermore, an analysis of the total number of observations for each nutrient (Appendix C: Fig. C5, Fig. C6, and Fig. C7 for nitrate, phosphate, and silicate, respectively) demonstrates that emulated profiles become especially necessary during periods with a complete absence of observations. Notable examples include a total lack of phosphate observations in the Indian Ocean from 2010 to 2014 across all seasons, as well as a similar absence of phosphate and silicate observations in the Arctic from 2020 to 2024. This marked improvement in global coverage is further illustrated by comparing the spatial distribution of the original observations (Appendix C, Fig. C8) with the augmented dataset containing the emulated profiles (Fig. C9).

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

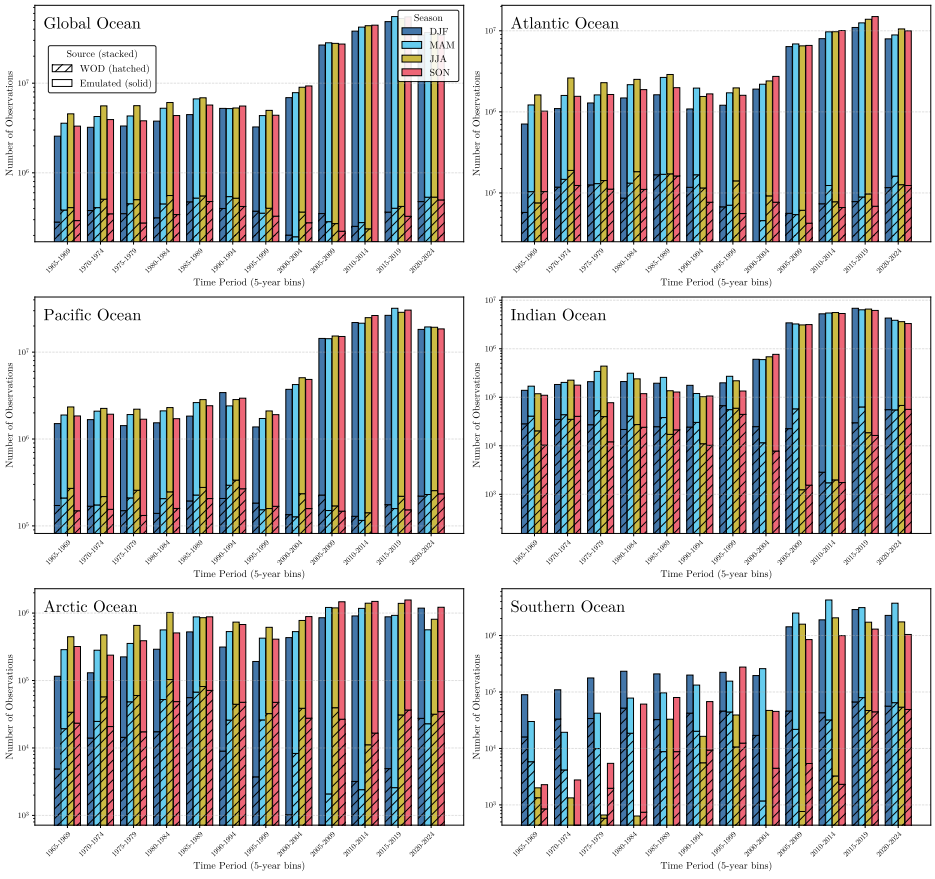


Figure 1. ***Data sampling by ocean region and season.*** Each panel corresponds to a different oceanic region. Hatched bars represent real observations, while solid bars indicate the contribution of emulated observations. Counts are shown on a logarithmic scale and are summed across all nutrients, depths, months, and grid cells within basin mask. The emulated nutrient data increase sampling coverage by one to two orders of magnitude across the different ocean basins.

2.2 Neural network emulator

2.2.1 Model architecture

Nutrient distributions are modeled as a nonlinear function of the physical ocean state, dissolved oxygen, and spatiotemporal coordinates. Formally, the target vector representing nutrient concentrations is defined as

$$\mathbf{c}(t,x,y,z) = \begin{bmatrix} \mathrm{NO_3^-}(t,x,y,z) \\ \mathrm{PO_4^{3-}}(t,x,y,z) \\ \mathrm{Si(OH)_4}(t,x,y,z) \end{bmatrix}. \tag{1}$$

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



We use a neural network $\mathcal{G}_\theta$, parameterized by weights θ, to approximate these concentrations as

$$\mathbf{c}(t,x,y,z) \approx \mathcal{G}_\theta\big(\mathbf{v}(t,x,y,z)\big), \tag{2}$$

where the input feature vector is given by

$$\mathbf{v}(t,x,y,z) = \Big[T(t,x,y,z),\ S(t,x,y,z),\ \mathrm{DO}(t,x,y,z),\ z,\ \phi,\ \lambda,\ m\Big]^T. \tag{3}$$

Here, z is depth, ϕ and λ represent latitude and longitude, and $m \in \{1,\ldots,12\}$ indicates the calendar month. To preserve periodicity, longitude (mapped to $[0°, 360°)$) and month are transformed using cyclical sinusoidal encodings as follows:

$$\lambda \longrightarrow \left(\sin\left(\frac{2\pi\lambda}{360}\right), \cos\left(\frac{2\pi\lambda}{360}\right)\right), \qquad m \longrightarrow \left(\sin\left(\frac{2\pi m}{12}\right), \cos\left(\frac{2\pi m}{12}\right)\right). \tag{4}$$

Our primary objective is to reconstruct nutrient distributions globally using only temperature and salinity profiles, which allows to leverage the spatiotemporal coverage of modern hydrographic platforms (particularly during the Argo era). Since direct DO observations remain comparatively sparse, strictly requiring them as inputs would severely constrain the spatiotemporal reach of the model. To overcome this limitation, we employ, as proposed in Ouala and Lachkar (2026a) a cascaded architecture where the DO input is substituted with estimates from an auxiliary neural network trained exclusively on T and S. As demonstrated by Ouala and Lachkar (2026a), this cascaded architecture is accurate and exhibits a desirable error-compensation property that helps mitigating errors in the input temperature profiles. Details regarding model hyperparameters and training protocols are provided in Appendix A.

2.2.2 Performance of the ML model

We evaluate the generalization and predictive skill of the trained model using 23 independent test regions spanning all major ocean basins and representing diverse biogeochemical provinces and nutrient regimes.. Figure 2 presents scatter plots comparing emulated versus observed concentrations across all depths for nitrate, phosphate, and silicate (left to right). The geographic distribution of the test regions is mapped within the silicate subplot. Overall, the neural network predictions demonstrate high accuracy, achieving root-mean-square errors (RMSE) below $1.8\ \mu\mathrm{mol/kg}$ for nitrate, $0.2\ \mu\mathrm{mol/kg}$ for phosphate, and $1.0\ \mu\mathrm{mol/kg}$ for silicate. Correlation coefficients consistently exceed 0.96 across all three nutrient variables. Furthermore, a comparisons of the averaged vertical profiles derived from *in situ* observations and ML emulations is given in the Appendix C, Fig. C1, Fig. C2, and Fig. C3 for nitrate phosphate and silicate respectively. These comparisons also demonstrate that the model accurately reconstructs the observed vertical structures and effectively captures regional variability for all nutrients.

2.3 Quality control of emulated data

The trained ML model can occasionally generate outliers, particularly when inferring nutrient concentrations in regimes poorly represented within the training data (i.e., out-of-distribution samples). These anomalies can significantly bias the final interpolated product, especially in regions where emulated profiles are dense but *in situ* observations remain sparse.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

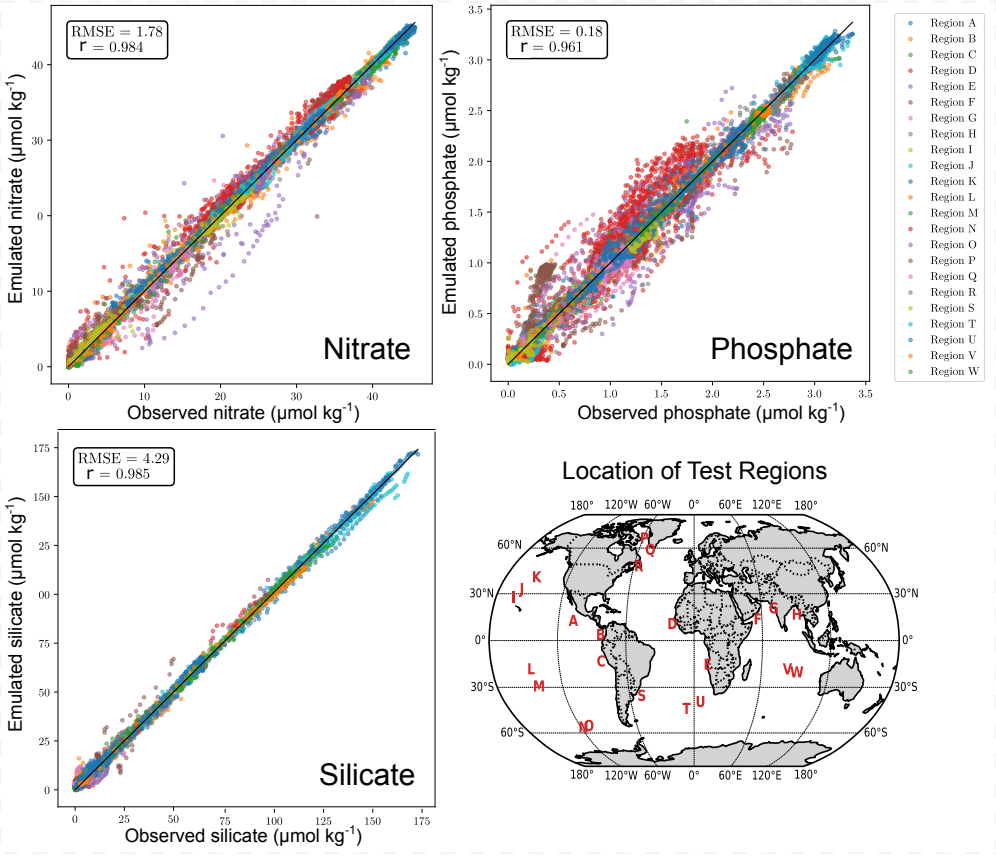


Figure 2. ***Emulated versus ground-truth nutrient concentrations.*** Scatter plots show nitrate, phosphate, and silicate concentrations across all depths within the independent test regions (from left to right and top to bottom). The bottom-right panel shows the locations of the test regions.

To mitigate this risk, we implement a QC framework to validate all emulated predictions prior to interpolation. Validity thresholds are derived directly from the historical distribution of *in situ* nutrient measurements. Specifically, historical observations within each biogeochemical province, delineated using the REgional Carbon Cycle Assessment and Processes Phase 2 (RECCAP2) region masks (Fay and McKinley, 2014; Laruelle et al., 2018; Resplandy et al., 2024), are aggregated to construct climatological minimum and maximum threshold fields (Appendix C, Fig. C4). During validation, every emulated prediction is evaluated against the corresponding local threshold field. Values falling outside the historically observed, biogeochemically plausible range are flagged as anomalous and strictly from the final interpolation step.

2.4 Optimal interpolation

Observed and emulated nutrient data are interpolated using a standard Optimal Interpolation (OI) method. As discussed in the data section, we construct two gridded products at different temporal resolutions: (i) a yearly product spanning 1965–2022,

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



suited to decadal and climate-change-related variability, and (ii) a monthly product spanning 2005–2022 (the Argo era), suited to seasonal and interannual variability. Both use 1° horizontal resolution, which is standard among objectively analyzed in-situ products (Cheng et al., 2017; Ito, 2022) and ML-based reconstructions of *in situ* data (Sharp et al., 2022; Ito et al., 2024), and 65 standard WOD depth levels from 0–2000 m, matching the typical vertical extent of Argo floats (Roemmich et al., 2009).

2.4.1 General formulation

Quality-controlled and emulated data are binned onto the global $1^\circ \times 1^\circ$, 65-level grid, then interpolated with OI to produce the analysis field and its uncertainty:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b), \tag{5a}$$

$$\mathbf{\Sigma}_a = \mathbf{\Sigma}_b - \mathbf{K}\mathbf{H}\mathbf{\Sigma}_b, \tag{5b}$$

where $\mathbf{x}_a$ and $\mathbf{\Sigma}_a$ are the analyzed field and its covariance, $\mathbf{x}_b$ and $\mathbf{\Sigma}_b$ are the background field and covariance, $\mathbf{y}$ are the binned observations, and $\mathbf{H}$ is the observation operator. The Kalman gain is

$$\mathbf{K} = \mathbf{\Sigma}_b\mathbf{H}^\top \left(\mathbf{H}\mathbf{\Sigma}_b\mathbf{H}^\top + \mathbf{\Sigma}_o\right)^{-1}, \tag{6}$$

with $\mathbf{\Sigma}_o$ the observation error covariance matrix.

2.4.2 Background construction

The background mean field $\mathbf{x}_b$ is obtained through a standard OI of all available profiles. The background covariance $\mathbf{\Sigma}_b$ is assumed diagonal, defined by a spatially varying background variance field. This variance field is estimated by pooling sample variances within oceanic subregions defined by the RECCAP2 region masks. Subregions with insufficient data fall back to a basin-wide or global pooled variance. A Gaussian smoothing filter is then applied over the open ocean to ensure smooth spatial transitions and avoid artificial discontinuities at subregional boundaries (an example of the background variance is shown in the upper panel of Figure 9).

2.4.3 Background and observation error covariances

The background and observation error covariance matrices are constructed following the same isotropic Gaussian-prior formulation used for the dissolved-oxygen product (Ouala et al., 2026), with an e-folding horizontal length scale L_h that we adapt to the observation density of each product. Specifically, and following Ito (2022), we use $L_h = 1000$ km for the yearly product (1965–2022), and we set $L_h = 300$ km for the monthly product, which is consistent with ARGO-based T/S products and reflects the larger number of available observations. The observation error covariance $\mathbf{\Sigma}_o$ is assumed diagonal, with each diagonal element decomposed into three independent variance sources, as detailed below.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



2.4.4 Observation error covariance

The variance at each grid point is the sum of three independent contributions: a *gridding* term (σ_g^2), arising from approximating the within-cell nutrient distribution by the empirical distribution of observed and emulated samples; a *measurement* term (σ_m^2), estimated as 3% of the gridded value following Sharp et al. (2022); and an *emulation* term (σ_e^2), computed analogously but with a slightly larger 4% factor to reflect the additional uncertainty introduced by the emulation step. Assuming independence, $(\mathbf{\Sigma}_o)_{n,n} = \sigma_{g,n}^2 + \sigma_{m,n}^2 + \sigma_{e,n}^2$.

2.4.5 Data aggregation

Observed and emulated data are aggregated over a time window matched to each product's resolution: for the yearly (respectively, monthly) product, each grid cell at year (respectively, month) t includes data from $t \pm 2$ years (respectively, $t \pm 2$ months).

3 Results

In this section, we evaluate the gridded OMNIE-Nutrients products. We first assess the annual-mean climatologies by comparing the spatial distributions and vertical structures of the reconstructed macronutrients with the established WOA23 baseline. We then investigate the seasonal variability of the nutrient fields across both hemispheres, exploring nutrient dynamics at the surface and in the subsurface ocean. Finally, we quantify product uncertainty and compare it with that of a baseline product constructed solely from observations, demonstrating that integrating machine-learning-emulated profiles into the observational network reduces mapping uncertainty.

3.1 Annual mean climatology

We first compare the climatological distribution of each nutrient with the WOA23 reference climatology. Figures 3, 4, and 5 show the annual-mean fields of nitrate, phosphate, and silicate, respectively. For each nutrient, the figures compare the reconstructed product with WOA23 in terms of both the surface distribution and the vertical structure along meridional sections through the Pacific, Indian, and Atlantic Oceans. Overall, our reconstruction shows close large-scale agreement with WOA23. It accurately captures the spatial extent of major nutrient-rich zones and the magnitude of their concentrations, including those associated with equatorial and eastern boundary upwelling systems, the Southern Ocean, and the North Pacific. Furthermore, the meridional cross-sections confirm that the product accurately reproduces the large-scale vertical and meridional distribution of all three nutrients.

3.2 Seasonal variability

To evaluate the climatological seasonal cycle of the reconstructed OMNIE-Nutrients product against the WOA23 baseline, we compare the amplitude of nutrient concentration differences between boreal winter and summer. Specifically, we map

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

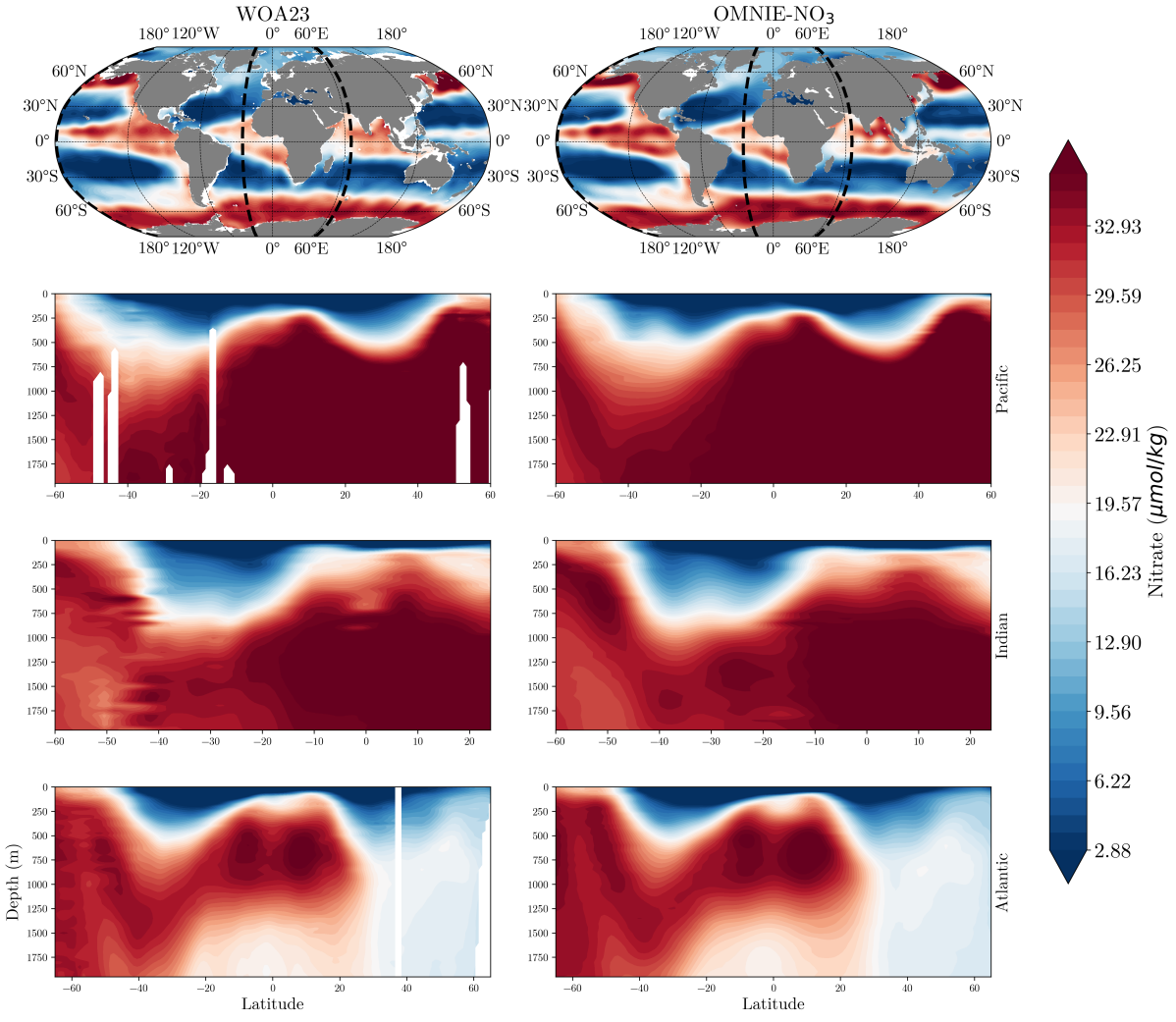


Figure 3. ***Spatial distribution of nitrate concentration in our product compared to WOA23.*** Top row: climatological annual-mean concentration at the ocean surface. Dashed black lines mark the meridional cross sections (surface to 2000 m) shown in the remaining rows for the Pacific ($-179°$E), Indian (65°E), and Atlantic ($-25°$E) basins, WOA23 (left) vs. our product (right).

the absolute seasonal difference between December–February (DJF) and June–August (JJA) climatological concentrations, ($\Delta_{\mathrm{DJF-JJA}}$), at both the surface (5 m) and subsurface (600 m) (Figs. 6 and 7, respectively).

At the surface, the two products show strong agreement in the large-scale pattern of seasonal nutrient differences, characterized by a pronounced hemispheric dipole. Positive values of $\Delta_{\mathrm{DJF-JJA}}$ (DJF $>$ JJA) prevail at mid-to-high latitudes in the Northern Hemisphere, whereas negative values (JJA $>$ DJF) dominate the corresponding latitudes in the Southern Hemisphere. These regions are separated by a broad band of relatively weak seasonal differences across much of the tropics. This canonical distribution reflects the seasonal phasing of thermal stratification and biological demand. During winter, reduced insulation

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

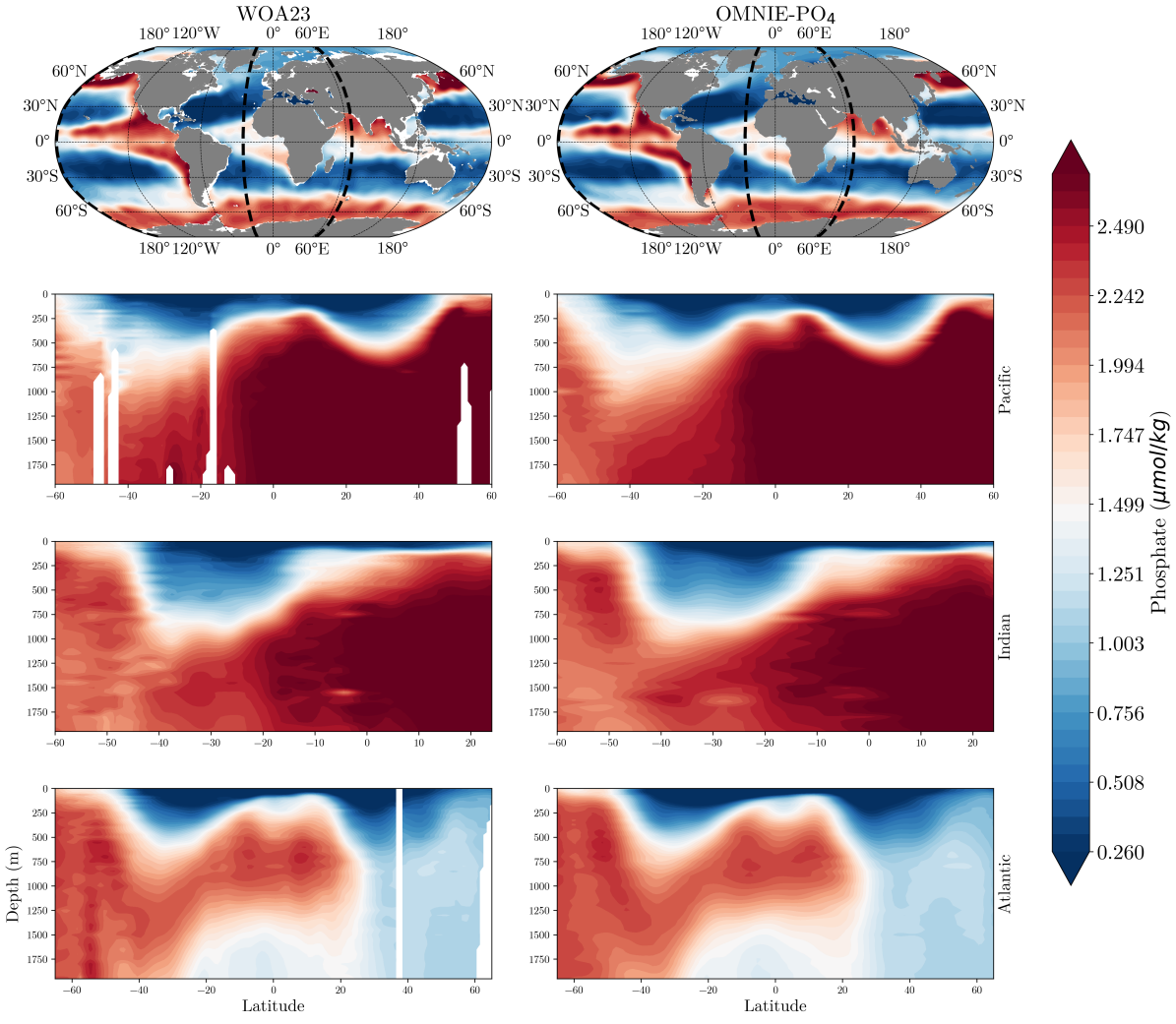


Figure 4. ***Spatial distribution of phosphate concentration in our product compared to WOA23.*** Top row: climatological annual-mean concentration at the ocean surface. Dashed black lines mark the meridional cross sections (surface to 2000 m) shown in the remaining rows for the Pacific (−179°E), Indian (65°E), and Atlantic (−25°E) basins, WOA23 (left) vs. our product (right).

suppresses phytoplankton growth, while deep convective overturning and deepening of the mixed layer replenish surface waters with nutrients from below. Conversely, during the summer months, surface heating enhances stratification and restricts vertical nutrient supply, while increased light availability promotes rapid biological uptake and the depletion of nutrients near the surface (Sarmiento and Gruber, 2006). Seasonal differences are generally weaker at low latitudes because insolation and upper-ocean stratification vary less strongly over the year, although pronounced regional cycles may occur in areas influenced by monsoons, equatorial upwelling, or seasonal shifts in circulation. While both products reproduce these broad spatial patterns, WOA23 generally exhibits slightly stronger seasonal differences. This discrepancy may arise from differences in the

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

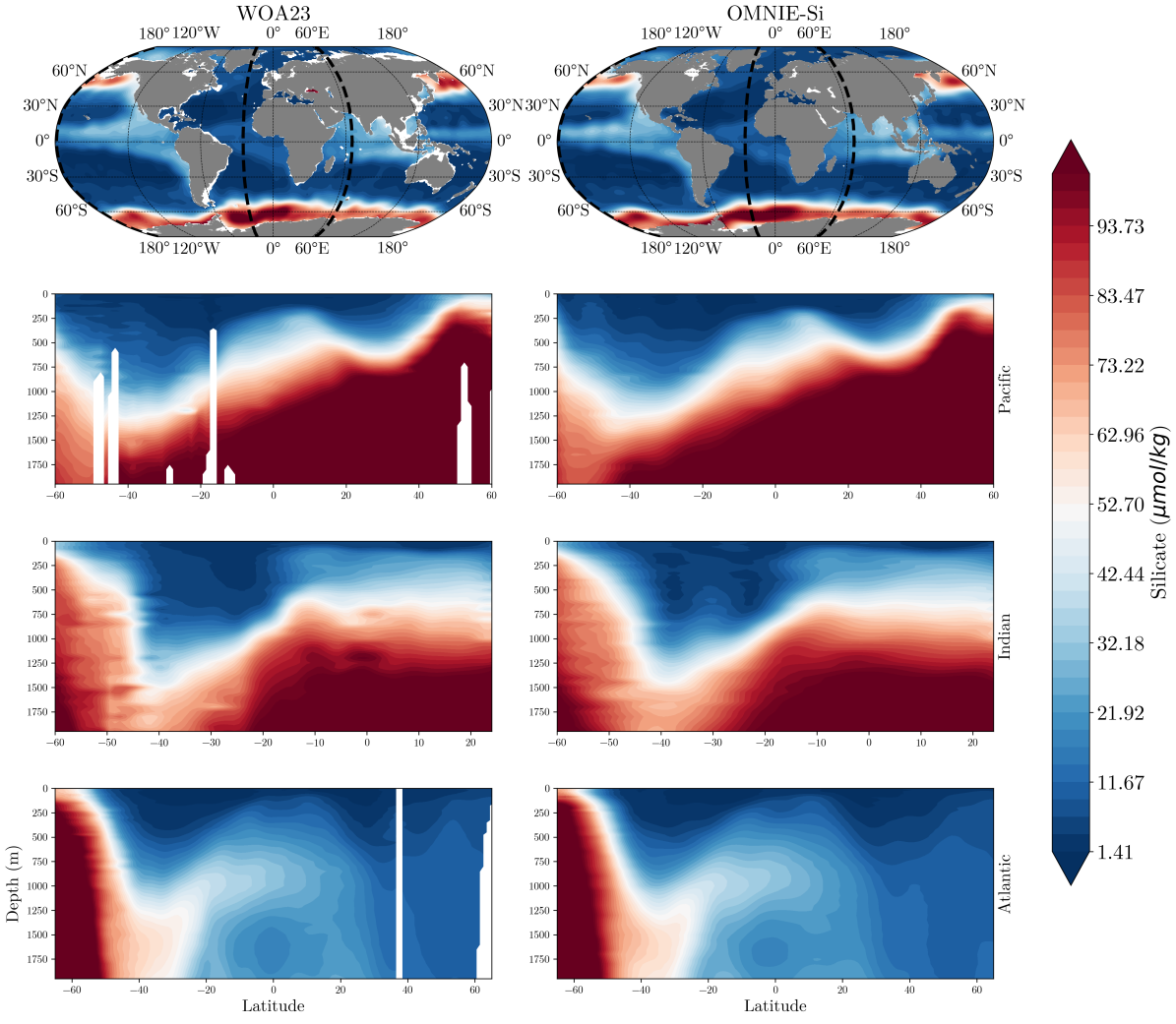


Figure 5. ***Spatial distribution of silicate concentration in our product compared to WOA23.*** Top row: climatological annual-mean concentration at the ocean surface. Dashed black lines mark the meridional cross sections (surface to 2000 m) shown in the remaining rows for the Pacific ($-179°$E), Indian ($65°$E), and Atlantic ($-25°$E) basins, WOA23 (left) vs. our product (right).

temporal aggregation windows. WOA23 bins historical data across a much longer period, whereas the OMNIE product restricts its monthly climatology to the Argo era (2005–2022), thereby capturing a different climatological baseline. Furthermore, the substantial disparities in observational sampling density between the two products likely contribute to this divergence (Fig. C9, Fig. C8, and Fig. C12).

However, pronounced differences between OMNIE-Nutrients and WOA23 emerge at depth (Fig. 7). As expected, the OMNIE product exhibits a strongly attenuated seasonal signal at 600 m, consistent with the limited penetration of surface-driven seasonality below the permanent nutricline. In contrast, WOA23 displays spatially heterogeneous, patchy seasonal anomalies

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



across several regions, in some cases exceeding those found at the surface (Fig. 6). Such intense deep-ocean seasonality is unrealistic and strongly suggests the presence of severe interpolation artifacts (including the aliasing of spatial or interannual variability into the seasonal climatology) rather than genuine seasonal variability at depth.

To investigate these discrepancies further, we examine vertical profiles of seasonal nitrate anomalies from OMNIE-NO_3 and WOA23, spatially averaged across the biogeochemical provinces defined by the RECCAP2 masks. Results for the Atlantic Ocean are presented in Fig. 8, with equivalent analyses for all global biogeochemical provinces provided in Appendix C (Figs. C10 and C11). The first row of Fig. 8 compares the fully interpolated OMNIE and WOA23 products. While both datasets agree reasonably well in the upper ocean of Provinces 1, 2, and 3, they diverge substantially in Provinces 4 and 5. In particular, the WOA23 vertical profiles are highly noisy and in Province 4, the seasonal anomaly between 300 m and 600 m is stronger than at the surface.

To isolate the source of this noise, we recomputed the vertical profiles by subsampling both the OMNIE and WOA23 fields exclusively at grid points containing the actual *in situ* observations used for interpolation (Fig. 8, second row). Under this observational mask, surface agreement in Provinces 1 and 2 remains intact. However, the previously smooth OMNIE profiles in the remaining provinces degrade, exhibiting the same erratic, noisy subsurface behavior seen in WOA23. This confirms that the unrealistic deep-ocean seasonal signals in WOA23 are artifacts of severe undersampling and interpolation noise over sparse data arrays. Furthermore, evaluating seasonal data coverage—defined as the spatial fraction containing ≥ 5 observations in both the summer and winter quarters—in the upper 600 m reveals a stark contrast across all provinces (Fig. C12, Appendix C). The OMNIE product maintains robust coverage, frequently exceeding 80% across most depths and regions, whereas WOA23 coverage generally falls below 5%. This demonstrates that WOA23 is heavily limited by seasonal sampling sparsity, positioning OMNIE as a significantly more reliable dataset for resolving seasonal biogeochemical variability in the upper ocean. By substantially expanding the effective observational network through ML emulation, OMNIE successfully suppresses this artificial noise, yielding an interpolated field that objectively outperforms the baseline at depth.

3.3 Uncertainty estimates

The uncertainty of the reconstructed nutrient products is quantified using the diagonal elements of the analysis covariance matrix, $\mathbf{\Sigma}_a$ (Sect. 2.4). Each diagonal entry represents the analysis variance of the estimated concentration at a specific grid point after data assimilation. This variance reflects both the uncertainty of the background field and the information gain provided by localized observations. Consequently, the analysis variance decreases proportionally with the density and quality of nearby data. Observational quality is encoded within the observation error covariance matrix, $\mathbf{\Sigma}_o$, which assigns lower error bounds (lower uncertainty) to direct *in situ* measurements than to emulated profiles.

By inferring nutrient profiles from T and S measurements, our cascaded framework substantially expands the effective observational network (Fig. 9). This expansion reduces the uncertainty of the reconstructed fields compared to an optimal interpolation relying solely on direct nutrient observations. This reduction is particularly pronounced from the early 2000s onward, driven by the global deployment of the Argo program. To visualize the impact of this augmented network on the interpolation uncertainty, Fig. 9 compares the nitrate standard deviations of three fields. Namely, the background climatology,

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

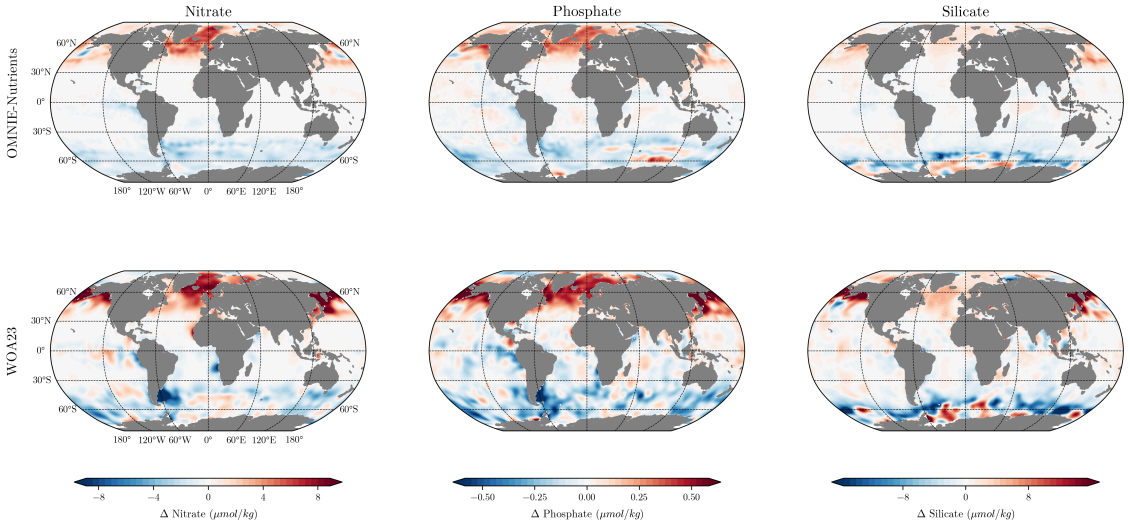


Figure 6. ***Surface seasonal variations of macronutrients (OMNIE vs. WOA23).*** Global maps displaying the absolute seasonal difference ($\Delta_{\mathrm{DJF-JJA}}$) for nitrate, phosphate, and silicate at a depth of 5 m. The top row shows the OMNIE-Nutrients reconstruction while the bottom row presents the WOA23 baseline.

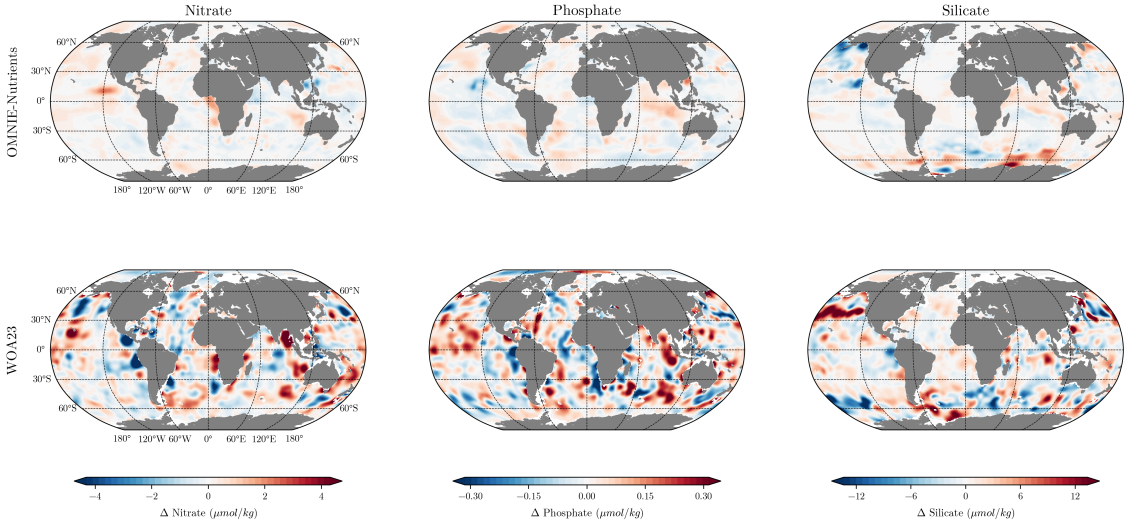


Figure 7. ***Subsurface seasonal variations of macronutrients (OMNIE vs. WOA23).*** Global maps displaying the absolute seasonal difference ($\Delta_{\mathrm{DJF-JJA}}$) for nitrate, phosphate, and silicate at a depth of 600 m. The top row shows the OMNIE-Nutrients reconstruction while the bottom row presents the WOA23 baseline.

a baseline interpolation restricted to *in situ* observations, and our proposed product integrating both *in situ* and emulated

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

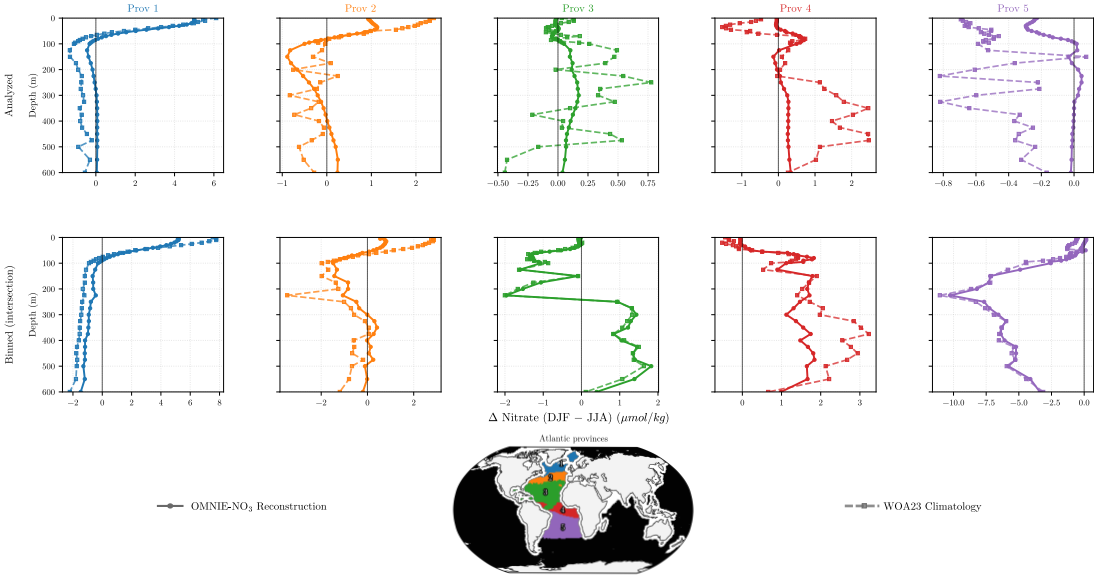


Figure 8. ***Vertical profiles of seasonal nitrate anomalies across Atlantic RECCAP2 provinces.*** Top row: Profiles derived from the fully interpolated OMNIE and WOA23 fields. Bottom row: Profiles derived by masking both products exclusively to grid cells containing actual observational data.

data. Overall, while the observations-only baseline exhibits high uncertainty across data-sparse regimes, the combined product demonstrates a marked, basin-scale reduction in uncertainty due to the significantly enhanced spatial coverage.

4 Conclusion and outlook

We have developed OMNIE-Nutrients, a global macronutrient product that combines sparse *in situ* observations with machine-learning-emulated profiles derived from physical ocean state variables. This framework, previously applied to develop a gridded dissolved oxygen product (Ouala et al., 2026), improves upon standard mapping approaches in several ways. It translates hydrographic measurements into biogeochemical estimates which increases spatial and temporal coverage and mitigates historical sampling biases in undersampled ocean basins. The combined *in situ* and emulated data then undergo optimal interpolation to produce the final gridded fields. We adapt the resolved spatiotemporal scales to observation availability. This yields a yearly product for the historical era (1965–2022) and a monthly product for the Argo era (2005–2022).

Validation against the WOA23 baseline demonstrates that OMNIE-Nutrients reproduces the large-scale spatial distribution and vertical gradients of the major marine macronutrients. Beyond reproducing static climatologies, the increased data density enables the reconstruction of seasonal variability that may be poorly constrained by conventional interpolation of sparse observations. In particular, OMNIE-Nutrients captures coherent upper-ocean seasonal nutrient dynamics while exhibiting the

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

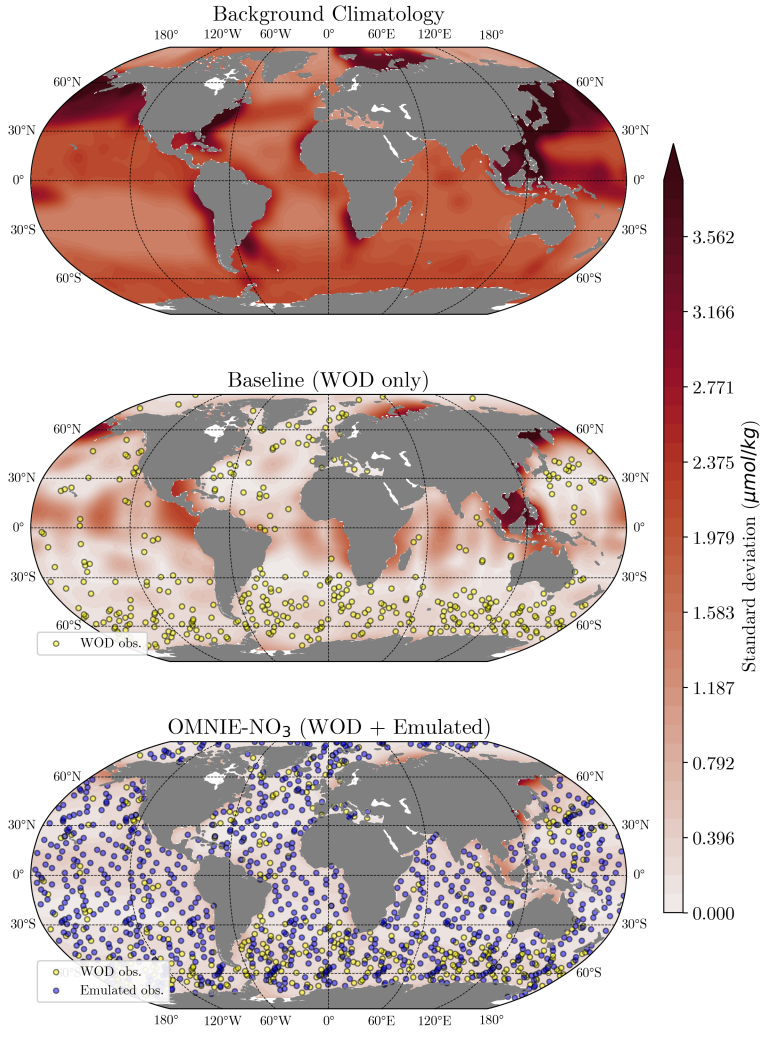


Figure 9. ***Uncertainty analysis of the gridded yearly product in year 2015 at 200m depth.*** We compare the standard deviations of three fields, the background prior (constructed as described in Sect. 2.4), a baseline interpolation using only *in situ* observations, and our proposed interpolated product combining both *in situ* and emulated observations. We also show the corresponding spatial distribution (subsampled for visual clarity) of the direct (green) and emulated (blue) observation networks.

expected attenuation of the seasonal signal with depth, thereby avoiding the spatially patchy and unrealistically large deep-
280 ocean variability apparent in WOA23.

The use of a cascaded neural network, where temperature and salinity inform oxygen, which together constrain nutrient predictions, offers a reliable method for marine biogeochemical state estimation. This approach can be extended to map the broader carbonate system, including total alkalinity, dissolved inorganic carbon, and pH. Generating an observationally grounded suite of these variables will support the evaluation and data-driven tuning of next-generation Earth System Models.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.


While this study centers on validating climatological and seasonal states, these high-resolution fields provide a basis for further climate research. By reducing interpolation uncertainties, the dataset supports efforts to quantify long-term trends in global nutrient inventories and to investigate regional shifts in primary productivity regimes under anthropogenic climate change.

5 Data availability

The gridded products developed in this study are available as NetCDF files (Ouala and Lachkar, 2026b). The corresponding DOI (https://doi.org/10.5281/zenodo.22205370) will be made public upon acceptance of the manuscript.

Author contributions. SO and ZL designed and structured the study. SO developed the model code and generated the gridded product. SO and ZL analyzed the data and prepared the manuscript.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. SO was supported by the Cluster SequoIA Chair HROMS funded by ANR (ANR-23-IACL-0009) and by the ANR Chair AI4Extremes. He also received funding from the Horizon project AI4PEX (grant agreement 101137682) and from the French National program LEFE (Les Enveloppes Fluides et l'Environnement) under project OL4ROMS. ZL was supported by funding provided by Tamkeen through grant CG009 to the Mubadala ACCESS Center and funding support from Mubadala Philanthropies under XR016; their support is greatly appreciated. Data processing and machine learning model training was performed at the High Performance Computing (HPC) cluster of NYUAD, Jubail.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



References

Aouni, A. E., Gaudel, Q., Regnier, C., Van Gennip, S., Drevillon, M., Drillet, Y., and Lellouche, J.-M.: GLONET: Mercator's End-to-End Neural Forecasting System, arXiv preprint arXiv:2412.05454, 2024.

Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land, Earth System Dynamics, 11, 139–159, 2020.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, 2023.

Bittig, H. C., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N. L., Sauzède, R., Körtzinger, A., and Gattuso, J.-P.: An alternative to static climatologies: Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks, Frontiers in Marine Science, 5, 328, 2018.

Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A., and Weber, T.: Multi-faceted particle pumps drive carbon sequestration in the ocean, Nature, 568, 327–335, 2019.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model, Geoscientific Model Development Discussions, pp. 1–21, 2019.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences, 113, 3932–3937, https://doi.org/10.1073/pnas.1517384113, 2016.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., et al.: FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead, arXiv preprint arXiv:2304.02948, 2023.

Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., and Zhu, J.: Improved estimates of ocean heat content from 1960 to 2015, Science Advances, 3, e1601 545, 2017.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., et al.: Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review, IEEE/CAA Journal of Automatica Sinica, 10, 1361–1387, 2023.

Claustre, H., Johnson, K. S., and Takeshita, Y.: Observing the global ocean with biogeochemical-Argo, Annual review of marine science, 12, 23–48, 2020.

Du, C., Zheng, N., Kao, S.-J., Dai, M., Cao, Z., Shi, D., Li, Q., Wang, H., Luo, X., and Li, X.: A historical nutrient dataset (1895–2024) for the North Pacific: reconstructed from machine learning and hydrographic observations, Earth System Science Data, 18, 2951–2969, 2026.

Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., and McGovern, A.: Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook, Artificial Intelligence for the Earth Systems, 1, e210 002, 2022.

Durand, C., Botvynko, D., Frezat, H., Grande, D., Storto, A., Greenberg, D., Fablet, R., Ouala, S., and Le Sommer, J.: Learning-Based Methods and the Future of Numerical Ocean and Sea Ice Modeling, 2026.

Falkowski, P. G., Barber, R. T., and Smetacek, V.: Biogeochemical controls and feedbacks on ocean primary production, science, 281, 200–206, 1998.

Fay, A. and McKinley, G.: Global open-ocean biomes: mean and temporal variability, Earth System Science Data, 6, 273–284, 2014.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Fourrier, M., Coppola, L., Claustre, H., D'Ortenzio, F., Sauzède, R., and Gattuso, J.-P.: A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the Mediterranean Sea: CANYON-MED, Frontiers in Marine Science, 7, 620, 2020.

Garcia, H. E., Bouchard, C., Cross, S. L., Paver, C. R., Reagan, J. R., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Seidov, D., et al.: World Ocean Atlas 2023, volume 4: dissolved inorganic nutrients (phosphate, nitrate, and silicate), 2024.

Ito, T.: Optimal interpolation of global dissolved oxygen: 1965–2015, 2022.

Ito, T., Cervania, A., Cross, K., Ainchwar, S., and Delawalla, S.: Mapping dissolved oxygen concentrations by combining shipboard and Argo observations using machine learning algorithms, Journal of Geophysical Research: Machine Learning and Computation, 1, e2024JH000 272, 2024.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs], http://arxiv.org/abs/1412.6980, arXiv: 1412.6980, 2014.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, Nature, pp. 1–7, 2024.

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J. R., Dunne, J. P., Gehlen, M., Ilyina, T., John, J. G., et al.: Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections, Biogeosciences, 17, 3439–3470, 2020.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, 2023.

Laruelle, G. G., Cai, W.-J., Hu, X., Gruber, N., Mackenzie, F. T., and Regnier, P.: Continental shelves as a variable but increasing global sink for atmospheric carbon dioxide, Nature communications, 9, 454, 2018.

Li, G., Cheng, L., Zhu, J., Trenberth, K. E., Mann, M. E., and Abraham, J. P.: Increasing ocean stratification over the past half-century, Nature Climate Change, 10, 1116–1123, 2020.

Liniger, G., Sharp, J. D., Takeshita, Y., and Johnson, K. S.: Two decades of increase in Southern Ocean net community production revealed by BGC-argo floats, Global Biogeochemical Cycles, 39, e2024GB008 371, 2025.

Lotze, H. K., Tittensor, D. P., Bryndum-Buchholz, A., Eddy, T. D., Cheung, W. W., Galbraith, E. D., Barange, M., Barrier, N., Bianchi, D., Blanchard, J. L., et al.: Global ensemble projections reveal trophic amplification of ocean biomass declines with climate change, Proceedings of the national academy of sciences, 116, 12 907–12 912, 2019.

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E. D., Geider, R. J., Guieu, C., Jaccard, S. L., Jickells, T. D., La Roche, J., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito, M. A., Thingstad, T. F., Tsuda, A., and Ulloa, O.: Processes and patterns of oceanic nutrient limitation, Nature Geoscience, 6, 701–710, https://doi.org/10.1038/ngeo1765, 2013.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: A foundation model for weather and climate, arXiv preprint arXiv:2301.10343, 2023.

Ouala, S. and Lachkar, Z.: Neural-BGC: An observation-driven emulator for hybrid physical-biogeochemical modeling, Geophysical Research Letters, 53, e2026GL123 123, 2026a.

Ouala, S. and Lachkar, Z.: OMNIE-Nutrients: A Global Gridded Ocean Nutrient Dataset (1965–2022) Derived from a Neural Network Emulator and In-Situ Observations, https://doi.org/10.5281/zenodo.22205370, 2026b.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Ouala, S., Fablet, R., Herzet, C., Chapron, B., Pascual, A., Collard, F., and Gaultier, L.: Neural network based kalman filters for the spatio-temporal interpolation of satellite-derived sea surface temperature, Remote Sensing, 10, 1864, 2018.

Ouala, S., Nguyen, D., Drumetz, L., Chapron, B., Pascual, A., Collard, F., Gaultier, L., and Fablet, R.: Learning latent dynamics for partially observed chaotic systems, Chaos: An Interdisciplinary Journal of Nonlinear Science, 30, 103 121, 2020.

Ouala, S., Brunton, S. L., Chapron, B., Pascual, A., Collard, F., Gaultier, L., and Fablet, R.: Bounded nonlinear forecasts of partially observed geophysical systems with physics-constrained deep learning, Physica D: Nonlinear Phenomena, 446, 133 630, 2023a.

Ouala, S., Chapron, B., Collard, F., Gaultier, L., and Fablet, R.: Extending the extended dynamic mode decomposition with latent observables: the latent EDMD framework, Machine Learning: Science and Technology, 4, 025 018, 2023b.

Ouala, S., Tandeo, P., Chapron, B., Collard, F., and Fablet, R.: End-to-End Kalman Filter in a High Dimensional Linear Embedding of the Observations, Stochastic Transport in Upper Ocean Dynamics, p. 211, 2023c.

Ouala, S., Chapron, B., Collard, F., Gaultier, L., and Fablet, R.: Online calibration of deep learning sub-models for hybrid numerical modeling systems, Communications Physics, 7, 402, 2024.

OUALA, S., Hidaoui, O., and Lachkar, Z.: A Novel Global Gridded Ocean Oxygen Product Derived from a Neural Network Emulator and in-situ observations, https://doi.org/10.5281/zenodo.15478088, 2025.

Ouala, S., Hidaoui, O., and Lachkar, Z.: A novel global gridded ocean oxygen product derived from a neural network emulator and in-situ observations, Earth System Science Data, 18, 287–308, 2026.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al.: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, arXiv preprint arXiv:2202.11214, 2022.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the national academy of sciences, 115, 9684–9689, 2018.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, 2019.

Resplandy, L., Hogikyan, A., Müller, J. D., Najjar, R. G., Bange, H. W., Bianchi, D., Weber, T., Cai, W.-J., Doney, S. C., Fennel, K., et al.: A synthesis of global coastal ocean greenhouse gas fluxes, Global Biogeochemical Cycles, 38, e2023GB007 803, 2024.

Roemmich, D., Johnson, G. C., Riser, S., Davis, R., Gilson, J., Owens, W. B., Garzoli, S. L., Schmid, C., and Ignaszewski, M.: The Argo Program: Observing the global ocean with profiling floats, Oceanography, 22, 34–43, 2009.

Sarmiento, J. L. and Gruber, N.: Ocean Biogeochemical Dynamics, Princeton University Press, Princeton, ISBN 9781400849079, https://doi.org/doi:10.1515/9781400849079, 2006.

Sauzède, R., Bittig, H. C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., and Johnson, K. S.: Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: A novel approach based on neural networks, Frontiers in Marine Science, 4, 128, 2017.

Sharp, J. D., Fassbender, A. J., Carter, B. R., Johnson, G. C., Schultz, C., and Dunne, J. P.: GOBAI-O 2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly two decades, Earth System Science Data Discussions, 2022, 1–46, 2022.

Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., Wang, H., Wang, S., Zhu, J., Xu, J., et al.: Xihe: A data-driven model for global ocean eddy-resolving forecasting, arXiv preprint arXiv:2402.02995, 2024.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., et al.: ClimateBench v1. 0: A benchmark for data-driven climate projections, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 954, 2022.

Zhong, W., Ma, X., Wu, Y., Li, C., Shi, T., Gong, W., and Qi, D.: Climatological fields of Southern Ocean interior carbonate system parameters and anthropogenic CO2 reconstructed and integrated from float-and ship-based observations, Earth System Science Data, 17, 7169–7201, 2025.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Appendix A: Parameterization and training of the ML model

To capture the biogeochemical coupling between dissolved oxygen and macronutrients—and to benefit from the error-compensation properties highlighted in Sect. 2.2.1—the neural network $\mathcal{G}_\theta$ is constructed as a cascaded architecture consisting of two distinct sub-networks.

The first sub-network, $\mathcal{G}_{\mathrm{DO}}$, serves as an auxiliary emulator designed to predict dissolved oxygen solely from the physical and spatiotemporal variables. We define this physical input vector as:

$$\mathbf{v}_{\mathrm{phys}}(t,x,y,z) = \Big[T(t,x,y,z),\ S(t,x,y,z),\ z,\ \phi,\ \lambda,\ m\Big]^T. \tag{A1}$$

Consequently, the oxygen prediction is formulated as:

$$\mathrm{DO} = \mathcal{G}_{\mathrm{DO}}(\mathbf{v}_{\mathrm{phys}}). \tag{A2}$$

The second sub-network, $\mathcal{G}_{\mathrm{Nut}}$, predicts the target macronutrient vector $\mathbf{c}$ (comprising $\mathrm{NO_3^-}$, $\mathrm{PO_4^{3-}}$, and $\mathrm{Si(OH)_4}$) using both the physical predictors and the oxygen concentration:

$$\mathbf{c} = \mathcal{G}_{\mathrm{Nut}}(\mathbf{v}_{\mathrm{phys}}, \mathrm{DO}). \tag{A3}$$

Both $\mathcal{G}_{\mathrm{DO}}$ and $\mathcal{G}_{\mathrm{Nut}}$ are implemented as fully connected multilayer perceptrons (MLPs). Each sub-network features an input layer, four hidden layers comprising 256, 128, 128, and 64 neurons respectively, and an output layer dimensioned to its target variable (a single neuron for DO, and three neurons for $\mathbf{c}$). The Rectified Linear Unit (ReLU) activation function is applied across all hidden layers.

The optimization process strictly differentiates between the training and inference phases. During training, $\mathcal{G}_{\mathrm{Nut}}$ is trained using ground-truth *in situ* DO observations as inputs. This stabilizes the learning process and ensures that the nutrient network learns the correct BGC coupling between oxygen and input physical variables. At inference time, when direct DO measurements are largely unavailable, the model operates in its fully cascaded mode: the synthetic DO output from $\mathcal{G}_{\mathrm{DO}}$ is fed directly into $\mathcal{G}_{\mathrm{Nut}}$, making the final nutrient predictions a function of $\mathbf{v}_{\mathrm{phys}}$ alone.

Both sub-networks are optimized using the Mean Squared Error (MSE) loss function via the Adam optimizer (Kingma and Ba, 2014). The network topology and hyperparameters were determined through empirical tuning, incrementally increasing the depth and width of the hidden layers until validation metrics reached a stable plateau.

Appendix B: Data sources

The data used in this study come from two databases. Table B1 lists sources providing collocated temperature, salinity, and dissolved oxygen measurements, used for training, validation, and testing of the neural network emulator; oxygen data from this dataset are also used directly in the optimal interpolation.

Table B2 lists sources providing temperature and salinity only, used to generate emulated oxygen data via the trained ML emulator.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

WOD dataset	Instrument type
OSD	Bottle, low-resolution CTD, low-resolution XCTD, plankton data
CTD	High-resolution CTD and XCTD data
PFL	Profiling float data, mainly from the Argo program
DRB	Drifting buoy data (surface drifters with thermistor chains, ice-tethered profilers)
UOR	Undulating Oceanographic Recorder data (towed CTD)
GLD	Glider data
MRB	Moored buoy data, mainly Equatorial buoy arrays (TAO)

Table B1. ***Data sources used to construct the training, validation, and test datasets of the ML-based emulator.*** Data selection handled using the WODselect web application.

WOD dataset	Instrument type
OSD	Bottle, low-resolution CTD, low-resolution XCTD, plankton data
CTD	High-resolution CTD and XCTD data
PFL	Profiling float data, mainly from the Argo program
DRB	Drifting buoy data (surface drifters with thermistor chains, ice-tethered profilers)
UOR	Undulating Oceanographic Recorder data (towed CTD)
GLD	Glider data
MRB	Moored buoy data, mainly Equatorial buoy arrays (TAO)
SUR	Surface-only data (bucket, thermosalinograph)
APB	Autonomous Pinniped Bathythermograph (TT-D recorders and CTDs on elephant seals)

Table B2. ***Data sources used to construct the temperature and salinity dataset used to generate emulated oxygen data.*** Data selection handled using the WODselect web application.

Appendix C: Additional figures

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

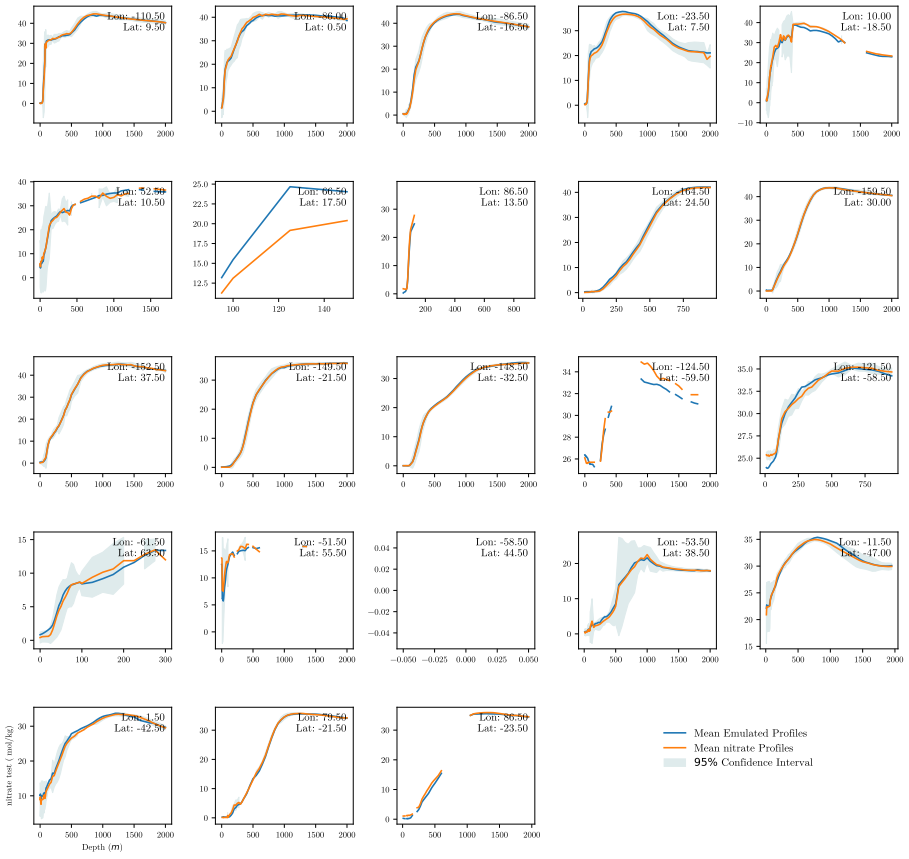


Figure C1. ***Visual comparison of the nitrate averaged vertical profiles and the emulated ones in the*** $1^\circ \times 1^\circ$ ***independent test regions.***
White panels indicate no observations in that particular test region.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

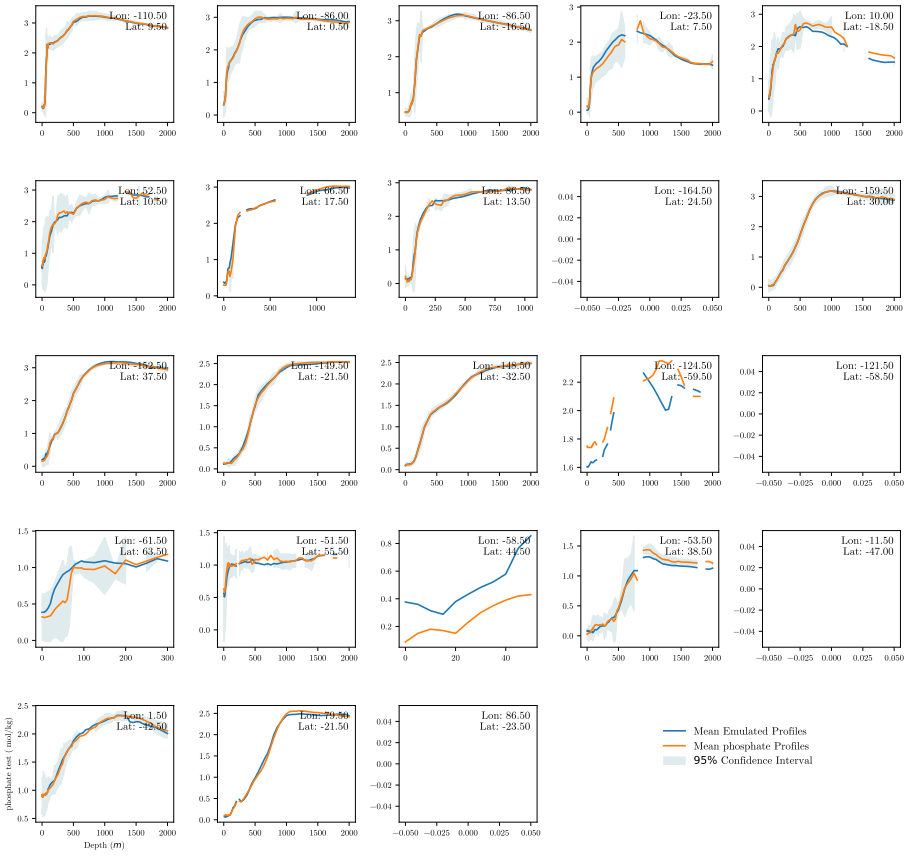


Figure C2. ***Visual comparison of the phosphate averaged vertical profiles and the emulated ones in the*** $1^\circ \times 1^\circ$ ***independent test regions.***
White panels indicate no observations in that particular test region.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

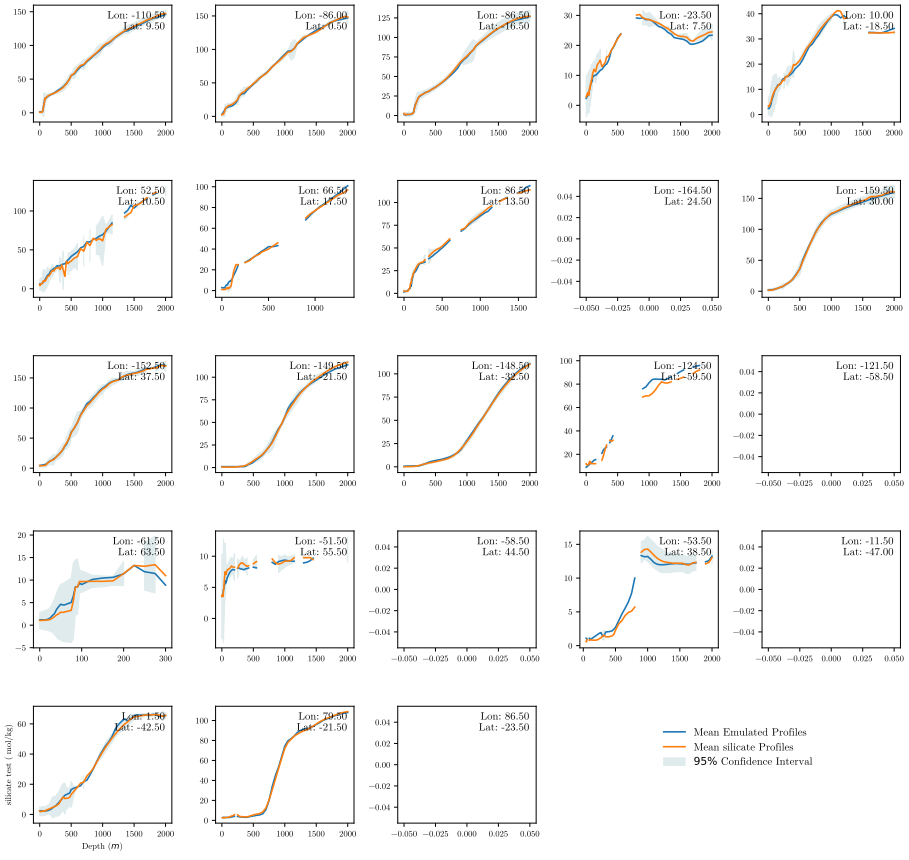
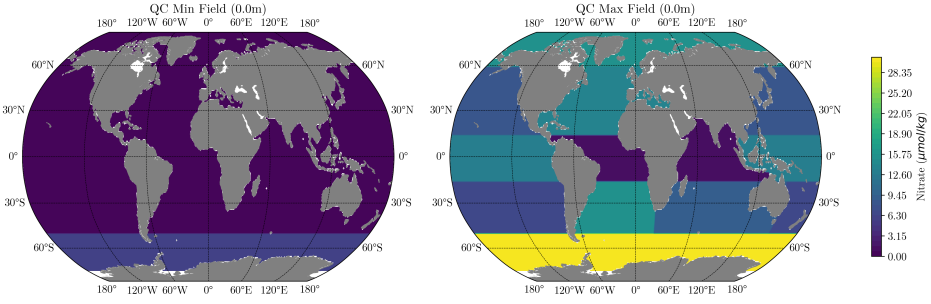


Figure C3. ***Visual comparison of the silicate averaged vertical profiles and the emulated ones in the*** $1^\circ \times 1^\circ$ ***independent test regions.*** White panels indicate no observations in that particular test region.

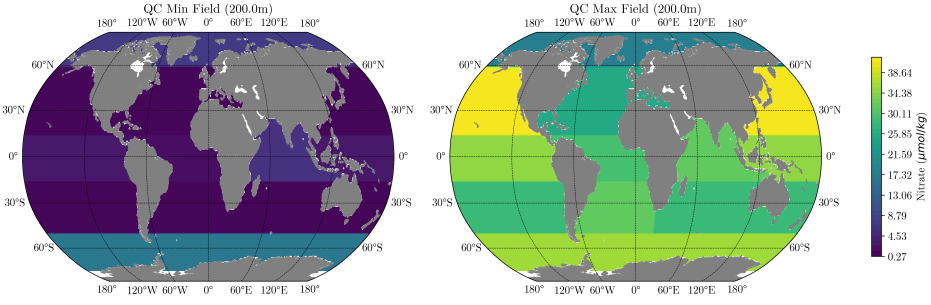
https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



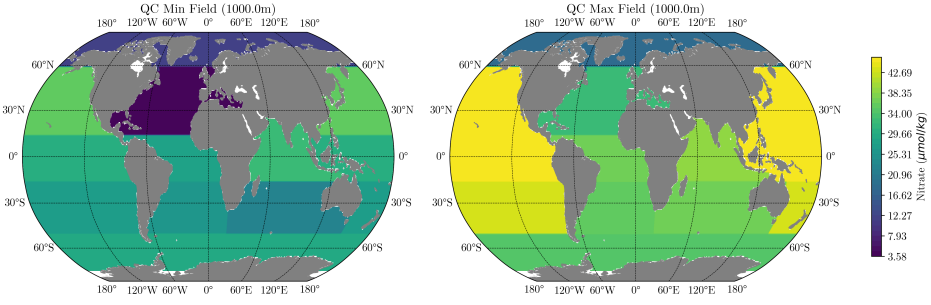
Open Access Earth System Science Data Discussions



(a) 0 m depth



(b) 200 m depth



(c) 1000 m depth

Figure C4. ***Minimum and maximum fields used in the QC of the emulated profiles (Sect. 2.3).*** Nitrate fields at 0 m (top), 200 m (middle), and 1000 m (bottom) depth.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

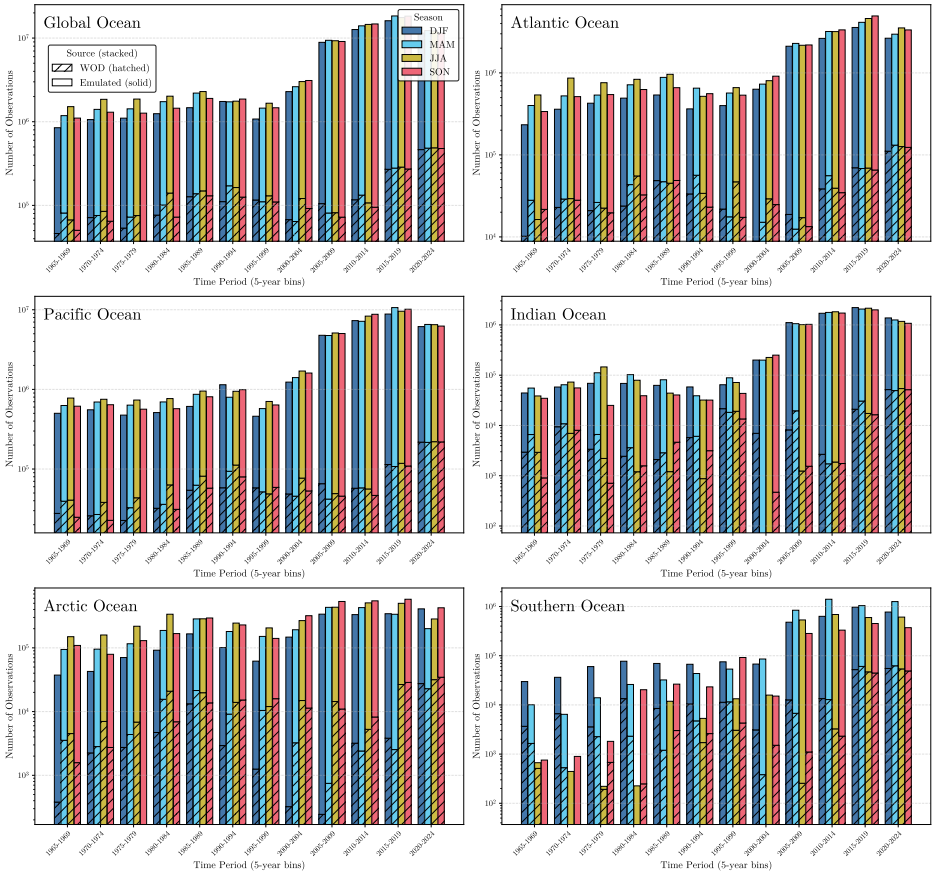


Figure C5. ***Nitrate data sampling by ocean region and by season.*** Each panel corresponds to a different oceanic region. Hatched bars represent real observations while solid bars indicate the contribution of emulated observations. Counts are summed over all depths, months and over all grid cells that fall inside the basin mask.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

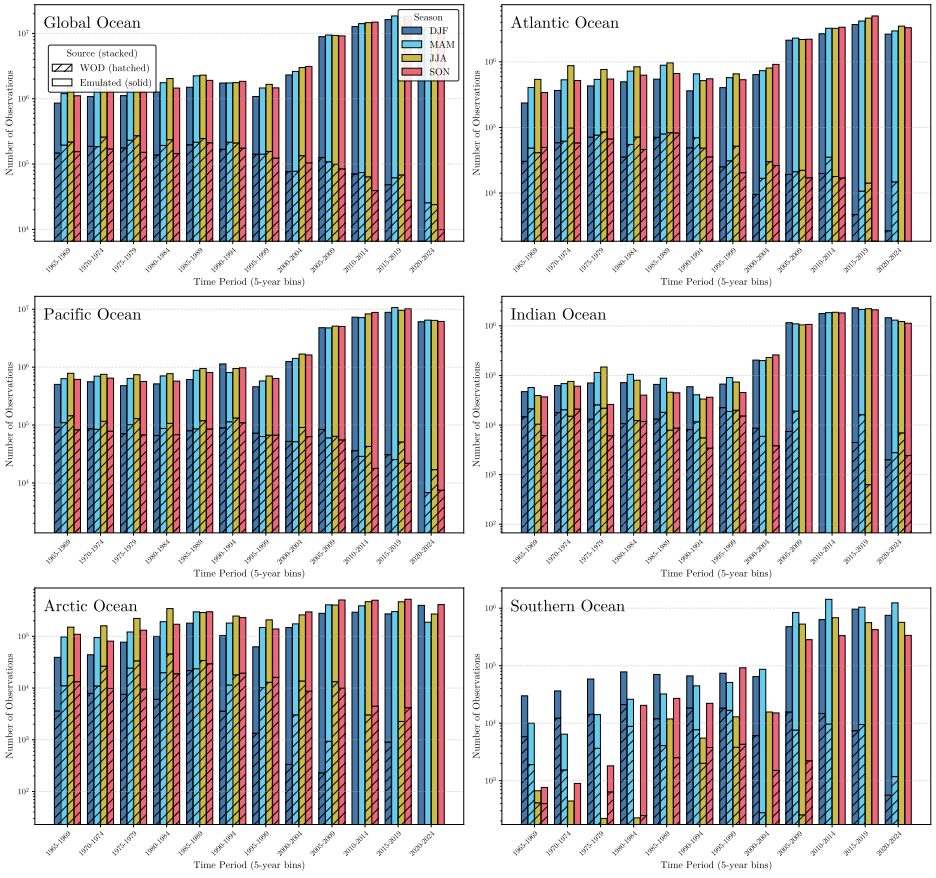


Figure C6. ***Phosphate data sampling by ocean region and by season.*** Each panel corresponds to a different oceanic region. Hatched bars represent real observations while solid bars indicate the contribution of emulated observations. Counts are summed over all depths, months and over all grid cells that fall inside the basin mask.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

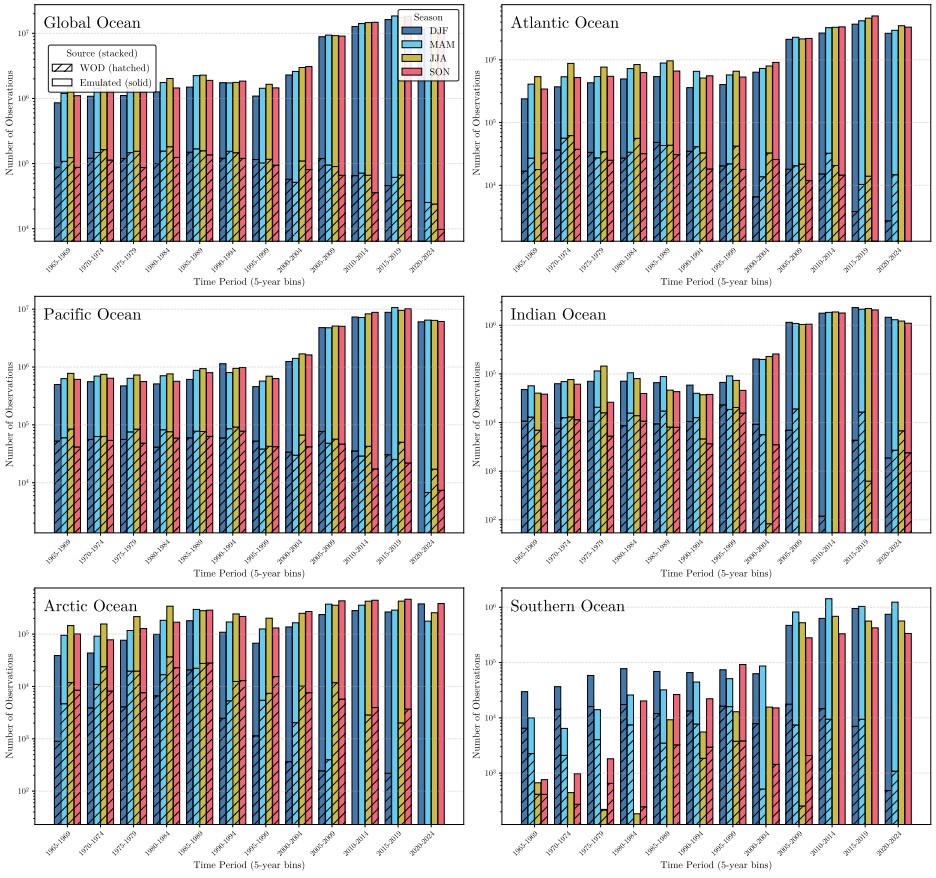


Figure C7. ***Silicate data sampling by ocean region and by season.*** Each panel corresponds to a different oceanic region. Hatched bars represent real observations while solid bars indicate the contribution of emulated observations. Counts are summed over all depths, months and over all grid cells that fall inside the basin mask.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

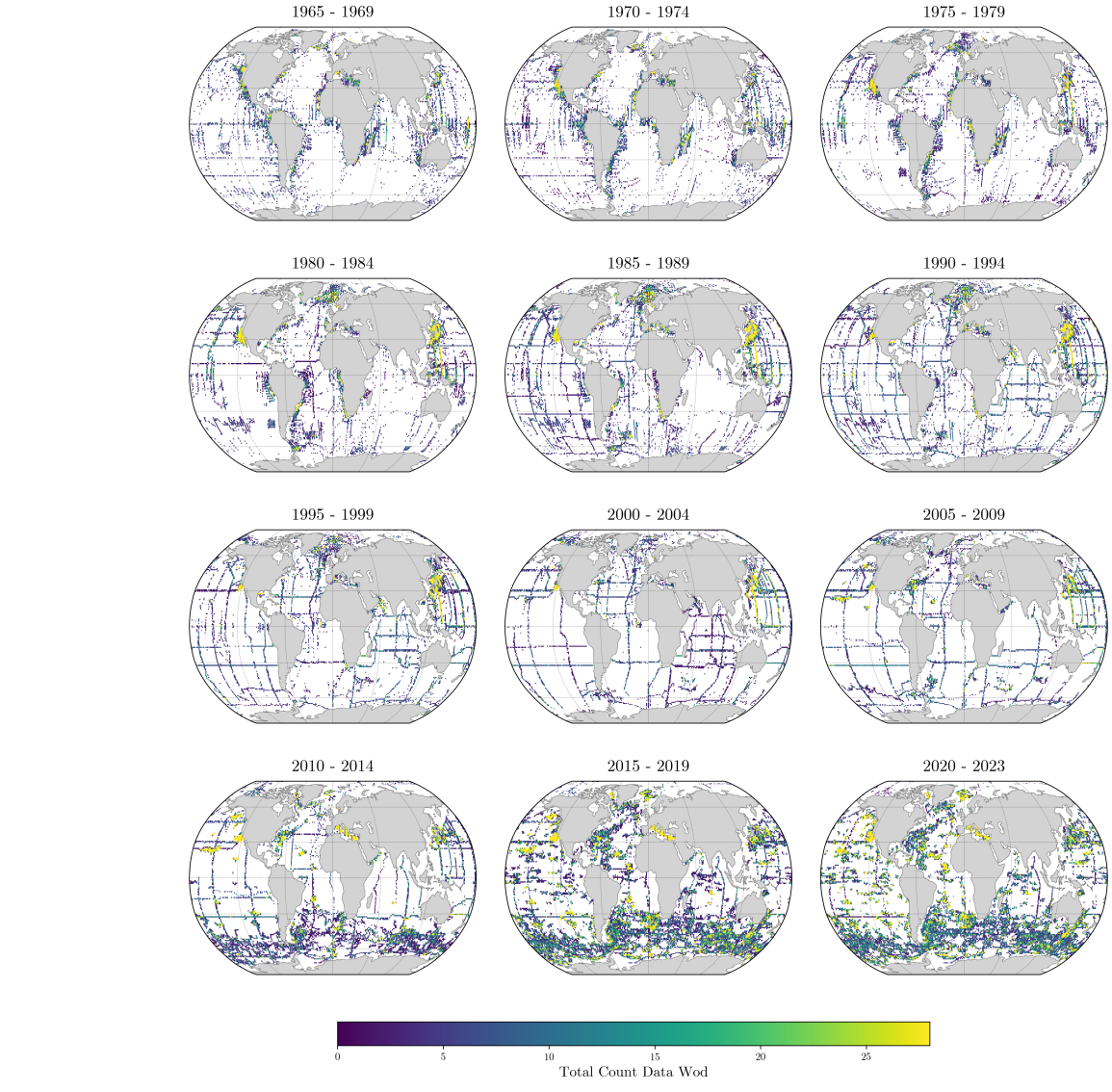


Figure C8. ***Historical nitrate observation density at 200 m depth.*** The maps display the cumulative number of quality-controlled *in situ* measurements recorded per $1^\circ \times 1^\circ$ grid cell, aggregated over consecutive 5-year intervals from 1965 to 2022.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

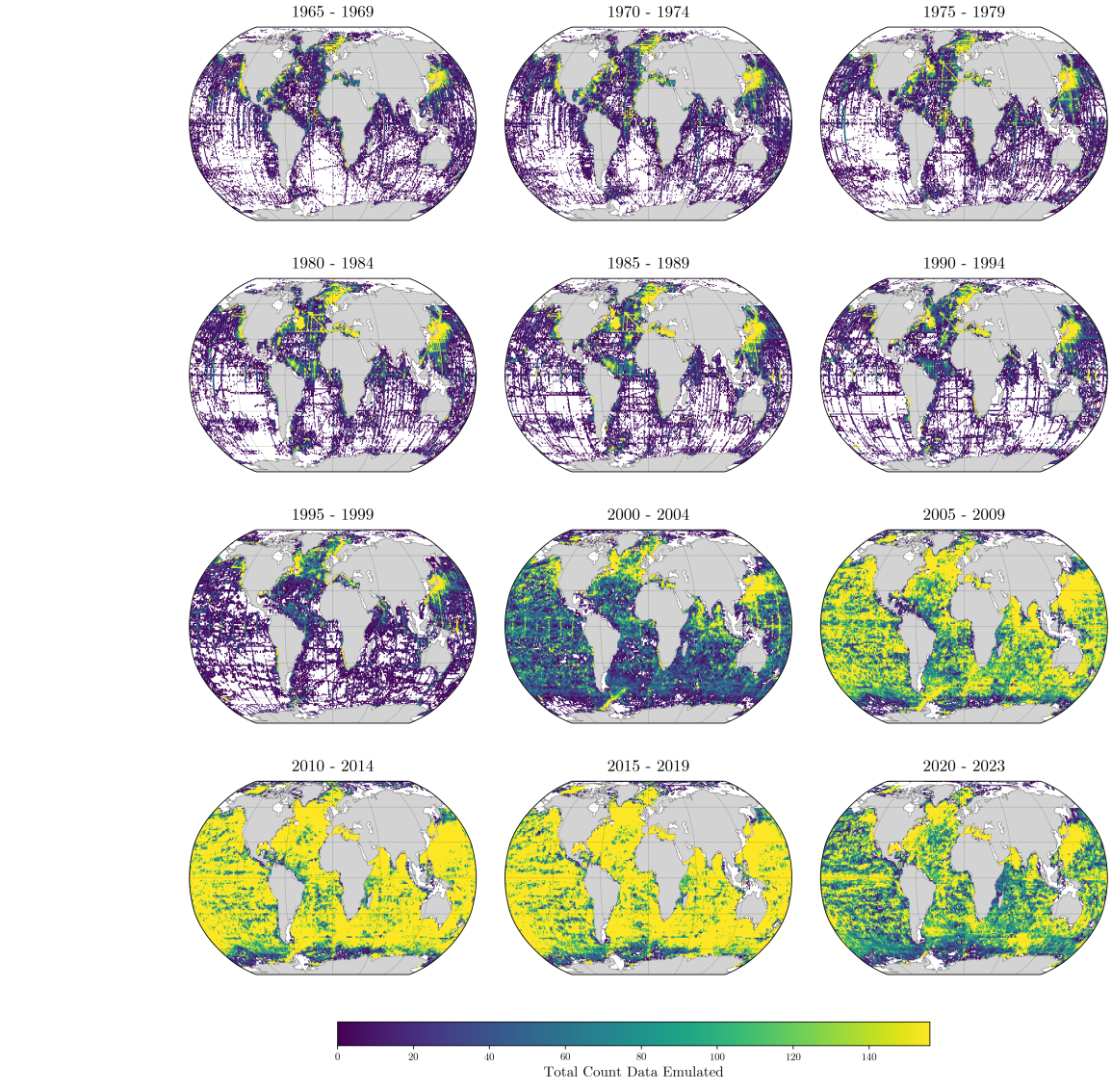


Figure C9. ***Emulated nitrate data density at 200 m depth.*** The maps display the cumulative number of neural-network-emulated profiles per $1^\circ \times 1^\circ$ grid cell, aggregated over consecutive 5-year intervals from 1965 to 2022.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

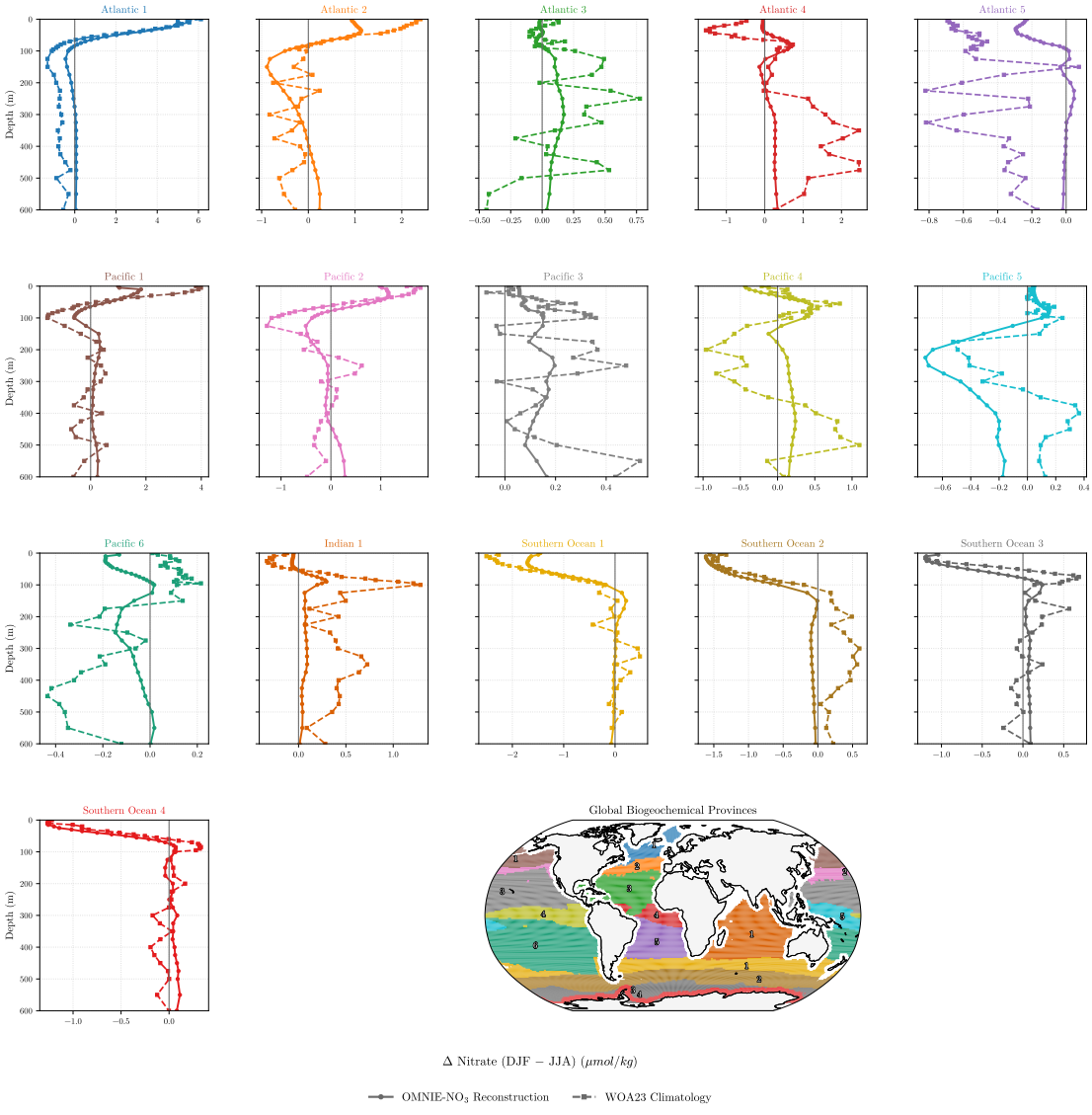


Figure C10. ***Vertical profiles of seasonal nitrate anomalies across major open-ocean biogeochemical provinces.*** Comparison of the absolute seasonal difference ($\Delta_{DJF-JJA}$) between the fully interpolated OMNIE and WOA23 products. Profiles are averaged across individual biogeochemical provinces defined by the RECCAP2 masks (restricted to open-ocean regions containing a minimum of 400 grid cells), extending from the surface down to 600 m depth.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

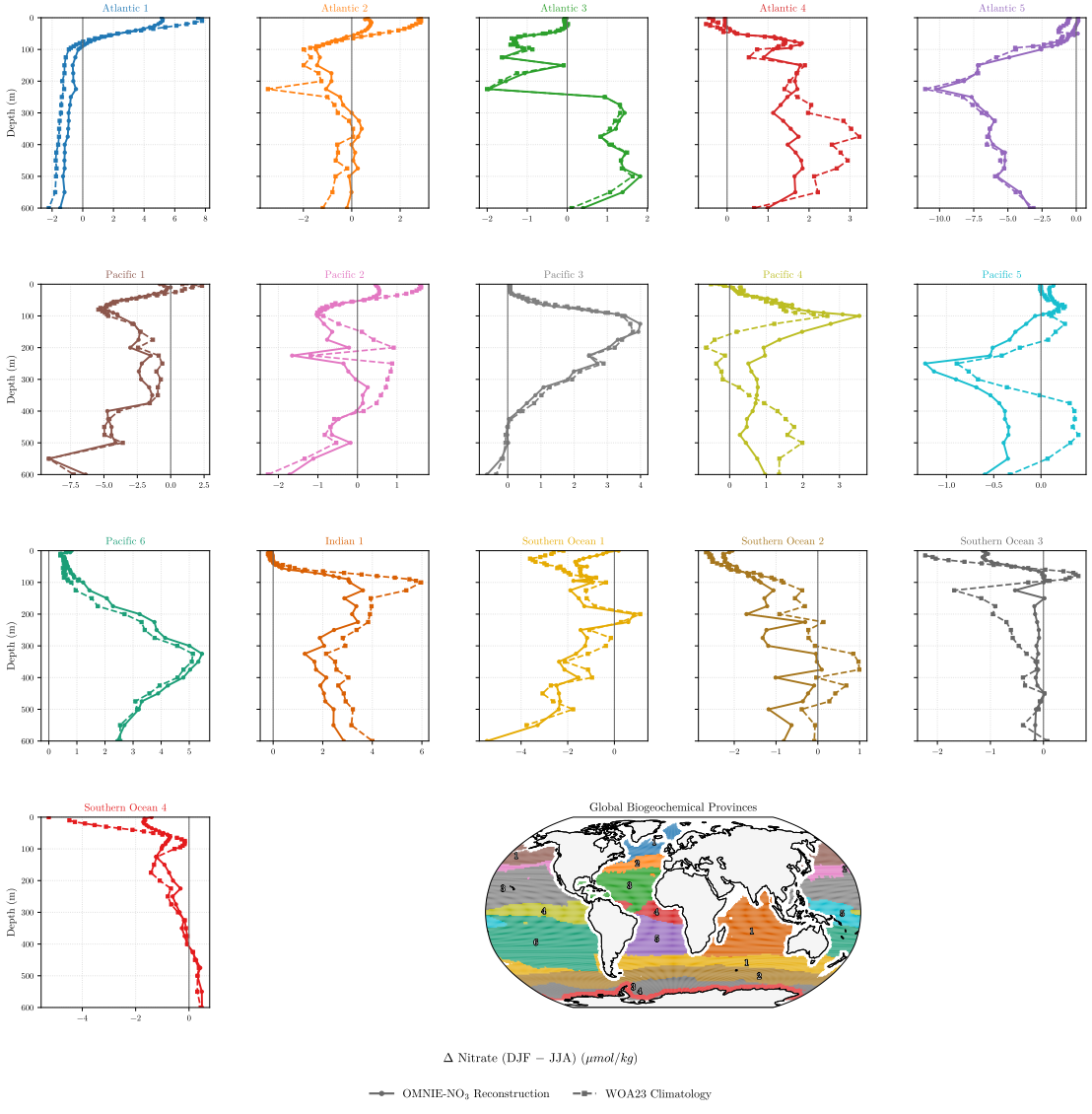


Figure C11. ***Observation-masked vertical profiles of seasonal nitrate anomalies across major open-ocean biogeochemical provinces.*** Comparison of the absolute seasonal difference ($\Delta_{\mathrm{DJF-JJA}}$) between the OMNIE and WOA23 products. To isolate the effects of sampling bias and interpolation noise, both products were strictly subsampled to include only grid cells containing actual *in situ* observational data prior to spatial averaging. Profiles are aggregated across individual biogeochemical provinces defined by the RECCAP2 masks (restricted to open-ocean regions containing a minimum of 400 grid cells), extending from the surface down to 600 m depth.

https://doi.org/10.5194/essd-2026-714
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

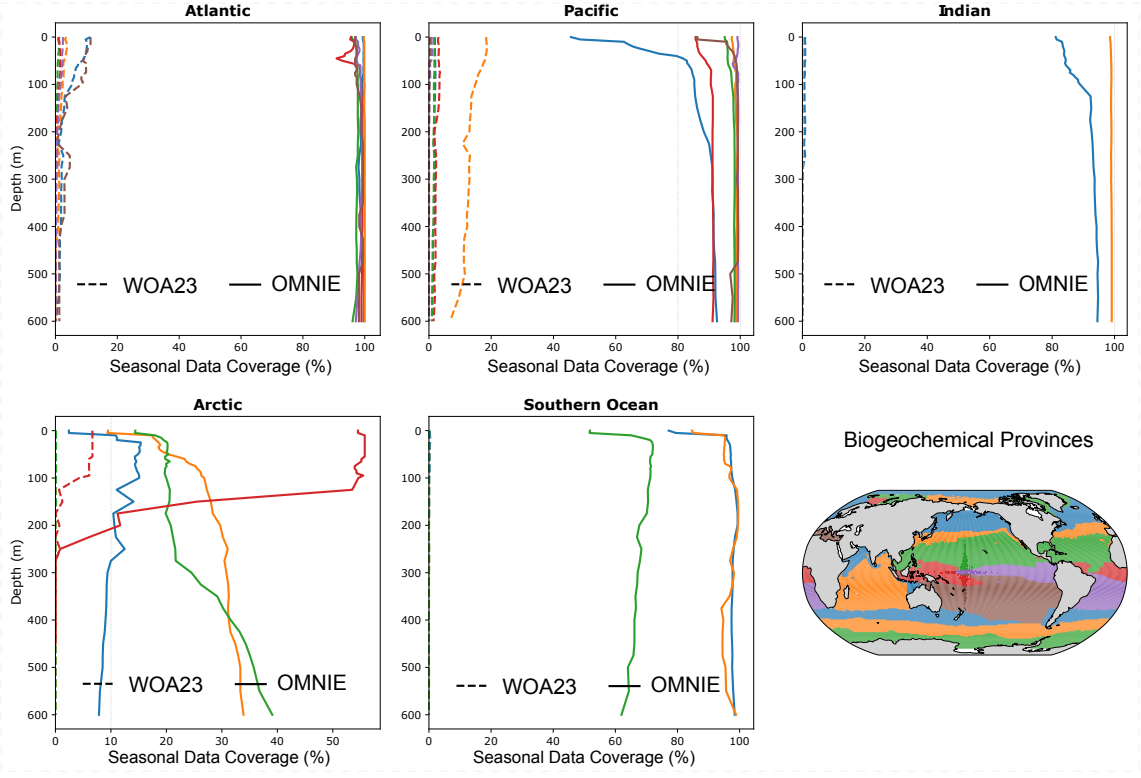


Figure C12. ***Seasonal data coverage.*** Seasonal data coverage in the upper 600 m across the five major ocean basins, aggregated by biogeochemical provinces. Panels (a)–(e) display depth profiles of the seasonal data coverage for the Atlantic, Pacific, Indian, Arctic, and Southern Oceans, respectively. Solid lines represent the observation density of the OMNIE product, while dashed lines correspond to WOA23. Line colors differentiate the specific regional provinces within each basin. Panel (f) provides a global map detailing the spatial boundaries and distribution of the RECCAP2 biogeochemical provinces. A grid cell is defined as having valid seasonal coverage if it contains ≥ 5 observations in both the January–February–March (JFM) and July–August–September (JAS) quarters.