A historical nutrient dataset (1895–2024) for the North Pacific: reconstructed from machine learning and hydrographic observations

Chuanjun Du[1*], Naiwen Zheng[1], Shuh-Ji Kao[1], Minhan Dai[2], Zhimian Cao[2], Dalin Shi[2], Qiancheng Li[1], Hao Wang[1], Xunlan Luo[1], and Xiaolin Li[2*]

[1]School of Marine Sciences, Hainan University, Haikou 570228, China

[2]State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361102, China

Manuscript resubmitted to *Earth System Science Data*

[*]***Corresponding Authors***: Chuanjun Du, cjdu@hainanu.edu.cn; Xiaolin Li, xlli@xmu.edu.cn

Key points:

- Rigorous data quality control procedures were applied to clean nutrient and hydrographic data collected from multiple sources in the North Pacific, following state-of-the-art practices.
- Three machine learning models demonstrated low errors across diverse validation strategies.
- We reconstructed a monumental database of ~473 million nutrient data points across 1.92 million stations (1895–2024), expanding the number of nutrient data points by a factor of 2,127–2,393 compared to original observations.

Abstract

Nutrients play a critical role in oceanic primary productivity and the biological pump. However, compared to hydrographic parameters such as temperature and salinity, nutrient observations are limited due to their labor-intensive and costly measurements. Thus, nutrient observations are several orders of magnitude sparser than hydrographic observations. In this study, we first established a rigorous data quality control procedure to clean the hydrographic and nutrient (including NO_3^-, NO_2^-, DIP, and $Si(OH)_4$) observations collected from World Ocean Database (WOD) and CLIVAR and Carbon Hydrographic Data Office (CCHDO) in the North Pacific. Subsequently, the cleaned and high-quality CCHDO dataset was used to train three machine learning models—Random Forest, Light Gradient Boosting Machine (LightGBM), and Gaussian Process Regression—to establish relationships between nutrient concentrations and key variables, including space coordinates (longitude, latitude, and depth), time variables (year and month), and water mass properties (indexed by potential temperature and salinity). Validation shows that the reconstruction closely matches the observations, with Root Mean Squared Errors (RMSEs) of <1.41, <0.071, <0.089 and <3.07 μmol kg^{-1} for NO_3^-, NO_2^-, DIP, and $Si(OH)_4$, respectively. The validated models were then applied to reconstruct nutrient concentrations from the hydrographic observations in WOD, most of which lacked direct nutrient measurements. This resulted in ~473 million reconstructed nutrient data points across 1.92 million stations for each nutrient, spanning from 1895 to 2024, representing a 2,127 to 2,393–fold increase compared to the original nutrient observations in the North Pacific (197,539 to 222,234). This new dataset will be valuable for studying nutrient transport and budgets, spinning up and validating ocean biogeochemical models, assessing long-term nutrients and their stoichiometric changes driven by anthropogenic forcing and climate change. The dataset generated in this study is openly available on Zenodo at https://zenodo.org/records/17451417.

1 Introduction

Bio-essential elements such as nitrogen, phosphorus, and silicon constitute the fundamental material basis for marine ecosystems. Their concentrations govern primary and new production (e.g., Browning et al., 2023; Lipschultz et al., 2002; Moore et al., 2013) and subsequently regulate oceanic uptake of atmospheric CO_2 (Deutsch and Weber, 2012; Sigman and Hain, 2012). However, traditional nutrient data collection relies heavily on ship-based cruises and subsequent sample analysis, which are labor-intensive, inefficient, and costly (Du et al., 2021). Consequently, compared to the abundant hydrographic data collected from multiple platforms such as Conductivity-Temperature-Depth (CTD) and the Array for Real-time Geostrophic Oceanography (Argo) profilers, etc., nutrient observations are sparse in the ocean. These sparse nutrient observations limit our understanding of both small-scale and long-term nutrient variations and our comprehensive understanding of the mechanisms driving changes in oceanic production and ecosystem dynamics (Bidigare et al., 2009; Yasunaka et al., 2021; Karl et al., 2021).

To address this data sparsity, two main approaches have been commonly employed to augment the spatiotemporal coverage of the observed nutrient data. The first is objective analysis, which interpolates field measurements to generate broader spatial coverage, as implemented in products such as the World Ocean Atlas (WOA) (e.g., Reagan et al., 2023; Lee et al., 2023). The second is data fusion, which establishes statistical relationships between nutrients and environmental predictors such as temperature (e.g., Kamykowski, 1987; Kamykowski et al., 2002; Kamykowski, 2008), density (e.g., Dugdale et al., 1989; Switzer et al., 2003), oxygen, salinity, and chlorophyll *a* (Goes et al., 1999; Palacios et al., 2013; Sarangi et al., 2011). Statistical methods including cubic regression, multiple linear regression (Steinhoff et al., 2010; Arteaga et al., 2015; Madani et al., 2024; Zhong et al., 2024), and generalized additive models (Palacios et al., 2013) are frequently used in these efforts.

Recent studies have demonstrated the potential of machine learning for enhancing the spatial and temporal coverage of nutrient data. For instance, Możejko and Gniot

(2008) used Artificial Neural Networks (ANNs) to model time series of total phosphorous concentrations in the Odra River. Self-organizing maps (SOMs) were used to estimate mixed layer nitrate and sea surface nutrients in the open ocean (Steinhoff et al., 2010; Yasunaka et al., 2014). Liu et al. (2022) applied Support Vector Regression, Random Forest Regression, and ANNs to reconstruct monthly surface nutrient concentrations in the Yellow and Bohai Seas from 2003 to 2019. Their results revealed pronounced seasonal and spatial variability in nutrient levels and underscored the influence of environmental drivers such as sea surface temperature and salinity. Similarly, Sundararaman and Shanmugam (2024) employed Gaussian Process Regression (GPR) models to estimate global ocean surface macronutrient concentrations using satellite-derived data, achieving high accuracy and demonstrating their suitability for large-scale marine ecosystem monitoring. Yang et al. (2024) employed a U-net and Earthformer to reconstruct the three-dimensional nitrate distribution by integrating surface data including wind speed, sea surface temperature, chlorophyll *a*, solar radiation, and precipitation in the Indian Ocean. These advancements highlight the expanding role of machine learning in marine biochemical data fusion and provide novel insights into nutrient dynamics and their ecological impacts.

However, many existing approaches rely solely on mathematical extrapolation or data fusion and often neglect the influence of physical seawater properties, such as water mass characteristics. Using the relationship between nutrient concentration and water masses (indexed by temperature and salinity), Du et al. (2021) successfully predicted the nutrient concentrations in the South China Sea. However, the water masses and their relationship with nutrients can also vary with space and time, which should also be taken into consideration. In addition, most research has predominantly focused on nutrient predictions at surface waters—driven by readily available remote-sensing measurements of sea surface temperature and chlorophyll *a*—while subsurface nutrient distributions remain poorly studied.

The North Pacific Ocean is one of the largest marine biomes in the global ocean (Karl and Church, 2017), spanning a broader longitudinal range than the other oceans in the world and a latitudinal range from tropical to polar regions. It includes a subtropical gyre characterized by extremely low surface nutrient concentrations due to Ekman convergence (e.g., Dave and Lozier, 2010; Browning et al., 2021; Dai et al., 2023), and subpolar gyres in the north with elevated nutrient concentrations driven by Ekman divergence. The atmospheric deposition (e.g., Martino et al., 2014; Qi et al., 2020), N_2-fixation (e.g., Dai et al., 2023), and denitrification (Bonnet et al., 2017) are thought to the main nutrient sources and sinks, which are decoupled in space and time in the North pacific. It has been reported that the North Pacific Subtropical Gyre (NPSG) plays an important role in fixed N inputs in summer, but also contributes disproportionately to losses due to intense water-column denitrification in the eastern Pacific low-oxygen zones (Eugster et al., 2012; Wang et al., 2019).

The North Pacific Ocean is influenced by multiple upwelling and current systems, including the equatorial and California upwelling systems, North Equatorial Current, Kuroshio Current, etc., which further change nutrient levels in these regions. In addition, the North Pacific Ocean exhibits abundant mesoscale eddies (Chelton et al., 2007), which play a critical role in redistributing nutrients and modulating biological activity (e.g., Benitez-Nelson et al., 2007; Ascani et al., 2013; Barone et al., 2022). The interaction of these multi-scale physical processes with biogeochemical processes results in highly dynamic nutrient variability in the upper ocean. Therefore, high-resolution and extensive nutrient datasets are essential to accurately resolve the nutrient dynamics. Although the WOA (Reagan et al., 2023) serves as a primary nutrient database and is widely used for boundary conditions in biogeochemical models, its applicability is constrained by relatively coarse spatial resolution (currently 1°) and climatological smoothing, which limit its ability to represent mesoscale and episodic features or to capture long-term variations.

In the North Pacific, Yasunaka et al. (2014) used the SOMs technique to generate monthly surface nutrient maps by integrating sea surface temperature, salinity, chlorophyll *a*, and mixed layer depth. These maps revealed seasonal and interannual variability in surface nutrient distributions in the northern North Pacific. To investigate long-term changes, Yasunaka et al. (2016) applied Optimal Interpolation to analyze the spatial and temporal evolution of surface nutrient concentrations. Lee et al. (2023) provided spatiotemporally gridded nitrate and phosphate data in northwest Pacific from 1980 to 2019 using the spatiotemporal kriging technique. Wang et al. (2023) used the deep neural network model to estimate nitrate concentrations in the upper northwestern Pacific Ocean using temperature and salinity as the primary input parameters.

In this study, we first collected nutrient data from public databases and applied rigorous quality control procedures. Using machine learning methods, we established relationships between nutrient concentrations and water mass properties, spatial coordinates, and temporal variables. We then evaluated the model performance through a comprehensive error analysis. Finally, the validated models were applied to reconstruct historical nutrient distributions across the North Pacific from 1895 to 2024.

2 Data and Methods

2.1 Observation data

Field observations were originally downloaded from the Climate and Ocean: Variability, Predictability, and Change (CLIVAR) and Carbon Hydrographic Data Office (CCHDO), which distributes vessel-based hydrographic data from programs such as the World Ocean Circulation Experiment (WOCE), Joint Global Ocean Flux Study (JGOFS), GO-SHIP, CLIVAR, and other repeat hydrography efforts (https://cchdo.ucsd.edu/). In total, 631 cruises were collected in the North Pacific, comprising 228,091, 197,617, 225,403, and 212,660 data points for $NO_3^- + NO_2^-$ (NO_x^-), NO_2^-, DIP, and $Si(OH)_4$, respectively (Table 1). The dataset spans from 1973 to 2022 and was downloaded on October 1 2024; any updates made after this date were not included in this study. The data cover a geographic range from 120.08°E to 95.17°W and from 2.05°S to 60.25°N. The study domain was slightly extended into the South Pacific to mitigate potential boundary effects during model development.

Table 1. Information on nutrients and their associated hydrographic data collected from CLIVAR and Carbon Hydrographic Data Office (CCHDO) and the data information after quality control (QC).

	Original data information		Data information after QC	
	Data	Stations	Data	Stations
Temperature	327792	15127	327688	15125
Salinity	328502	15274	328275	15269
NO_x^-	217725	9588	213962	9021
NO_2^-	197617	8233	197539	8228
DIP	225403	9623	222234	9474
$Si(OH)_4$	212660	8220	210447	8121

Hydrographic data for nutrient reconstruction were obtained from the World Ocean Database (WOD; Mishonov et al., 2024), which compiles observations from various platforms, including Autonomous Pinniped Bathythermograph (APB), Conductivity-Temperature-Depth profiler (CTD), Drifting Buoy (DRB), Glider (GLD), Mechanical Bathythermograph (MBT), Moored Buoy (MRB), Ocean Station Data (OSD), Profiling Float (PFL), and Undulating Oceanographic Recorder (UOR). Since nutrient reconstruction models rely on relationships with water masses, only samples containing both temperature and salinity measurements were used; therefore, most APB observations, which record only temperature, were excluded. Among these platforms, CTD, OSD, and PFL provided the majority of usable data. Additionally, several marginal seas—including the South China Sea, the Yellow Sea, the Sea of Japan, and the Sea of Okhotsk—were excluded from this study because they are semi-enclosed and strongly influenced by terrestrial inputs. The spatial domain was consistent with that used for the CCHDO dataset, while the temporal coverage extended from 1875 to 2024. In total, 577,215,683 data points from 2,284,448 stations across 40,113 original cruises were collected (Table 2). In addition, the OSD data before 1970 were extracted for nutrient validation in section 3.1. A total of 102,424, 125,142, 447,335, and 294,734 data points were collected for NO_3^-, NO_2^-, DIP, and $Si(OH)_4$, respectively.

Table 2. Information on hydrographic data collected from World Ocean Database, and the data information after quality control (QC). See main text for acronyms' full name.

Platform	Original data information			Data information after QC		
	Data	Stations	Cruises	Data	Stations	Cruises
APB	692302	46454	189	543714	37209	154
CTD	157914052	315177	8785	135584007	297036	8415
GLD	119302218	288840	384	69834989	285778	380
OSD	8885341	592225	21169	6942902	505780	17671
PFL	284781001	700798	9511	255423345	680531	9099
UOR	3373799	26699	7	3304158	25813	6
MRB	1459032	293734	65	1019565	88487	19
DRB	807938	20521	3	0	0	0
Total	577215683	2284448	40113	472652680	1920634	35744

2.2 Data quality control

Given that the data were collected from multiple platforms using various methods over a long-time span and broad spatial range, quality control (QC) was essential (Du et al., 2021; Wang et al., 2025). Following the QC procedures developed by the World Ocean Database (WOD) (Garcia et al., 2024), we applied comprehensive QC protocols (Fig. 1) to both CCHDO and WOD datasets, including hydrographic and nutrient variables.

Four levels of QC were applied to identify and remove potentially erroneous or low-quality records from the CCHDO and WOD datasets. The first level targeted individual measurements, including several checks. (1) A range check was conducted by defining depth-dependent acceptable value ranges for each parameter; data falling outside these ranges were flagged as invalid. This check was applied to temperature, salinity, NO_x^-, NO_2^-, DIP, and $Si(OH)_4$. Note that the NO_x^- denotes the sum concentration of NO_2^- and

NO_3^-. At stations lacking direct NO_x^- measurements, NO_x^- concentrations were derived by summing discrete NO_2^- and NO_3^- observations. (2) An empirical relationship check was performed to verify consistency among paired variables based on predefined acceptable domains, including temperature–salinity, temperature–NO_x^-, temperature–NO_2^-, temperature–DIP, temperature–$Si(OH)_4$, salinity–NO_x^-, salinity–NO_2^-, salinity–DIP, salinity–$Si(OH)_4$, NO_x^-–DIP, and NO_x^-–$Si(OH)_4$. (3) A six-standard-deviation check was conducted by calculating the mean and standard deviation at each depth level; values falling beyond six standard deviations were flagged as outliers. (4) A gradient check assessed the vertical gradients of each parameter at each depth level across stations; data showing abnormal gradients exceeding five standard deviations from the mean were flagged as questionable. (5) A depth/potential density (σ_θ) inversion check was applied to detect unrealistic reversals in parameters such as temperature and nutrients, which typically exhibit monotonic relationships with depth or σ_θ in stratified waters; measurements violating preset thresholds for depth–temperature, depth–NO_x^-, depth–DIP, depth–$Si(OH)_4$, σ_θ–temperature, σ_θ–NO_x^-, σ_θ–DIP, and σ_θ–$Si(OH)_4$ were flagged. (6) A spike check was implemented to identify abrupt deviations (spikes) between a measurement and its adjacent vertical neighbors; if the difference exceeded a defined threshold, the data point was flagged as suspect. This check was applied to temperature, NO_x^-, DIP, and $Si(OH)_4$. (7) Only measurements with an original quality flag of ‘good’ from CCHDO and WOD were retained, while those marked as questionable or erroneous were flagged as outliers.

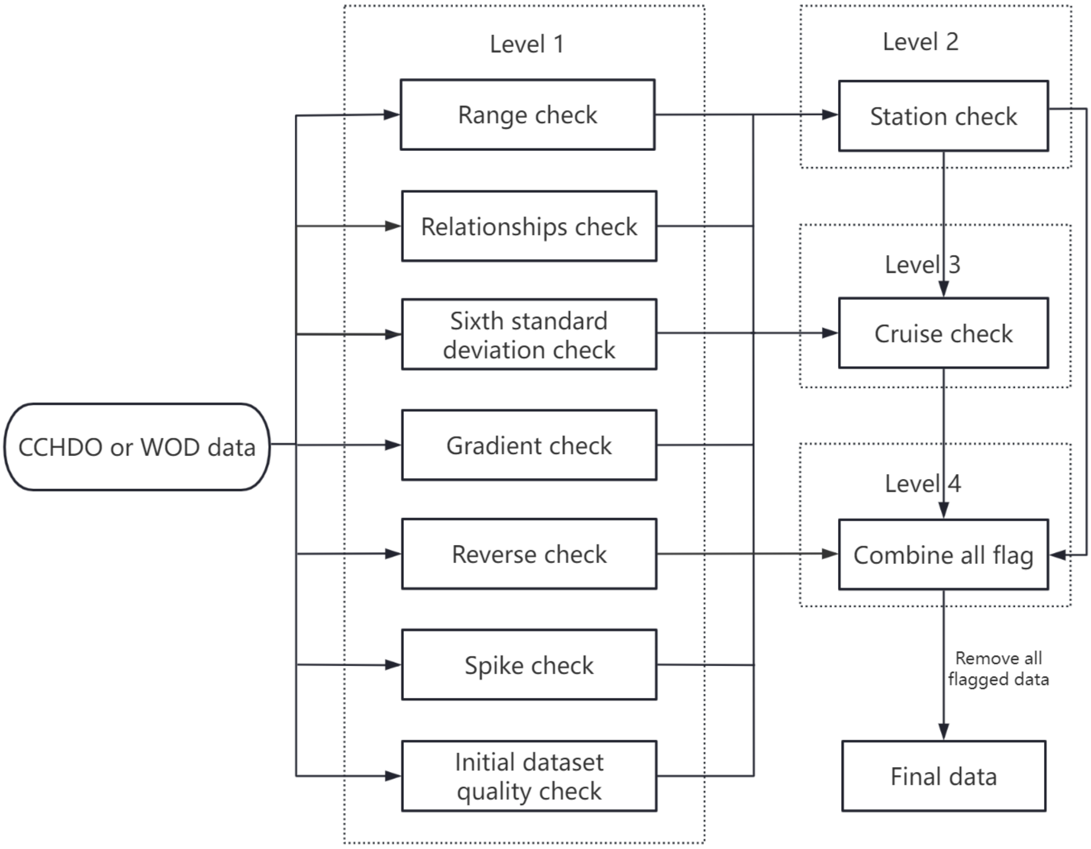

Figure 1. Data quality control procedures for temperature, salinity and nutrients collected from the CLIVAR and Carbon Hydrographic Data Office (CCHDO) and the World Ocean Database (WOD) datasets.

Building on the individual-level QC, we implemented additional QC at the station and cruise levels. At the station level, if a station profile contained more than 20% flagged data points, all data from that station were flagged as questionable. At the cruise level, if over 30% of a cruise's data were flagged, all data from that cruise were flagged. The final step integrated flags from all three levels (individual, station, and cruise), and any data flagged at any level were excluded. This hierarchical QC protocol effectively eliminates low-quality data. Although this approach may discard some high-quality measurements, the large volume of available data necessitates strict QC to ensure reliability.

After quality control, the CCHDO dataset retained 214,943 (9,120), 197,539 (8,228), 222,234 (9,457) and 210,447 (8,123) data points (stations), accounting for 94.2% (95.1%), 100.0% (99.9%), 98.6% (98.5%) and 99.0% (98.8%) of the original data points (stations) for NO_x^-, NO_2^-, DIP, and $Si(OH)_4$, respectively (Table 1). The retained stations cover nearly the entire North Pacific Ocean (Fig. 2a), spanning from 1972 to 2023. Most observations were collected after 1980, with a substantial increase after 1990 (Fig. 2b). Seasonally, the number of stations in June, July, and August was approximately three times greater than that in March and December (Fig. 2c).

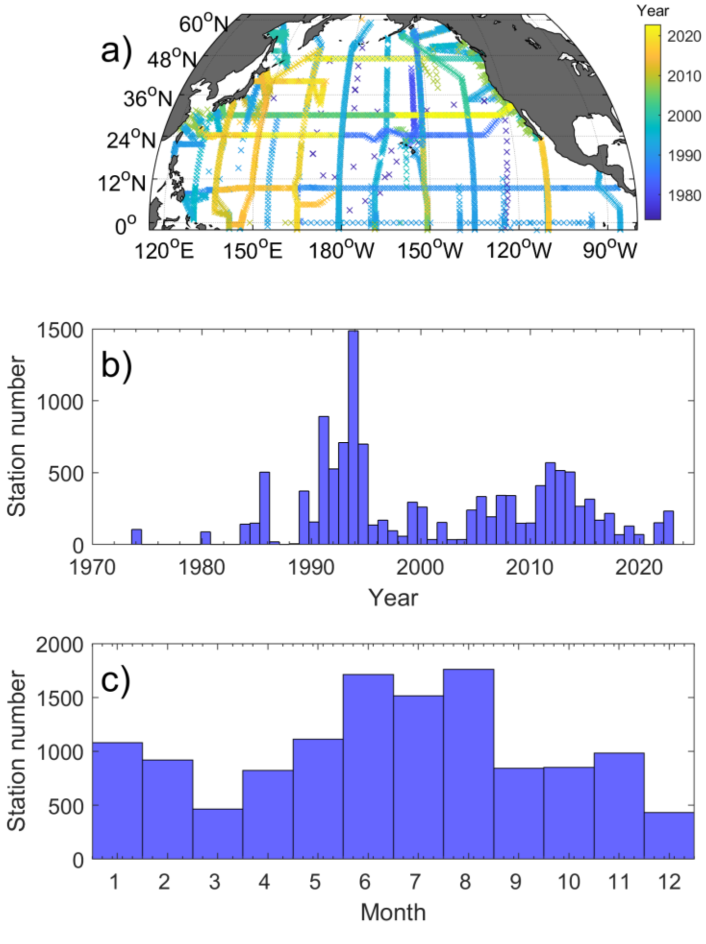

Figure 2. Spatial and temporal distributions of NO_x^- (nitrate plus nitrite) after quality control in the North Pacific. a) Distribution of NO_x^- data locations, with points color-coded by year; b) station counts per year; c) station counts per month.

Following quality control, the final WOD dataset comprised 472,652,680 temperature and salinity data points from 1,920,634 stations across 35,744 cruises, spanning 1895 to 2024. These represent 81.9% of the original observations, 84.1% of the original stations, and 89.1% of the original cruises, respectively (Table 2). Spatially, station counts per 1°×1° grid cell range from 1 to 31,851, with a mean of 249 stations per cell (Fig. 3a). High sampling densities are found off eastern Japan and western North America, resulting from high frequency observations from CTD and OSD platforms, whereas elevated counts in the southwestern North Pacific primarily result from MRB observations. Temporally, fewer than 300 stations per year were collected before 1930. The annual number of stations exceeds 10,000 after 1964 and peaked at approximately 100,000 in 2021 (Fig. 3b). Seasonally, station numbers are highest from May to August (Fig. 3c). Overall, the collected WOD dataset provides 2127–2393 times more observations and 202 times more station records than the CCHDO dataset.

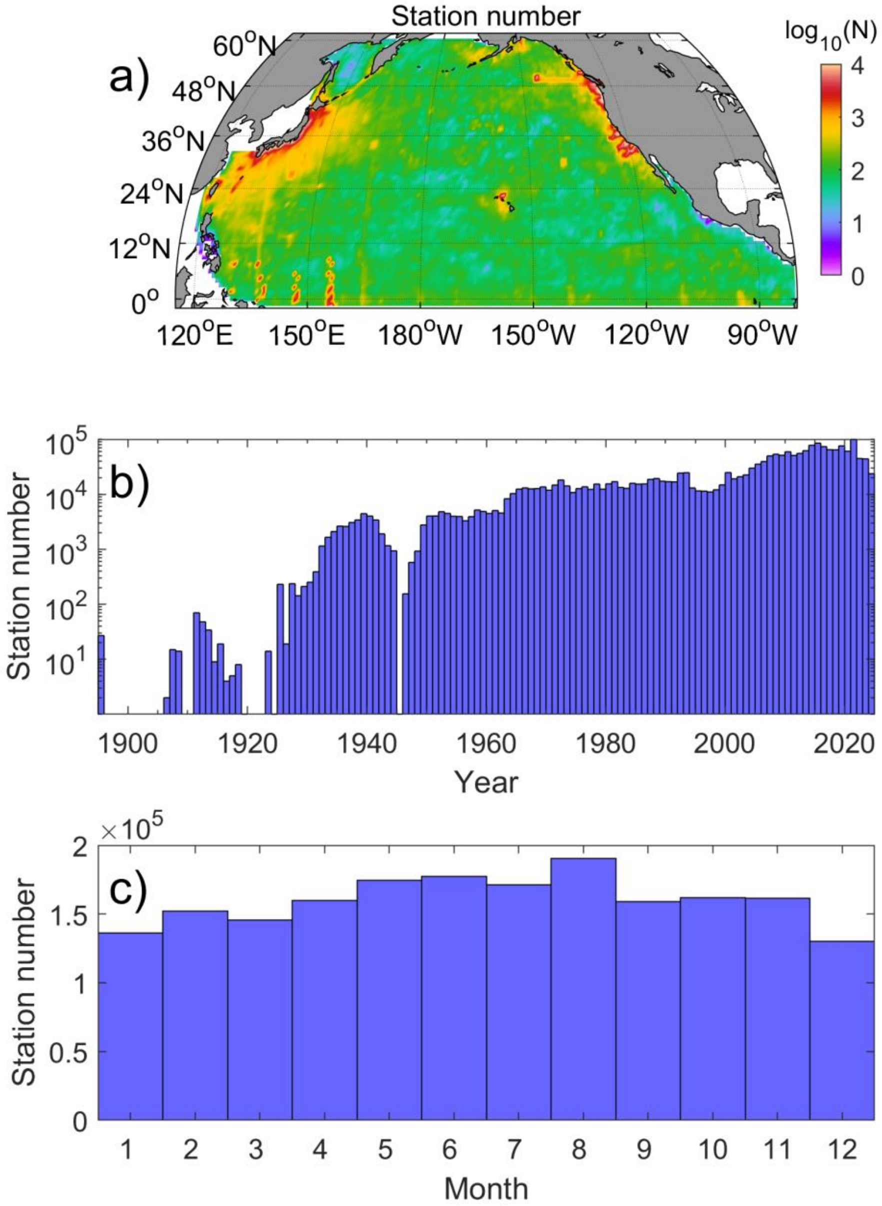

Figure 3. Spatial and temporal distribution of the World Ocean Database (WOD) data after quality control. a) Station counts per 1°×1° grid cell; b) station counts per year; c) station counts per month.

2.3 Machine learning and nutrient reconstruction

After rigorous data quality control, CCHDO data were used to train machine learning models. Three algorithms including Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Gaussian Process Regression (GPR) were applied to establish the relationship between environmental parameters and nutrient concentrations. These methods are widely used in marine science (Hu et al., 2021; Huang et al., 2022; Yu et al., 2022; Chen et al., 2023; Sundararaman and Shanmugam, 2024). The use of diverse models helps decrease algorithm selection bias. RF is an ensemble technique based on bagging, which builds multiple independent decision trees and aggregates their outputs by voting or averaging (Liaw and Wiener, 2002). Its strengths include high predictive accuracy and reduced overfitting owing to the large number of trees. RF has been applied to predict global primary production (Huang et al., 2021), chlorophyll concentrations (Madani et al., 2024), nutrients (Chen et al., 2023; Chen et al., 2024), dissolved iron (Huang et al., 2022), surface ocean $p\mathrm{CO}_2$ (Chen et al., 2019), and N_2 fixation rates (Yu et al., 2024).

LightGBM is an ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT). Compared to standard GBDT, LightGBM employs a leaf-wise tree growth strategy and a histogram-based binning technique to improve predictive accuracy and computational efficiency (Ke et al., 2017). It has been successfully applied to predict water levels (Gan et al., 2021), salinity (Dong et al., 2022; Wang et al., 2022), and chlorophyll *a* concentration (Su et al., 2021). GPR is a non-parametric Bayesian approach that infers relationships by defining a prior distribution over functions via kernel-based covariance matrices, rather than estimating fixed coefficients. This flexibility allows GPR to capture complex, nonlinear input–output relationships and to quantify prediction uncertainty. GPR has been used in oceanography to estimate global dissolved oxygen and nutrient concentrations (Sundararaman and Shanmugam, 2024).

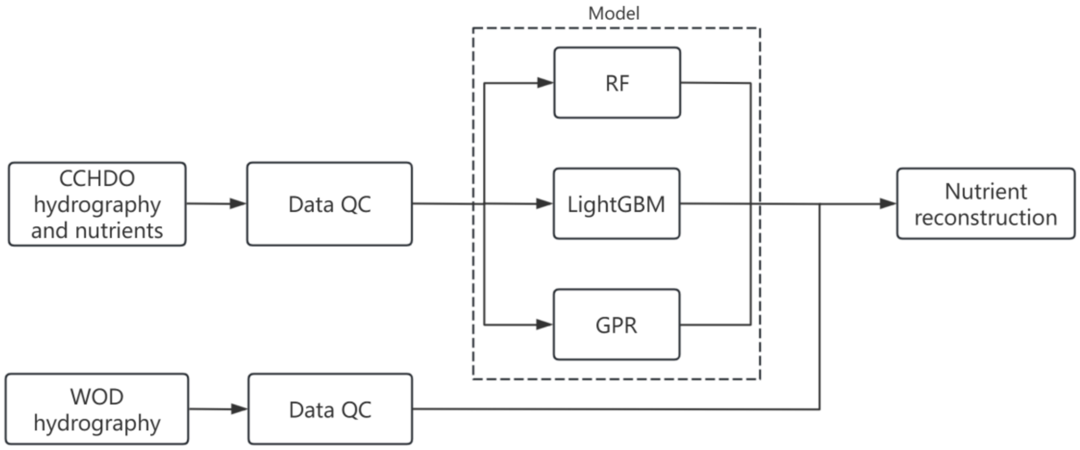

Figure 4. Flowchart of the machine learning framework and its application to WOD hydrographic data for nutrient reconstruction.

In this study, we used spatial coordinates (longitude, latitude, depth), temporal variables (month and year), and water mass properties (represented by potential temperature and salinity) as environmental predictors of nutrient concentrations. The time predictors used month and year with decimals to capture seasonal, interannual, and long-term variability. The North Pacific contains distinct water masses, including North Pacific Subtropical Water, North Pacific Intermediate Water, Antarctic Intermediate Water, Western South Pacific Central Water, North Pacific Deep Water, and Pacific Deep Water, as well as Circumpolar Deep Water (e.g., Talley et al., 2011; Fuhr et al., 2021). These water masses mix to form different water types associated with distinct nutrient concentrations (Fig. 5). Water types have been found to be an important parameter to reconstruct nutrient concentrations in the South China Sea (Du et al., 2021). Thus, potential temperature and salinity serve as proxies for water mass identification.

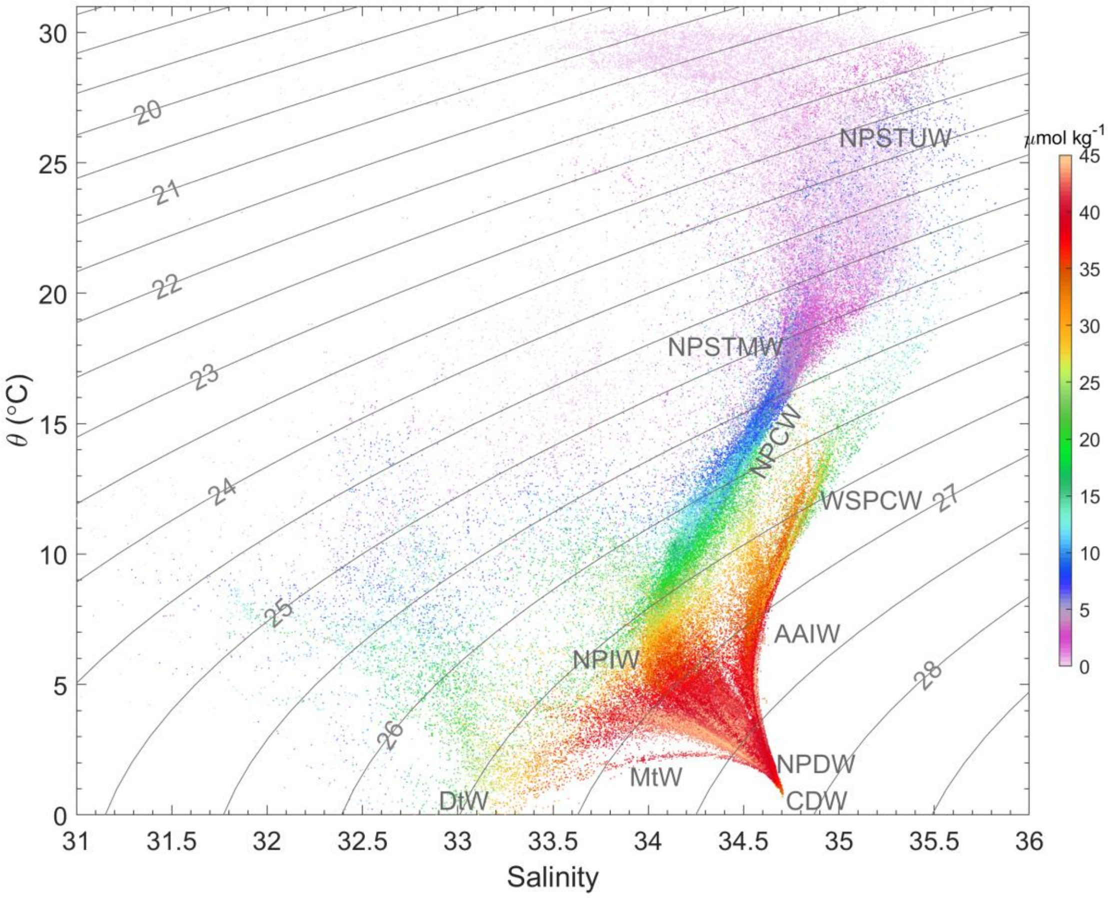

Figure 5. The water masses (indicated by salinity and potential temperature (θ)) and NO_x^- ($NO_3^- + NO_2^-$; color shading) relationships in the North Pacific. The temperature and salinity data were collected from the CCHDO dataset. The gray contour lines and number denote the potential density anomaly. The typical water masses are shown as follows: North Pacific Central Water (NPCW), North Pacific Subtropical Underwater (NPSTUW), North Pacific Subtropical Mode Water (NPSTMW), North Pacific Intermediate Water (NPIW), Dichothermal Water (DtW), Mesothermal Water (MtW), Antarctic Intermediate Water (AAIW), Western South Pacific Central Water (WSPCW), Pacific Deep Water (PDW), and Circumpolar Deep Water (CDW). The water masses and their acronyms are follow the classifications in Talley et al. (2011) and Fuhr et al. (2021).

3 Results

3.1 Error estimation

Leave-one-out cross-validation was primarily used to quantify model reconstruction errors. The CCHDO dataset was divided into training and testing subsets for model development and performance evaluation, respectively. To assess how data partitioning affects error metrics, we implemented four validation methods based on different data-selection strategies (Fig. 6a). The first three methods involved partitioning the CCHDO dataset into training (80%) and testing (20%) subsets. These methods employed three data selection strategies: (1) sample-random, by withholding 20% of individual samples; (2) station-random, by withholding 20% of stations; and (3) cruise-random, by withholding 20% of cruises. Predictions for the held-out subsets, generated using their respective spatial, temporal, and water mass property data, were compared against the actual withheld nutrient measurements to calculate error metrics. These partitioning strategies were designed to evaluate potential errors under the sparse and non-uniform spatiotemporal distribution of observations: Error 1 represented an optimistic estimate (validation data are likely colocated with training data in space and time), Error 3 represented a conservative, generalized scenario (validation data are independent of training data), Error 2 provided an intermediate estimate (validation data may share spatial/temporal context with training data within the same cruise). The choice of error metric (Error 1, 2, or 3) should be guided by the degree of extrapolation in the intended application relative to the training data's spatiotemporal distribution.

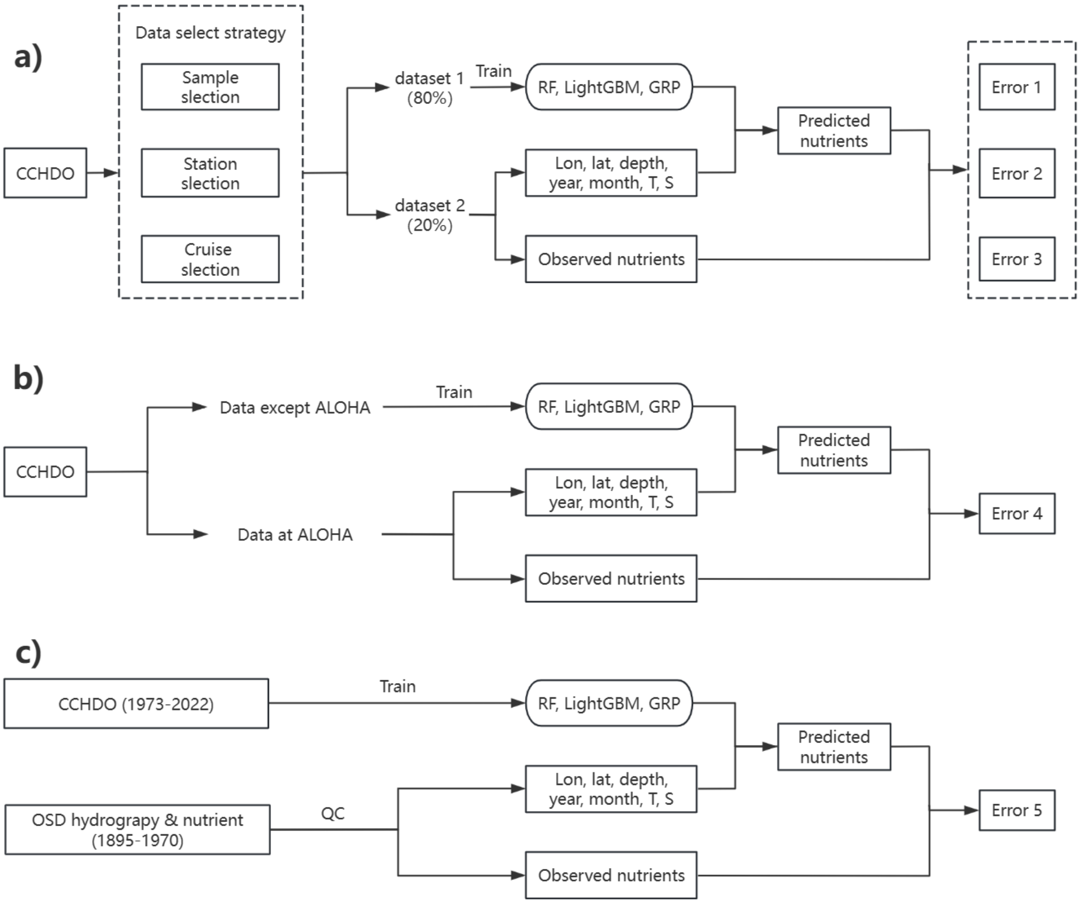

Figure 6. Schematic of the error estimation procedure. a) Error estimation based on three types of data selection strategy; b) assessing temporal error evolution by excluding the data at Station ALOHA; c) examining the models' reconstruction error using the hydrographic and nutrient data before 1970. The T and S denote the potential temperature and salinity, respectively.

The validation results for reconstructed NO$_x^-$ versus observations under the first three data-selection strategies are shown in Fig. 7. RF and GPR exhibited nearly identical performance, with regression slopes of 0.992–0.998, R^2 >0.992, and Root Mean Squared Errors (RMSEs) between 0.734 and 1.313 µmol kg^{-1} (Fig. 7a, c, d, f, g, i). LightGBM showed slightly lower accuracy (slope: 0.991–0.995; R^2: 0.991–0.996; RMSEs: 0.780–1.419 µmol kg^{-1}) (Fig. 7b, e, h). Across different data-selection strategies, sample-random (Error 1) yielded the lowest errors (RMSEs: 0.734–0.983 µmol kg^{-1}) (Fig. 7a–c), station-random (Error 2) was intermediate (RMSEs: 0.908–1.313 µmol kg^{-1}) (Fig. 7d–f), and cruise-random (Error 3) produced the highest errors

(RMSEs: 1.243–1.424 µmol kg⁻¹) (Fig. 7; Table 3). This gradient in error estimates underscores the necessity of employing different data-selection strategies for a comprehensive error assessment. The high slopes and R^2 values (>0.99) achieved across all algorithms and data-selection strategies confirmed the robustness of the nutrient reconstructions.

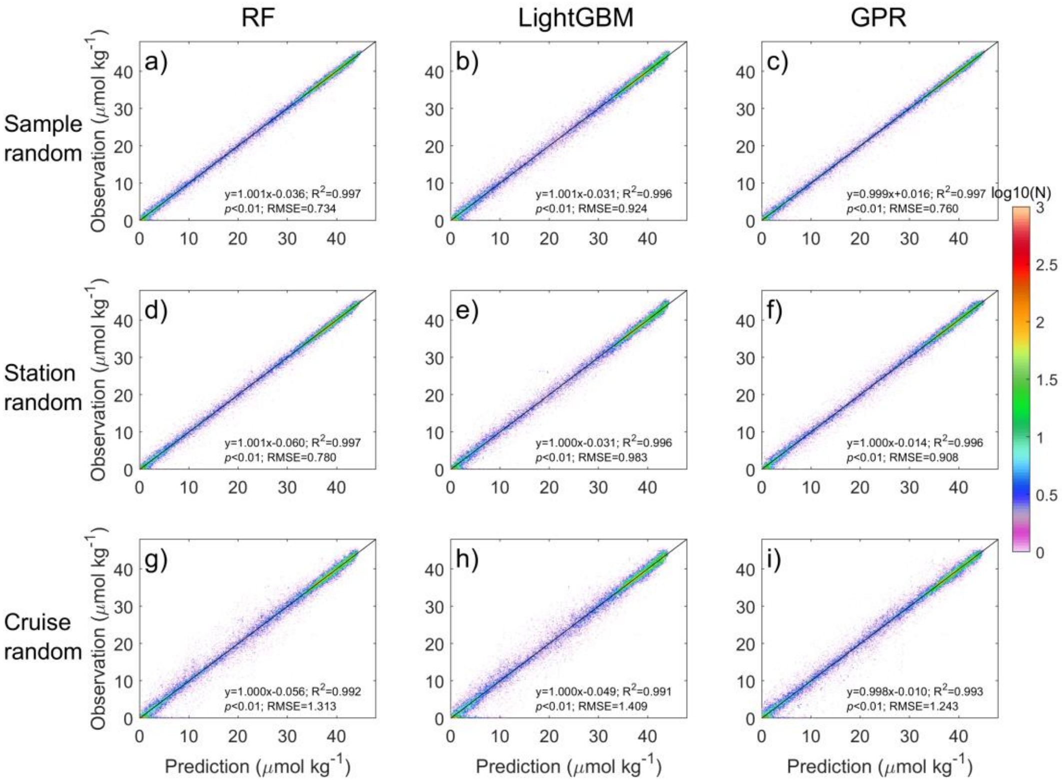

Figure 7. Validating the reconstructed NO_x^- concentrations using leave-one-out cross-validation with different data selection strategies and machine learning methods. Plots shown in row 1 correspond to the sample random strategy (a-c), row 2 correspond to the station random strategy (d-e), and row 3 correspond to the cruise random strategy (g-i). Plots shown in column 1 correspond to the Random Forest (RF; a, d, and g), column 2 correspond to the LightGBM (b, e, and h), and column 3 correspond to the Gaussian Process Regression (GPR; c, f, and i). The black lines and text show the fitted linear regressions, regression equations, coefficient of determination (R^2), p values, and Root Mean Squared Errors (RMSEs). The color represents the data density (N, number of observations). Note that the logarithmic scale of N is applied.

Reconstruction errors for NO_2^-, DIP, and $Si(OH)_4$ are summarized in Figs. S1–S3 and Table 3. Across methods, the RMSEs were below 0.079 µmol kg⁻¹ for NO_2^-, 0.089

μmol kg^{-1} for DIP, and 3.07 μmol kg^{-1} for $Si(OH)_4$. DIP and $Si(OH)_4$ exhibited similar error trends: RMSEs increased from sample-random to station-random to cruise-random selection. In contrast, NO_2^- reconstruction exhibited lower accuracy than NO_x^-, DIP, and $Si(OH)_4$, with regression slopes of 0.48–0.68 and R^2 values of 0.32–0.72. RF and LightGBM outperform GPR for NO_2^-. The poorer NO_2^- performance likely reflects its generally low concentrations (mostly <0.5 μmol kg^{-1}) and high biological variability. Thus, we highlight NO_2^- as a high-uncertainty reconstruction.

Table 3 The Root Mean Squared Errors of nutrient reconstruction from different error evaluation strategies (unit: μmol kg^{-1}).

Data selection strategy	NO_x^-			NO_2^-			DIP			$Si(OH)_4$		
	RF	Light GBM	GPR	RF	Light GBM	GPR	RF	Light GBM	GPR	RF	Light GBM	GPR
Sample random	0.724	0.924	0.760	0.049	0.054	0.079	0.056	0.070	0.055	1.90	2.30	1.53
Station random	0.780	0.983	0.908	0.065	0.068	0.072	0.058	0.071	0.065	2.07	2.45	2.20
Cruise random	1.313	1.409	1.243	0.054	0.057	0.071	0.080	0.089	0.084	2.79	3.07	2.94
ALOHA validation	0.701	0.842	0.674	—	—	—	0.066	0.079	0.064	2.13	2.48	2.32

Understanding the spatiotemporal structure of reconstruction errors is also important for assessing the models' reconstruction applicability. As shown in Figs. S4-S7, the reconstruction errors of NO_3^-, DIP, and $Si(OH)_4$ are generally small in the surface layer, increase with depth to maxima at the nutricline, and then decrease to low values in deep layers. However, the random errors associated with individual cruise observations for $Si(OH)_4$ display no evident vertical pattern. Horizontally, we paid particular attention to surface waters due to their greatest concentration gradients. The horizontal distribution shows that the errors are small in the western NPSG (a nutrient-depleted region) but are large in the subarctic gyre and close to the equatorial regions (nutrient-replete regions; Figs. S8-S11). Here, we particularly examined the nutrient reconstruction errors in the oligotrophic NPSG. The oligotrophic regimes are defined as regions where NO_3^-, NO_2^-, DIP, and $Si(OH)_4$ concentrations are <0.2, <0.2, <0.2, and <5.0 μmol kg^{-1}, respectively. As shown in Table 4, the reconstruction errors in these regimes are <0.574, <0.056, <0.084, and <1.88 μmol kg^{-1} for NO_3^-, NO_2^-, DIP, and

$Si(OH)_4$, respectively, which are evidently lower than the overall RMSEs for the entire North Pacific (Table 3). Among these models, the RF generally performs the best compared to the others. This confirms that absolute errors decrease in oligotrophic regimes. Since the number of summer observations is up to three times greater than that in winter and spring, we further examined the seasonal variation of errors. Overall, no evident seasonal variations are displayed. Only in the case of random cruise selection was the NO_3^- error shown to be greater in spring (March to May) than in other seasons (Fig. S12). For other cases and nutrients, seasonal variation in error was not evident. On a decadal timescale, the reconstruction errors display a slight decreasing trend, particularly for DIP, from 1973 to 2020 (Fig. S13), implying that the errors might be smaller in recent decades than in previous ones.

Table 4 The Root Mean Squared Errors of nutrient reconstruction from different error evaluation strategies in surface oligotrophic regimes (unit: μmol kg^{-1}).

Data selection strategy	NO_x^-			NO_2^-			DIP			$Si(OH)_4$		
	RF	Light GBM	GPR	RF	Light GBM	GPR	RF	Light GBM	GPR	RF	Light GBM	GPR
Sample random	0.290	0.567	0.444	0.018	0.035	0.048	0.028	0.042	0.039	1.19	0.90	1.30
Station random	0.303	0.457	0.474	0.030	0.030	0.043	0.036	0.045	0.043	1.24	1.51	1.51
Cruise random	0.378	0.457	0.574	0.030	0.029	0.056	0.075	0.077	0.084	1.85	1.88	1.75

A fourth validation step assessed the model's temporal performance at Station ALOHA (Error 4; Fig. 6b). To test this, we withheld all observations from ALOHA (which, since 1988, represent 8.52%, 8.45%, and 8.11% of the total $Si(OH)_4$, NO_x^-, and DIP records, respectively) from model training. We then reconstructed nutrient concentrations using space, time, and water-type predictors at Station ALOHA. NO_2^- was excluded due to insufficient observations. For NO_x^-, the regression slopes between reconstruction and observations were 0.99, 0.98, and 0.99, with RMSEs of 0.701, 0.842, and 0.674 μmol kg^{-1} for RF, LightGBM, and GPR, respectively; R^2 values exceeded 0.997 for all models (Fig. 8a). RF and GPR slightly outperformed LightGBM. All models accurately reproduced the NO_x^- profiles (Fig. 8b). The reconstruction errors for DIP were 0.066, 0.079, and 0.064 μmol kg^{-1} for RF, LightGBM, and GPR, respectively.

The corresponding errors for $Si(OH)_4$ were 2.13, 2.48, and 2.32 μmol kg^{-1} (Table 3, Figs. S14–S15).

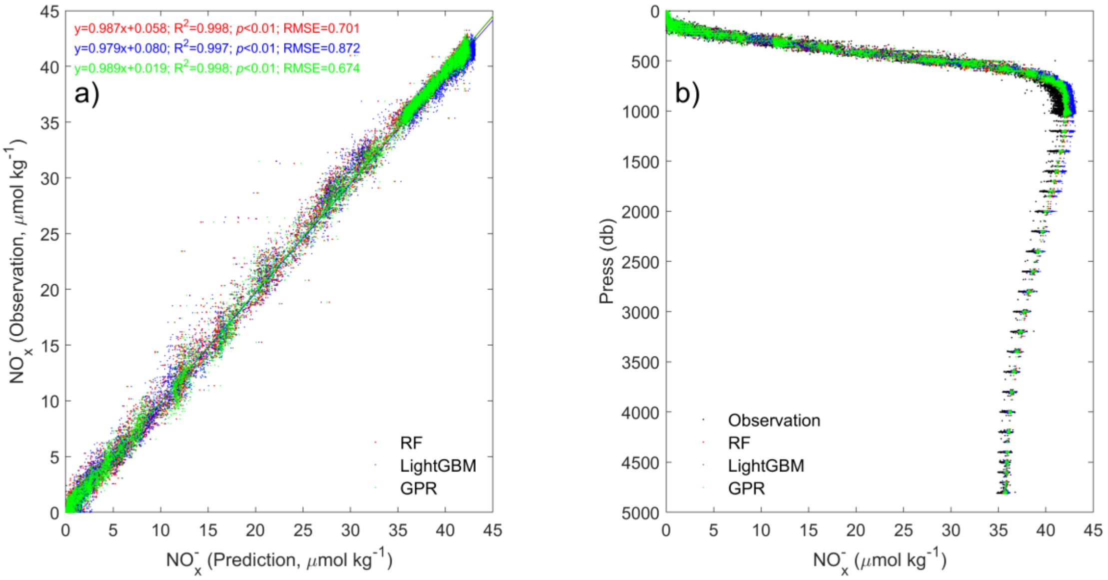

Figure 8. Validating the reconstructed nutrient concentrations at Station ALOHA. a) Reconstructed NO_3^- + NO_2^- (NO_x^-) vs. observations: Random Forest (RF; red dots), LightGBM (blue dots), and Gaussian Process Regression (GPR; green dots). b) Profiles of observed (black dots) and reconstructed NO_x^- from RF (red dots), LightGBM (blue dots), and GPR (green dots).

Since the variations of nutrients primarily occur in the upper water column, we focused on the nutrient reconstruction in the upper 300 m at Station ALOHA. Overall, the models reproduced the profiles of NO_x^- from 1988 to 2021 well (Fig. 9a-d). The reconstruction errors were low at the surface and increased with depth, with most of the values <3.0 μmol kg^{-1} (Fig. S16a-d). To evaluate models' ability to reconstruct nutrient variations in time, the nutrient concentrations were averaged monthly over the upper 300 m. As compared to observations, RF, LightGBM, and GPR all well reconstructed the interannual variations of NO_x^- with most of the absolute errors < 0.5 μmol kg^{-1} (Figs. 9e and S16e) at Station ALOHA. Similarly, the validation of DIP and $Si(OH)_4$ are shown in Figs. S17-S20.

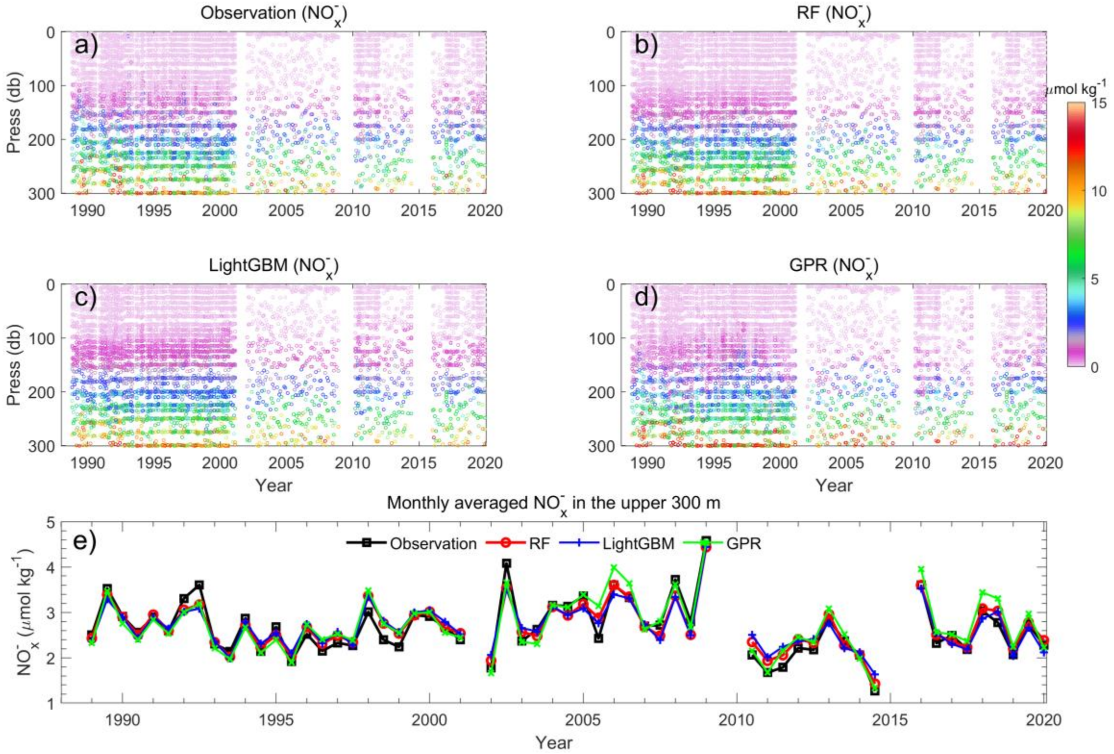

Figure 9. Temporal variations of NO_x^- concentrations in the upper 300 m at Station ALOHA from 1988 to 2021 for observed (a) and reconstructed NO_x^- by Random Forest (RF; b), LightGBM (c), and Gaussian Process Regression (GPR; d). (e) Time series of monthly averaged NO_x^- concentrations in the upper 300 m from observations, and reconstructions by RF, LightGBM, and GPR.

A fifth validation step evaluates the models' reconstruction for the period before 1970 (Error 5; Fig. 6c). This is necessary because the training data (CCHDO) spans 1973–2022, while the reconstructions are extrapolated back to 1895. We argue that this extrapolation should be reasonable because the variations of temperature-salinity-nutrient relationships in the ocean's interior might be small over the past century, providing a basis for temporal extrapolation. First, the residence time of nitrogen in deep and intermediate waters can be up to 2000 years in the North Pacific. Consequently, the imprint of centennial-scale change on nutrient inventories is attenuated. Second, the long-term variations of nutrient concentrations are not evident within our core training period (1973–2022; Figs. 9e and 17). Finally, the mean nutrient profiles derived from the 1920-1970 and 1973-2022 periods are not evidently different in the central North Pacific (Fig. S21). Therefore, while the North Pacific may experience long-term variability, it might be masked by the reconstruction error, and the use of hydrographic properties as predictors for nutrients is justified for historical reconstructions.

However, when assessing the reconstruction errors before 1970, we first consider data quality issues. Prior to the standardization of modern oceanographic methods, nutrient measurements—particularly from earlier decades—were subject to greater analytical errors, inconsistent sampling protocols, and varied determination techniques. The data quality concern is evident in the sporadic and sometimes physically implausible deep nutrient profiles found in WOD for that era (Fig. S22). This is also the primary reason that nutrient data pre-1973 collected from sources like the OSD from WOD were not incorporate into model training. To evaluate data quality in earlier decades, we selected five specific years with more abundant observations: 1929, 1947, 1953, 1958, and 1966 (Fig. S23). After applying the same quality-control criteria outlined in Section 3.1, we used the historical hydrographic data (temperature and salinity) from those years to predict nutrient concentrations. A total of 52,277, 119,137, 284,472, and 193,339 data points were collected for NO_3^-, NO_2^-, DIP, and $Si(OH)_4$, respectively, after QC. The comparison between these predictions and the quality-controlled observations yields the prediction errors for the pre-1970 period (Fig. 6c). The RMSEs from different models suggested values <5.7, <0.40, and <22.9 μmol kg^{-1} for NO_3^-, DIP, and $Si(OH)_4$, respectively (Figs. S24–S26), which are much larger than the corresponding errors for the period after 1970. We recommend that these values be considered a conservative estimate of the upper error bound, as they incorporate both nutrient observations and prediction errors. In addition, the hydrographic data are also less reliable in the earlier period. Thus, we acknowledge that reconstruction errors are likely higher for the pre-1973 period, and the error estimated here should be considered as a "best estimate" with quantified uncertainties, and encourage users to consider these error bounds when applying the dataset to early twentieth-century conditions.

3.2 Reconstructed nutrients

The final reconstructed nutrient dataset aligns with the spatiotemporal coverage of the quality-controlled WOD hydrographic dataset, comprising 472,652,680 data points for each nutrient (NO_x^-, NO_2^-, DIP, and $Si(OH)_4$) from 1,920,634 stations across 35,744 cruises, spanning from 1895 to 2024 (Table 2). Most data points are located above 2,000 m, with fewer observations at greater depths due to hydrographic platform limitations.

It is important to clarify the nature of the reconstructed dataset, which is fundamentally different from gridded products. This product provides nutrient concentrations linked to each hydrographic observations: nutrient values are reconstructed precisely at the locations, depths, and times of original hydrographic observations (sourced from WOD) where direct nutrient measurements might be unavailable or of poor quality. This approach yields a point-wise dataset that aligns with the original hydrographic observations, rather than a spatially or temporally interpolated field—an important distinction for users interpreting and applying the data.

3.3 Climatology of nutrient distributions

To evaluate the reliability of our product, we binned and averaged the predicted nutrients within 1°×1° grid cells for each month to produce a monthly climatology. This climatology represents a mean field that depends heavily on the spatiotemporal distribution of the underlying data and may be influenced by uneven data sampling. This reconstructed climatology was compared with the World Ocean Atlas 2023 (WOA23), which is derived from quality-controlled and objectively analyzed observational data. Since the large-scale patterns of NO_3^-, DIP, and $Si(OH)_4$ are similar among different models (Figs. 10–13, S27–S36), we focus on NO_3^- reconstructed by the RF model in this section unless stated otherwise.

Figs. 10–13 present the monthly climatology of NO_x^- at 5 m, 100 m, 500 m, and 1,000 m in the North Pacific. At 5 m, the reconstructed NO_x^- accurately captures the established spatial patterns, with elevated concentrations in the subpolar gyre, Bering Sea, and equatorial regions, and depleted concentrations in the NPSG (Fig. 10). Seasonally, the basin-averaged surface NO_x^- concentrations display the highest value of 3.50 μmol kg^{-1} in March, in contrast to the lowest value of 1.82 μmol kg^{-1} in

September. These results agree with Yasunaka et al. (2014, 2021), who, using extensive surface nutrient observations (up to 14,000 for nitrate) in the North Pacific, reported similar spatial and seasonal patterns.

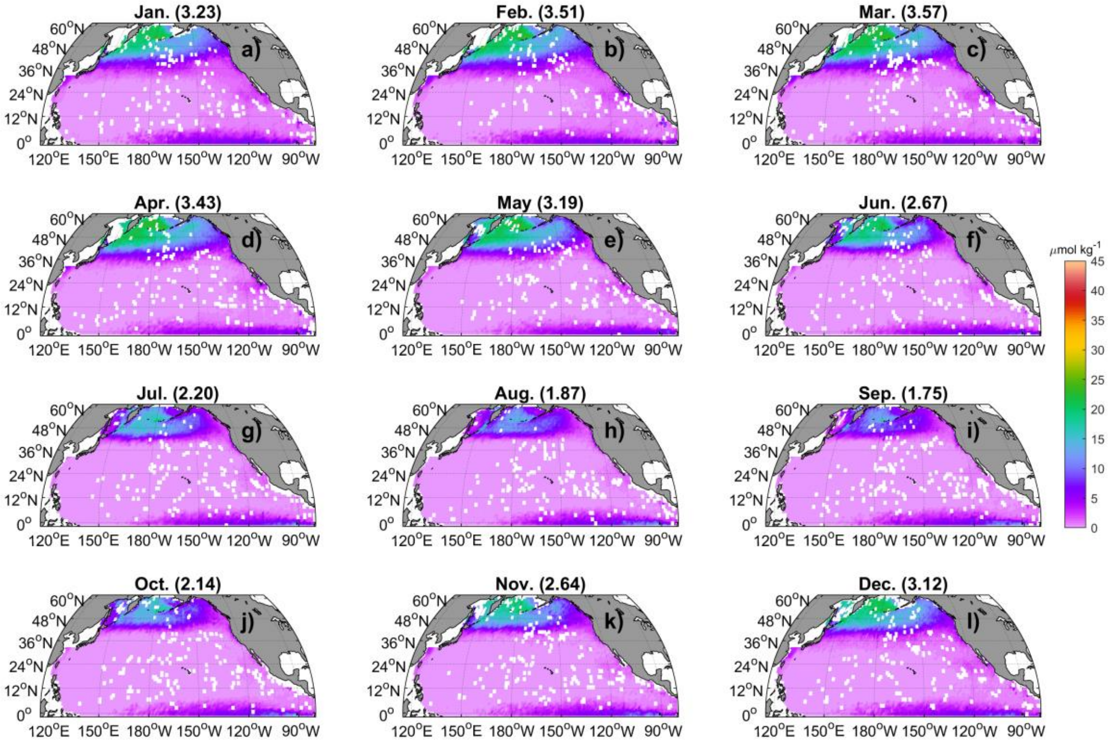

Figure 10. The monthly climatology of NO_x^- at 5 m in the North Pacific. Data are binned and averaged within 1×1° grid cells. The values in the title represent the spatial mean values.

At 100 m, NO_x^- concentrations are elevated particularly in the subarctic gyre, north of the Equator, and the eastern North Pacific, while the central regions, particularly the NPSG, exhibit lower values. At 500 m, NO_x^- concentrations display patterns similar to those at 100 m, except that the NO_x^- concentrations in the western NPSG are evidently lower than those in other regions (Fig. 13). At 1000 m, concentrations in the southwestern North Pacific Ocean are markedly lower than those in other regions (Fig. 12). Below 100 m depth, seasonal variability in NO_x^- is minimal (Figs. 11–13). These results display patterns similar to WOA23 (Figs. S36–S44). The differences between the averaged values of these two climatologies are generally <0.7 μmol kg^{-1} at the surface and <1.5 μmol kg^{-1} at 100 m and 500 m. The maximum differences are found in July at a depth of 500 m (Figs. 13g and S38g). In that month and layer, WOA23 shows a notably low mean NO_3^- value (31.94 µmol kg^{-1}) compared to its values in other months (33.15 to 34.64 µmol kg^{-1}; Fig. S38) and compared to our climatology (33.34 to 33.56 µmol kg^{-1}; Fig. 13). This discrepancy arises because the WOA23 climatology for July features a pronounced low-NO_3^- patch (down to 20 µmol kg^{-1}) within the eastern subarctic gyre, surrounded by waters with concentrations of >35 µmol kg^{-1} (Fig. S38g). These regional differences are clearly visible in the difference maps between the two products (Figs. S45–S47). Generally, our reconstructions capture finer spatial detail, exhibit less oversmoothing, and avoid artificial "bull's-eye" patterns.

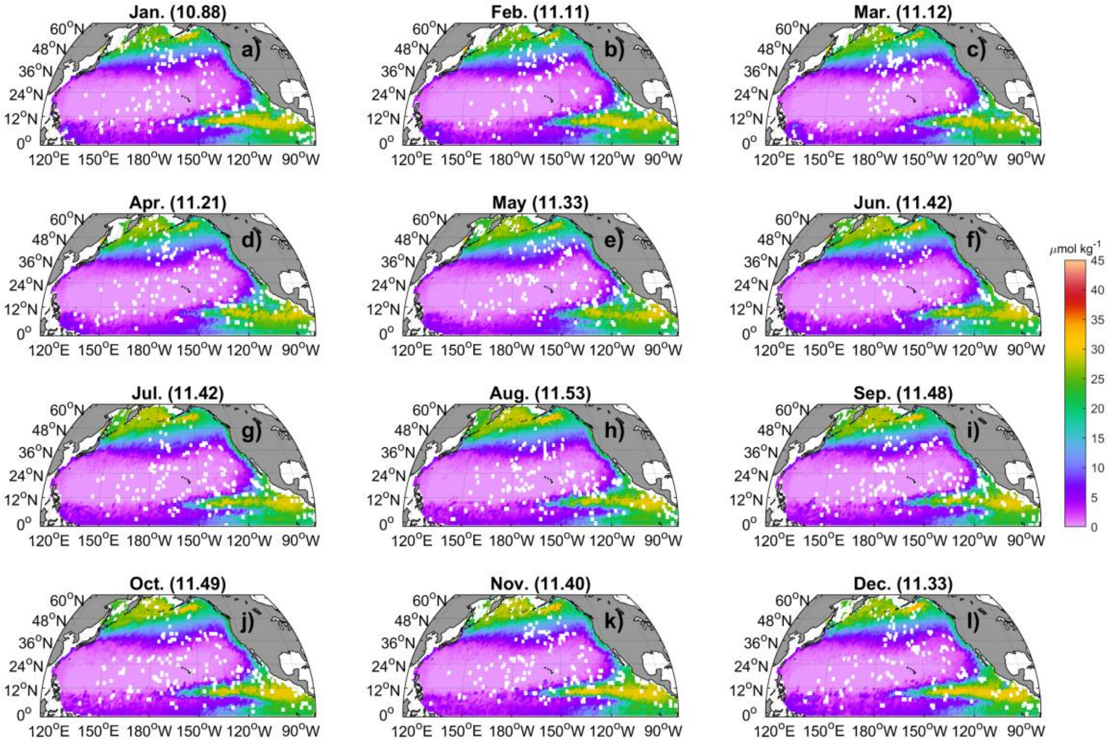

Figure 11. The monthly climatology of NO_x^- at 100 m in the North Pacific. Data are binned and averaged within 1×1° grid cells. The values in the title represent the spatial mean values.

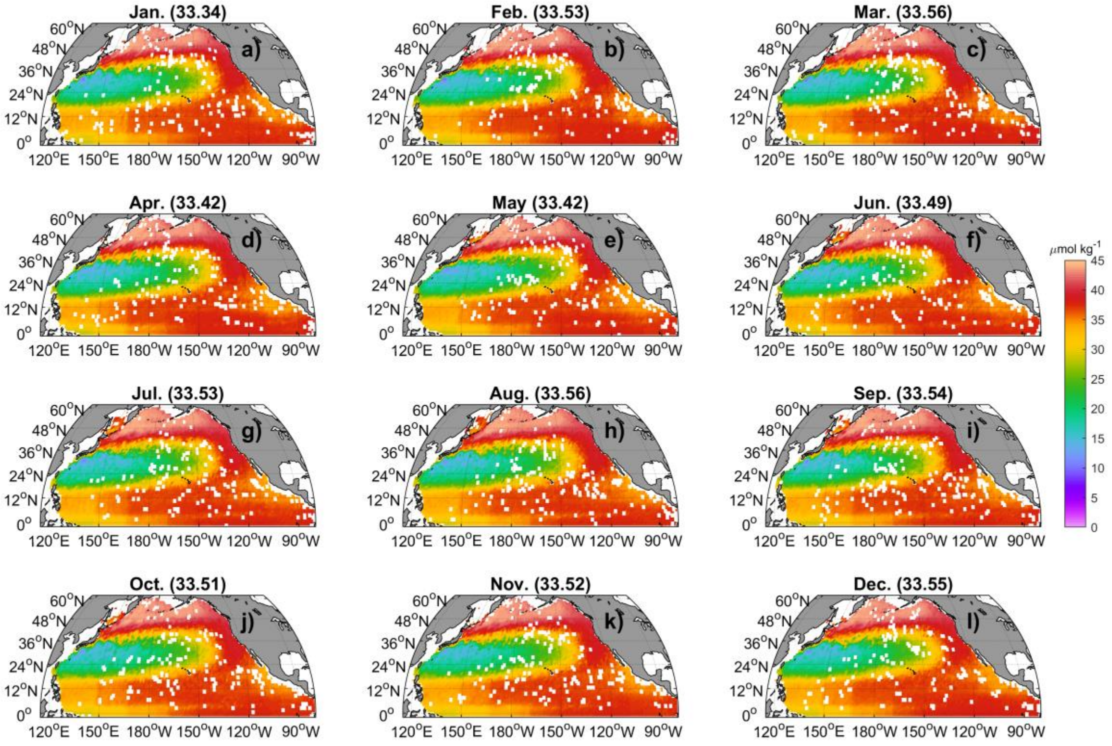

Figure 12. The monthly climatology of NO_x^- at 500 m in the North Pacific. Data are binned and averaged within 1×1° grid cells. The values in the title represent the spatial mean values.

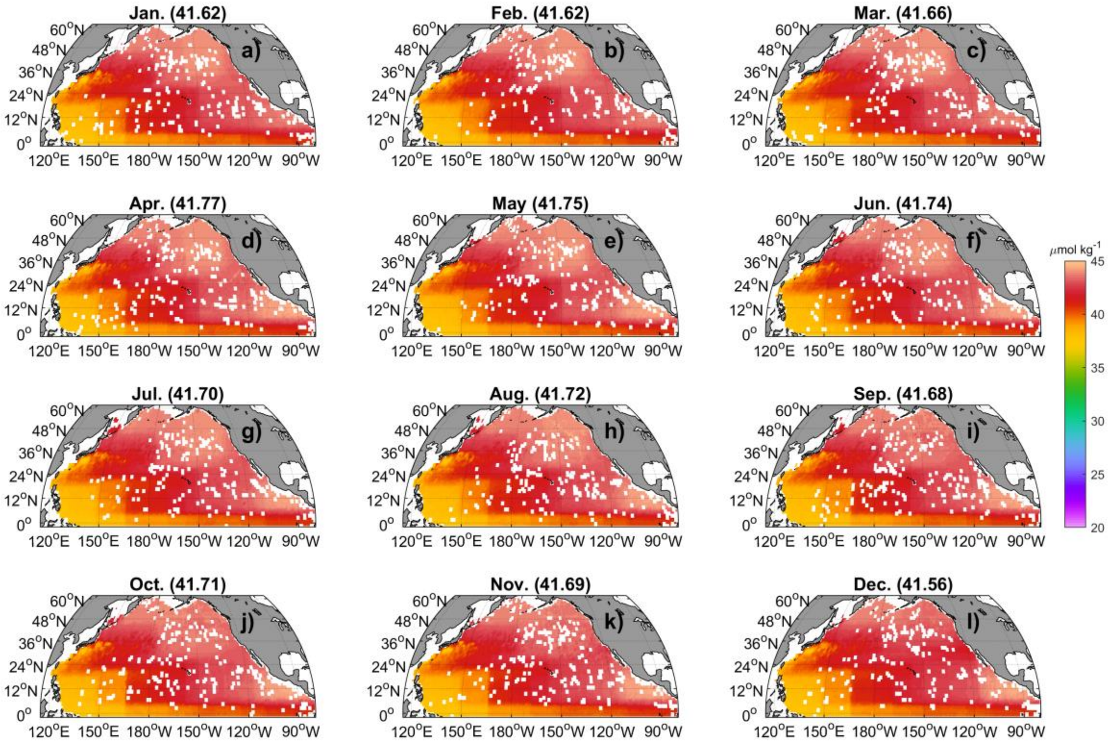

Figure 13. The monthly climatology of NO_x^- at 1000 m in the North Pacific. Data are binned and averaged within 1×1° grid cells. The values in the title represent the spatial mean values.

Sectional distributions of NO_x^- in the upper 2000 m along 10° N and 180° E were used as examples to illustrate the vertical profile distributions of nutrients within the North Pacific. At 10° N, NO_x^- concentrations increase from ~0.0 μmol kg^{-1} at the surface to ~45.0 μmol kg^{-1} at ~1000 m, followed by a decrease to ~38.0 μmol kg^{-1} at 2000 m. NO_x^- concentrations increase from west to the east in the North Pacific in the upper 300 m (Fig. 14). At 180° E, in the upper 500 m, meridional NO_x^- concentrations increase from the equator to the North Equatorial Current (~10° N), decline within the subtropical gyre, and then increase toward the subarctic region (Fig. 15). Generally, seasonal differences of NO_x^- concentrations along both sections are not evident.

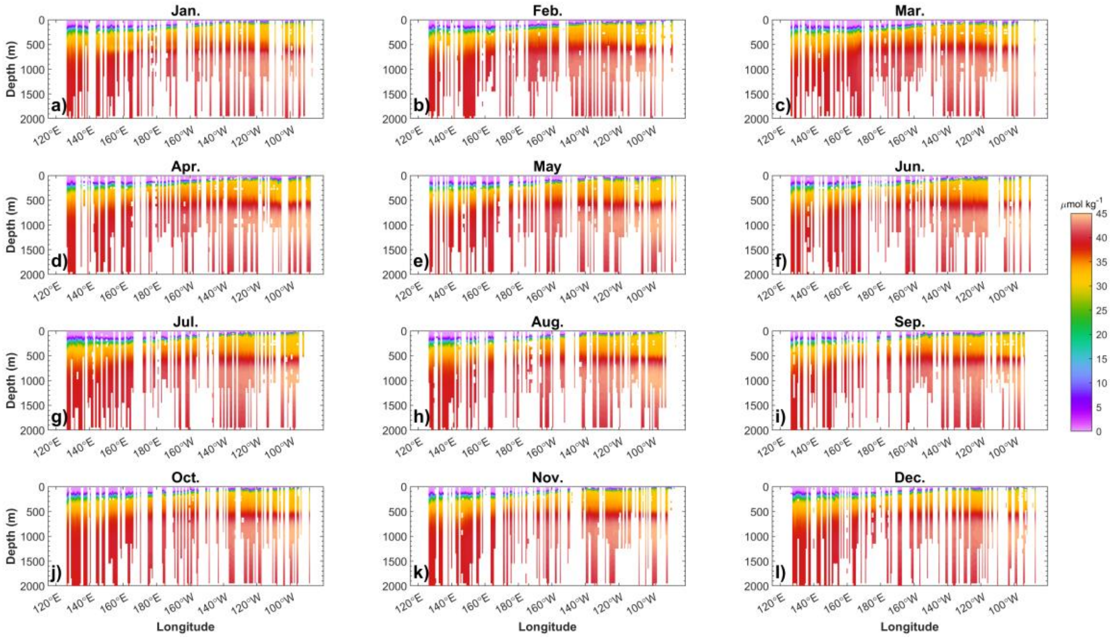

Figure 14. Zonal and monthly climatology of NO_x^- in the upper 2000 m at 10 ºN in the North Pacific. Data were binned and averaged within 1°×1° grid cells.

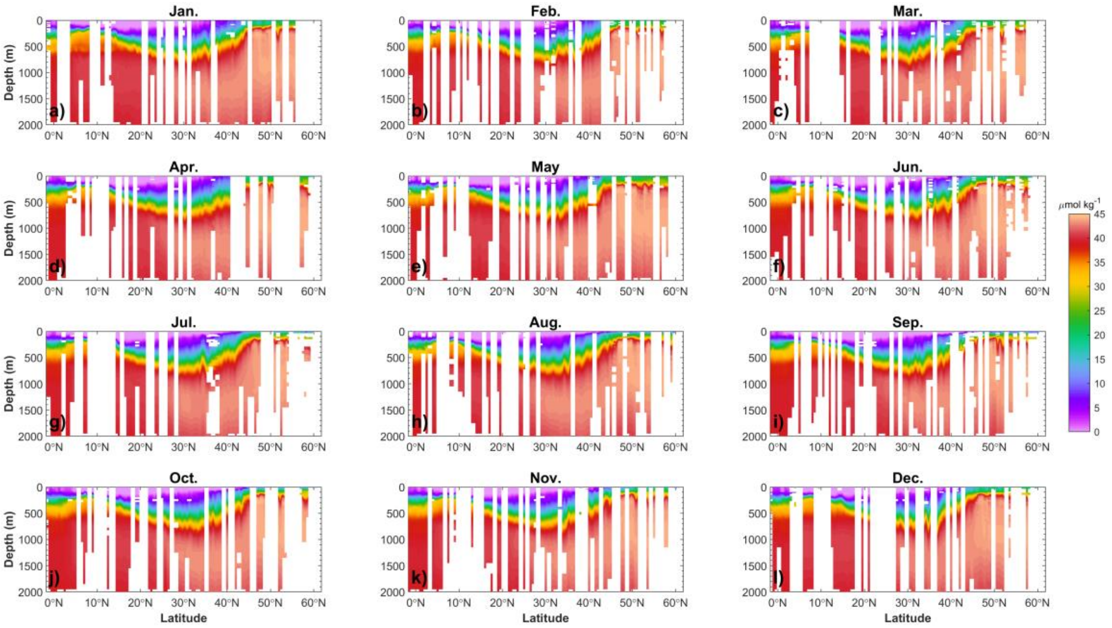

Figure 15. The monthly climatology of NO_x^- in the upper 2000 m at 170 ºE section in the North Pacific. Data were binned and averaged within 1°×1° grid cells.

3.4 Long-term variations of nutrients

We present an initial analysis of long-term nutrient changes by examining five representative regions in the North Pacific, covering the subarctic gyre, the subtropical gyre, and equatorial areas (Fig. 16). The data are binned by region, month, and depth (10 m, 100 m, 200 m, 300 m, 500 m, and 1000 m) for regions 1–5. As shown in Fig. 17, these time series reveal notable interannual fluctuations of NO_3^- (with 2–5-year oscillations), providing a first-order view of low-frequency variability captured by the reconstruction. However, no evident long-term trend is found for nutrients. DIP and $Si(OH)_4$ display patterns similar to NO_3^- (Figs. S48–S49). In contrast, at depths of 200 m and 300 m, NO_2^- displays an increasing trend in the central NPSG and a decreasing trend in the eastern NPSG during the 1970–2005 period (Fig. S50). More sophisticated trend analyses and basin-scale integrations are promising avenues for future work based on this newly reconstructed dataset.

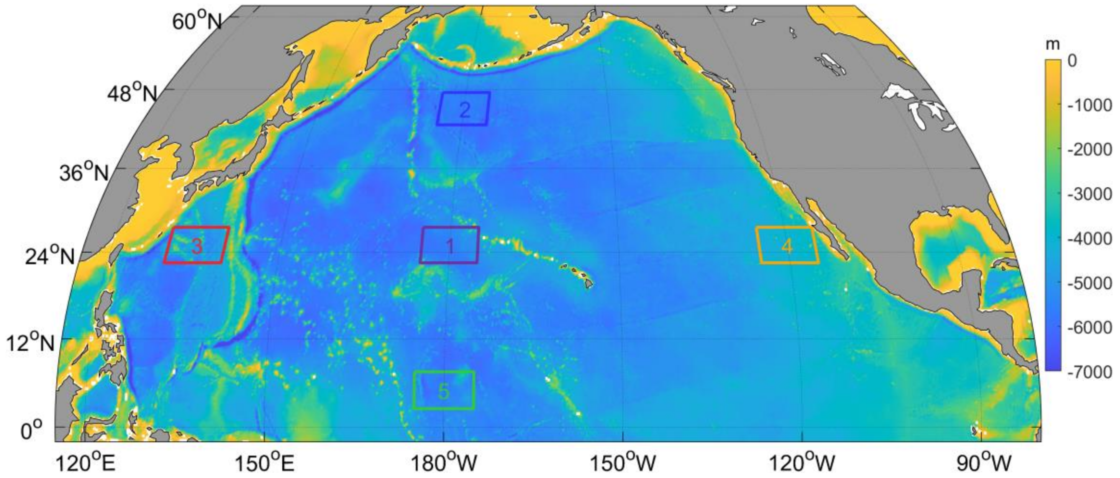

Figure 16. Locations of five representative regions for analyzing long-term nutrient variations.

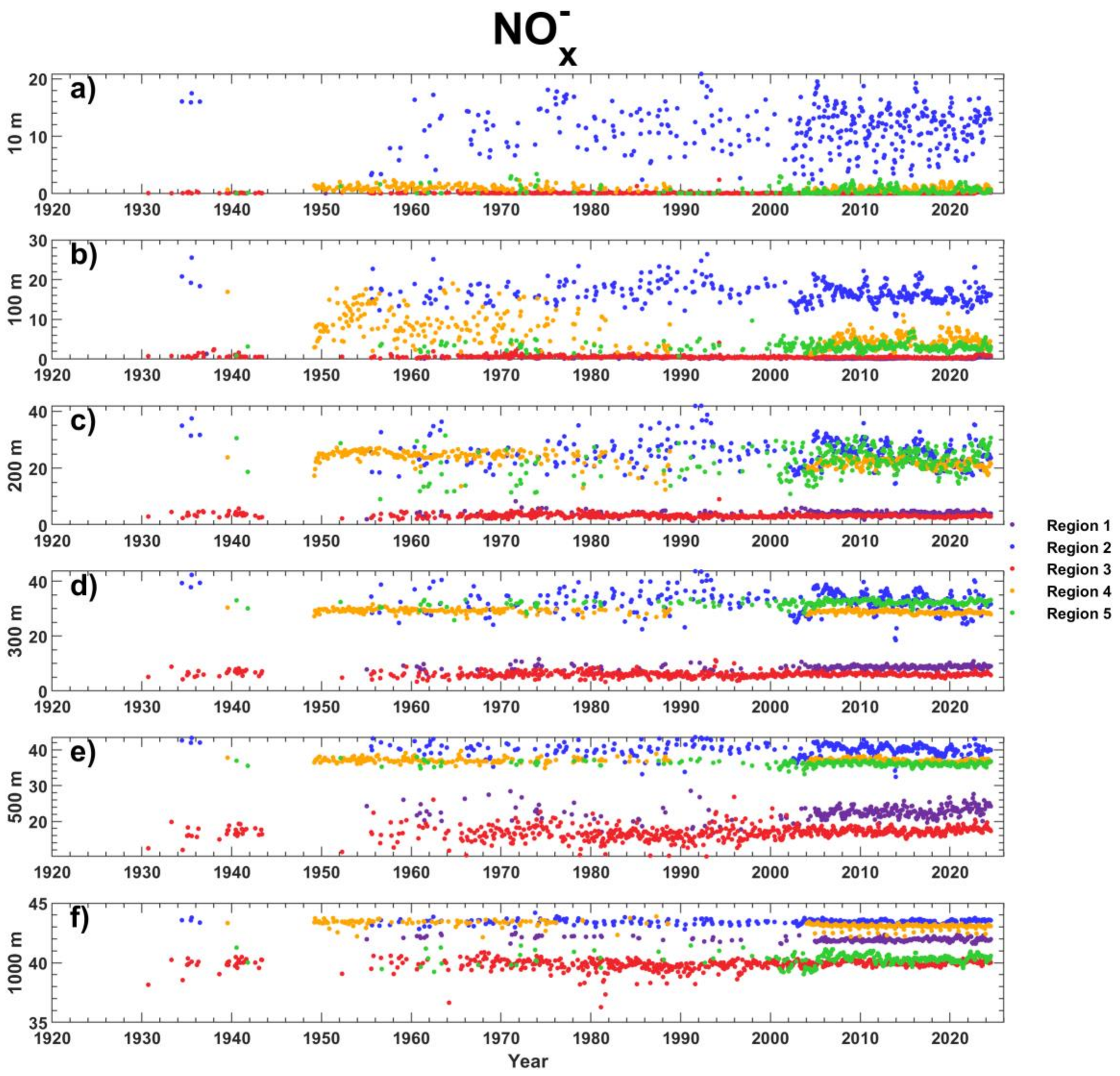

Figure 17. Time series of reconstructed NO_3^- concentrations at 10 m (a), 100 m (b), 200 m (c), 300 m (d), 500 m (e), and 1000 m (f) for regions 1–5 (see Fig. 16). Data were binned by depth and region and then averaged by month.

4 Data availability

The database is available in a data repository (Du et al., 2025; https://zenodo.org/records/17140658). Although the reconstruction results from RF, LightGBM, and GPR are generally consistent, RF yields the best performance. To avoid redundancy and minimize storage requirements—given the large volume of the data files—only the nutrient data reconstructed by RF have been uploaded. Researchers may contact the corresponding authors to request the reconstructions generated by LightGBM and GPR.

5 Conclusion

In this study, we applied rigorous quality control procedures to clean hydrographic and nutrient observations from CCHDO and WOD datasets. The cleaned CCHDO data were then used to train three machine-learning models to relate nutrient concentrations to spatial, temporal, and water-mass predictors. The models were applied to reconstruct nutrient concentrations from hydrographic observations collected from WOD, though most of which lack direct nutrient measurements. We assessed the model performance using four data-partition strategies, and found that all models reproduced held-out data with low RMSEs. RF and GPR slightly outperformed LightGBM. The application of these models to WOD hydrography yielded 472,652,680 reconstructed nutrient concentrations across 1,920,634 stations and 35,744 cruises, spanning from 1895 to 2024. This represents a 2,127– to 2,393-fold increase compared to the original volume of CCHDO nutrient data. The reconstruction captured the spatial, seasonal, and interannual variations of water column nutrients in the North Pacific Ocean well. Compared to the WOA23 climatology, the reconstruction-based nutrient climatology exhibited more realistic spatial structures than WOA23. This high-quality and high-resolution nutrient dataset adds historical nutrient estimation for locations and times with solely hydrographic measurements. The additionally potential application of this dataset include: 1) investigating nutrient transport and budget in the north Pacific; 2) spinning up and validating ocean biogeochemical models; 3) assessing long-term nutrient trends driven by anthropogenic forcing and climate change; 4) investigating nutrient stoichiometric changes and their ecological impacts under climate variability. Collectively, this resource facilitates advanced studies on marine biogeochemical cycles, ecosystem dynamics, and climate-nutrient interactions.

Author contributions

CD and XL designed the study and dataset. CD, SK, MD, ZC, DS, and XL conceived the project and secured the funding. CD, NZ, QL, HW and XL collected and processed the data, developed the code, and performed the analysis. SK, MD, ZC, and DS provided methodological guidance and advice. CD and NZ wrote the original draft. All authors reviewed, edited the manuscript.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (Grants 42494885), National Key Research and Development Program of China (Grant 2023YFF0805001), This study was funded by the National Natural Science Foundation of China (Grants 42576215, 42494881), Innovational Fund for Scientific and Technological Personnel of Hainan Province (Grant KJRC2023B04), and Natural Science Foundation of Hainan Province (Grant 624MS037). We thank the CCHDO (https://cchdo.ucsd.edu/) and the WOD (https://www.ncei.noaa.gov/products/world-ocean-database) for providing the data used in this study. Special thanks are owed to all scientists involved in data collection, analysis, and management for these programs.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the authors used deepseek to check the spelling and grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Arteaga, L., Pahlow, M., and Oschlies, A.: Global monthly sea surface nitrate fields estimated from remotely sensed sea surface temperature, chlorophyll, and modeled mixed layer depth, Geophys. Res. Lett., 42, 1130–1138, 2015.

Ascani, F., Richards, K. J., Firing, E., Grant, S., Johnson, K. S., Jia, Y., et al.: Physical and biological controls of nitrate concentrations in the upper subtropical North Pacific Ocean, Deep-Sea Res. Pt. II, 93, 119–134, 2013.

Barone, B., Church, M. J., Dugenne, M., Hawco, N. J., Jahn, O., White, A. E., et al.: Biogeochemical dynamics in adjacent mesoscale eddies of opposite polarity, Global Biogeochem. Cy., 36, e2021GB007115, 2022.

Benitez-Nelson, C. R., Bidigare, R. R., Dickey, T. D., Landry, M. R., Leonard, C. L., et al.: Mesoscale Eddies Drive Increased Silica Export in the Subtropical Pacific Ocean, Science, 316, 1017–1021, 2007.

Bidigare, R. R., Chai, F., Landry, M. R., Lukas, R., Hannides, C. C. S., Christensen, S. J., Karl, D. M., Shi, L., and Chao, Y.: Subtropical ocean ecosystem structure changes forced by North Pacific climate variations, J. Plankton Res., 31, 1131–1139, 2009.

Bonnet, S., Caffin, M., Berthelot, H., and Moutin, T.: Hot spot of N_2 fixation in the western tropical South Pacific pleads for a spatial decoupling between N_2 fixation and denitrification, Proc. Natl. Acad. Sci. USA, 114, E2800–E2801, 2017.

Browning, T. J. and Moore, C. M.: Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation, Nat. Commun., 14, 5014, 2023.

Chelton, D. B., Schlax, M. G., Samelson, R. M., and de Szoeke, R. A.: Global observations of large oceanic eddies, Geophys. Res. Lett., 34, L15606, 2007.

Chen, S., Hu, C., Barnes, B. B., Wanninkhof, R., Cai, W., Barbero, L., and Pierrot, D.: A machine learning approach to estimate surface ocean $p\mathrm{CO}_2$ from satellite measurements, Remote Sens. Environ., 228, 203–226, 2019.

Chen, S., Meng, Y., Lin, S., Yu, Y., and Xi, J.: Estimation of sea surface nitrate from space: Current status and future potential, Sci. Total Environ., 899, 165690, 2023.

Chen, S., Meng, Y., Shang, S., Zheng, M., Wang, Y., and Chai, F.: Remote estimates of sea surface nitrate and its trends from ocean color in the northwest Pacific, J. Geophys. Res., 129, e2023JC019846, 2024.

Dai, M., Luo, Y., Achterberg, E. P., Browning, T. J., Cai, Y., Cao, Z., Chai, F., Chen, B., Church, M. J., Ci, D., Du, C., Gao, K., Guo, X., Hu, Z., Kao, S., Laws, E. A., Lee, Z., Lin, H., Liu, Q., et al.: Upper Ocean biogeochemistry of the oligotrophic North Pacific subtropical gyre: From nutrient sources to carbon export, Rev. Geophys., 61, e2022RG000800, 2023.

Du, C., Zheng, N., Kao, S.-J., Dai, M., Cao, Z., Shi, D., Li, Q., Wang, H., and Li, X.: Validated temperature and salinity data, and reconstructed nutrient concentrations in the North Pacific (1895 – 2024). Zenodo, https://zenodo.org/records/17451417, 2025.

Dave, A. C. and Lozier, M. S.: Local stratification control of marine productivity in the subtropical North Pacific, J. Geophys. Res., 115, C12032, 2010.

Deutsch, C. and Weber, T.: Nutrient Ratios as a Tracer and Driver of Ocean Biogeochemistry, Annu. Rev. Mar. Sci., 4, 113–138, 2012.

Dong, L., Qi, J., Yin, B., Zhi, H., Li, D., Yang, S., Wang, W., Cai, H., and Xie, B.: Reconstruction of subsurface salinity structure in the South China Sea using satellite observations: a LightGBM-Based Deep forest method, Remote Sens., 14, 3494, 2022.

Du, C., He, R., Liu, Z., Huang, T., Wang, L., Yuan, Z., Xu, Y., Wang, Z., and Dai, M.: Climatology of nutrient distributions in the South China Sea based on a large data set derived from a new algorithm, Prog. Oceanogr., 195, 102586, 2021.

Dugdale, R. C., Morel, A., Bricaud, A., and Wilkerson, F. P.: Modeling new production in upwelling centers: A case study of modeling new production from remotely-sensed temperature and color, J. Geophys. Res., 94, 18119–18132, 1989.

Eugster, O. and Gruber, N.: A probabilistic estimate of global marine N-fixation and denitrification, Glob. Biogeochem. Cycles, 26, GB4013, 2012.

Fuhr, M., Laukert, G., Yu, Y., Nürnberg, D., and Frank, M.: Tracing water mass mixing from the Equatorial to the North Pacific Ocean with dissolved neodymium isotopes and concentrations, Front. Mar. Sci., 7, 603761, 2021.

Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., and Zhu, X.: Application of the Machine Learning LightGBM model to the prediction of the water levels of the Lower Columbia River, J. Mar. Sci. Eng., 9, 496, 2021.

Garcia, H. E., Boyer, T. P., Locarnini, R. A., Reagan, J. R., Mishonov, A. V., Baranova, O. K., Paver, C. R., Wang, Z., Bouchard, C. N., Cross, S. L., Seidov, D., and Dukhovskoy, D.: World Ocean Database 2023: User's Manual. A.V. Mishonov, Technical Ed., NOAA Atlas NESDIS, 98, 129 pp., 2024.

Goes, J. I., Saino, T., Oaku, H., and Jiang, D. L.: A Method for Estimating Sea Surface Nitrate Concentrations from Remotely Sensed SST and Chlorophyll - A Case Study for the North Pacific Ocean Using OCTS/ADEOS Data, IEEE Trans. Geosci. Remote Sens., 37, 1633–1644, 1999.

Hu, C., Feng, L., and Guan, Q.: A machine learning approach to estimate surface chlorophyll *a* concentrations in global oceans from satellite measurements, IEEE Trans. Geosci. Remote Sens., 59, 4590–4607, 2021.

Huang, Y., Nicholson, D., Huang, B., and Cassar, N.: Global estimates of marine gross primary production based on machine learning upscaling of field observations, Global Biogeochem. Cy., 35, e2020GB006718, 2021.

Huang, Y., Tagliabue, A., and Cassar, N.: Data-Driven Modeling of Dissolved Iron in the Global Ocean, Front. Mar. Sci., 9, 837183, 2022.

Kamykowski, D., Zentara, S.-J., Morrison, J. M., and Switzer, A. C.: Dynamic global patterns of nitrate, phosphate, silicate, and iron availability and phytoplankton community composition from remote sensing data, Global Biogeochem. Cy., 16, 1077, 2002.

Kamykowski, D.: A preliminary model of the relationship between temperature and plant nutrients in the upper ocean, Deep-Sea Res., 34, 1067–1079, 1987.

Kamykowski, D.: Estimating upper ocean phosphate concentrations using ARGO float temperature profiles, Deep-Sea Res. Pt. I, 55, 1580–1589, 2008.

Karl, D. M. and Church, M. J.: Ecosystem structure and dynamics in the North Pacific Subtropical Gyre: new views of an old ocean, Ecosystems, 20, 433–457, 2017.

Karl, D. M., Letelier, R. M., Bidigare, R. R., Björkman, K. M., Church, M. J., Dore, J. E., and White, A. E.: Seasonal-to-decadal scale variability in primary production and particulate matter export at Station ALOHA, Prog. Oceanogr., 195, 102563, 2021.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst., 30, 3147–3155, 2017.

Lee, G. S., Lee, J. H., and Cho, H. Y.: Spatiotemporal estimation of nutrient data from the northwest pacific and east Asian seas, Sci. Data, 10, 2023.

Liaw, A. and Wiener, M.: Classification and regression by randomForest, R News, 2, 18–22, 2002.

Lipschultz, F., Bates, N. R., Carlson, C. A., and Hansell, D. A.: New production in the Sargasso Sea: History and current status, Global Biogeochem. Cy., 16, 1001, 2002.

Liu, H., Lin, L., Wang, Y., Du, L., Wang, S., Zhou, P., Yu, Y., Gong, X., and Lu, X.: Reconstruction of Monthly Surface Nutrient Concentrations in the Yellow and Bohai Seas from 2003–2019 Using Machine Learning, Remote Sens., 14, 5021, 2022.

Madani, N., Parazoo, N. C., Manizza, M., Chatterjee, A., Carroll, D., Menemenlis, D., Fouest, V. L., Matsuoka, A., Luis, K. M., Serra-Pompei, C., and Miller, C. E.: A machine learning approach to produce a continuous Solar-Induced chlorophyll fluorescence over the Arctic Ocean, J. Geophys. Res. Machine Learn. Comput., 1, 2024.

Martino, M., Hamilton, D. S., Baker, A. R., Jickells, T., Bromley, T., Nojiri, Y., Quack, B., and Boyd, P. W.: Western Pacific atmospheric nutrient deposition fluxes, their impact on surface ocean productivity, Glob. Biogeochem. Cycles, 28, 712–728, 2014.

Mishonov, A. V., Boyer, T. P., Baranova, O. K., Bouchard, C. N., Cross, S. L., Garcia, H. E., Locarnini, R. A., Paver, C. R., Reagan, J. R., Wang, Z., Seidov, D., Grodsky, A. I., and Beauchamp, J. G.: World Ocean Database 2023, C. Bouchard, Technical Ed., NOAA Atlas NESDIS, 97, 2024.

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., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito, M. A., Thingstad, T., Tsuda, A., and Ulloa, O.: Processes and patterns of oceanic nutrient limitation, Nat. Geosci., 6, 701–710, 2013.

Możejko, J. and Gniot, R.: Application of Neural Networks for the Prediction of Total Phosphorus Concentrations in Surface Waters, Pol. J. Environ. Stud., 17, 363–368, 2008.

Palacios, D. M., Hazen, E. L., Schroeder, I. D., and Bograd, S. J.: Modeling the temperature-nitrate relationship in the coastal upwelling domain of the California Current, J. Geophys. Res., 118, 1–17, 2013.

Qi, J., Yu, Y., Yao, X., Yuan, G., and Gao, H.: Dry deposition fluxes of inorganic nitrogen and phosphorus in atmospheric aerosols over the Marginal Seas and Northwest Pacific, Atmos. Res., 245, 105076, 2020.

Reagan, J. R., Boyer, T. P., García, H. E., Locarnini, R. A., Baranova, O. K., Bouchard, C., Cross, S. L., Mishonov, A. V., Paver, C. R., Seidov, D., Wang, Z., and

Dukhovskoy, D.: World Ocean Atlas 2023, NOAA National Centers for Environmental Information, Dataset, NCEI Accession 0270533, 2024.

Sarangi, P. K., Thangaradjou, T., Kumar, A. S., and Balasubramanian, T.: Development of nitrate algorithm for the southwest bay of bengal water and its implication using remote sensing satellite datasets, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 4, 983–991, 2011.

Sigman, D. M. and Hain, M. P.: The Biological Productivity of the Ocean, Nat. Educ. Knowl., 3, 21, 2012.

Steinhoff, T., Friedrich, T., Hartman, S. E., Oschlies, A., Wallace, D. W. R., and Körtzinger, A.: Estimating mixed layer nitrate in the North Atlantic Ocean, Biogeosciences, 7, 795–807, 2010.

Su, H., Lu, X., Chen, Z., Zhang, H., Lu, W., and Wu, W.: Estimating Coastal Chlorophyll-A Concentration from Time-Series OLCI Data Based on Machine Learning, Remote Sens., 13, 576, 2021.

Sundararaman, H. K. K. and Shanmugam, P.: Estimates of the global ocean surface dissolved oxygen and macronutrients from satellite data, Remote Sens. Environ., 311, 114243, 2024.

Switzer, A. C., Kamykowski, D., and Zentara, S.-J.: Mapping nitrate in the global ocean using remotely sensed sea surface temperature, J. Geophys. Res., 108, 345–359, 2003.

Talley, L. D., Pickard, G. L., Emery, W. J., and Swift, J. H.: Descriptive Physical Oceanography, An Introduction, Sixth Edition, Academic Press, 350–362 pp., 2011.

Wang, C., Su, B., Sun, J., Hu, X., and Liu, J.: A regional ocean database for the Coastal China Sea. Sci Data, 12, 1550, 2025.

Wang, L., Xu, Z., Gong, X., Zhang, P., Hao, Z., You, J., Zhao, X., and Guo, X.: Estimation of nitrate concentration and its distribution in the northwestern Pacific Ocean by a deep neural network model, Deep Sea Res. I, 195, 104005, 2023.

Wang, Z., Wang, G., Guo, X., Hu, J., and Dai, M.: Reconstruction of High-Resolution Sea Surface Salinity over 2003–2020 in the South China Sea Using the Machine Learning Algorithm LightGBM Model, Remote Sens., 14, 6147, 2022.

Yang, G. G., Wang, Q., Feng, J., He, L., Li, R., Lu, W., Liao, E., and Lai, Z.: Can three-dimensional nitrate structure be reconstructed from surface information with artificial intelligence? – A proof-of-concept study, Sci. Total Environ., 924, 171365, 2024.

Yasunaka, S., Mitsudera, H., Whitney, F., and Nakaoka, S.: Nutrient and dissolved inorganic carbon variability in the North Pacific, J. Oceanogr., 77, 3–16, 2021.

Yasunaka, S., Nojiri, Y., Nakaoka, S., Ono, T., Whitney, F. A., and Telszewski, M.: Mapping of sea surface nutrients in the North Pacific: Basin-wide distribution and seasonal to interannual variability, J. Geophys. Res. Oceans, 119, 7756–7771, 2014.

Yasunaka, S., Ono, T., Nojiri, Y., Whitney, F. A., Wada, C., Murata, A., Nakaoka, S., and Hosoda, S.: Long-term variability of surface nutrient concentrations in the North Pacific, Geophys. Res. Lett., 43, 3389–3397, 2016.

Yu, X. R., Wen, Z., Jiang, R., Yang, J.-Y. T., Cao, Z., Hong, H., Zhou, Y., and Shi, D.: Assessing N2 fixation flux and its controlling factors in the (sub)tropical western North Pacific through high-resolution observations, Limnol. Oceanogr. Lett., 9, 716 ‑ 724, 2024.

Zhong, A., Wang, D., Gong, F., Zhu, W., Fu, D., Zheng, Z., Huang, J., He, X., and Bai, Y.: Remote sensing estimates of global sea surface nitrate: Methodology and validation, Sci. Total Environ., 950, 175362, 2024.