Supplement of Earth Syst. Sci. Data, 18, 2951–2969, 2026
https://doi.org/10.5194/essd-18-2951-2026-supplement

Open Access Earth System Science Data

Supplement of

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

Chuanjun Du et al.

Correspondence to: Chuanjun Du (cjdu@hainanu.edu.cn) and Xiaolin Li (xlli@xmu.edu.cn)

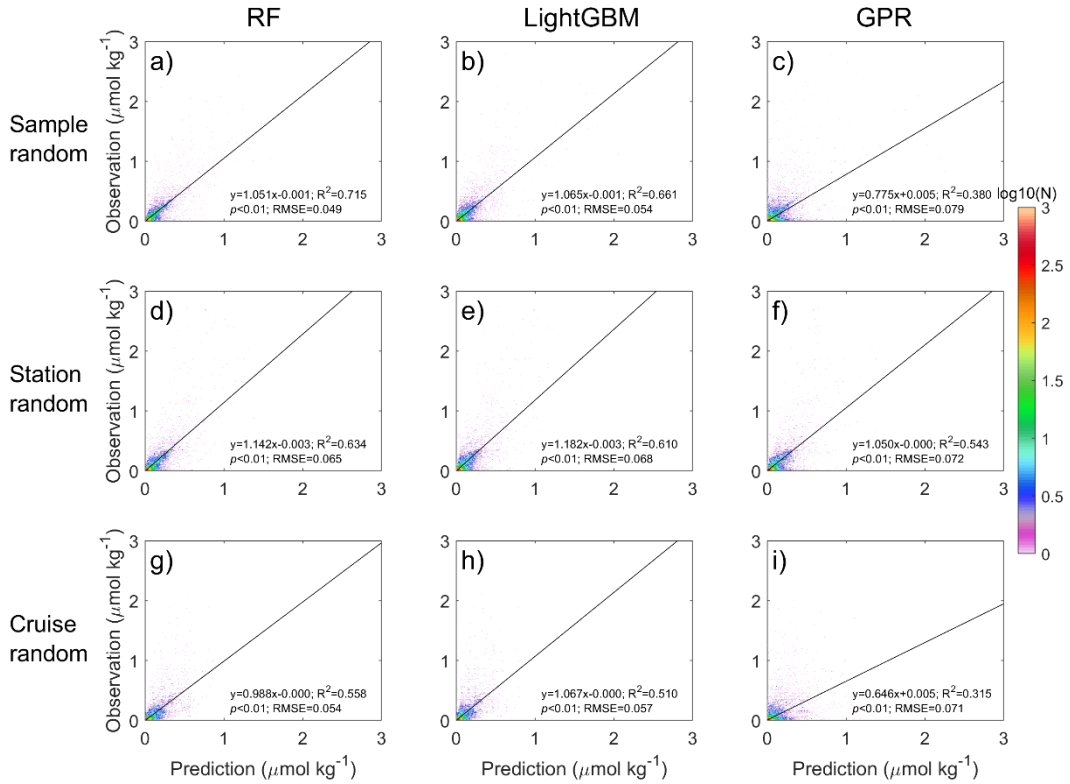


Figure S1. Validating the reconstructed NO_2^- concentrations using leave-one-out cross-validation with different data selection strategies and machine learning methods. Plots shown in row 1 correspond to the sample random strategy (a-c), row 2 correspond to the station random strategy (d-e), and row 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 (RMSE). The color represents the data density (N, number of observations). Note that the logarithmic scale of N is applied.

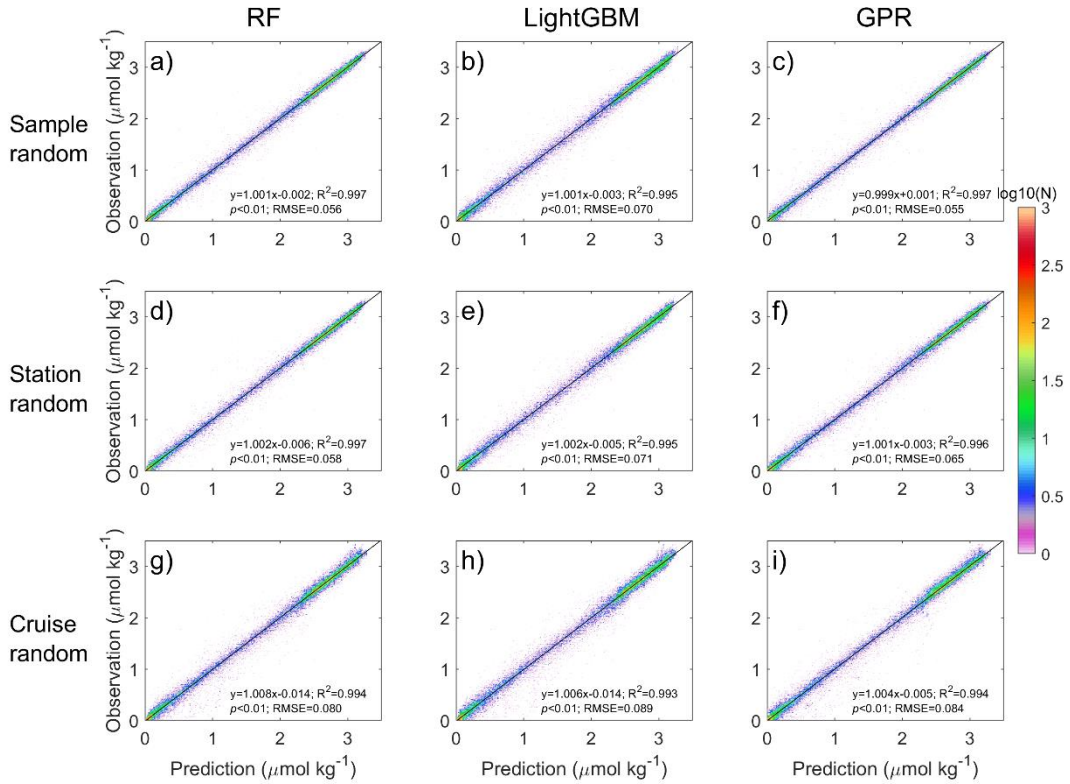


Figure S2. Similar to Fig. S1, but for DIP.

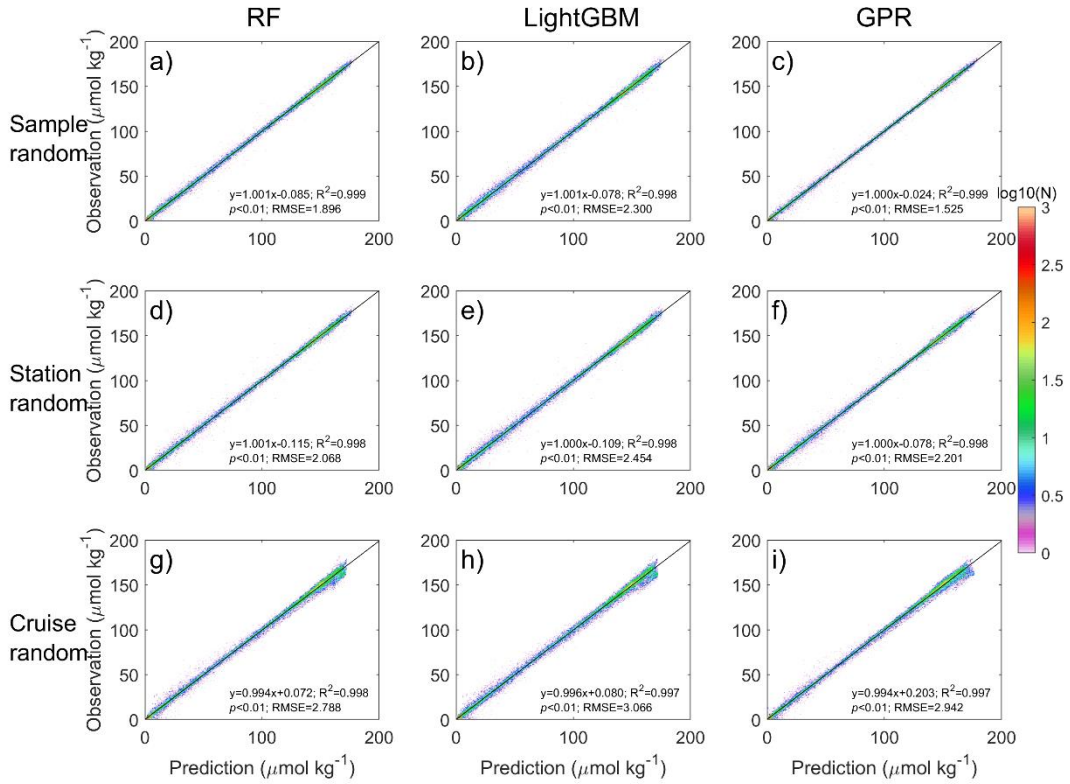


Figure S3. Similar to Fig. S1, but for $Si(OH)_4$.

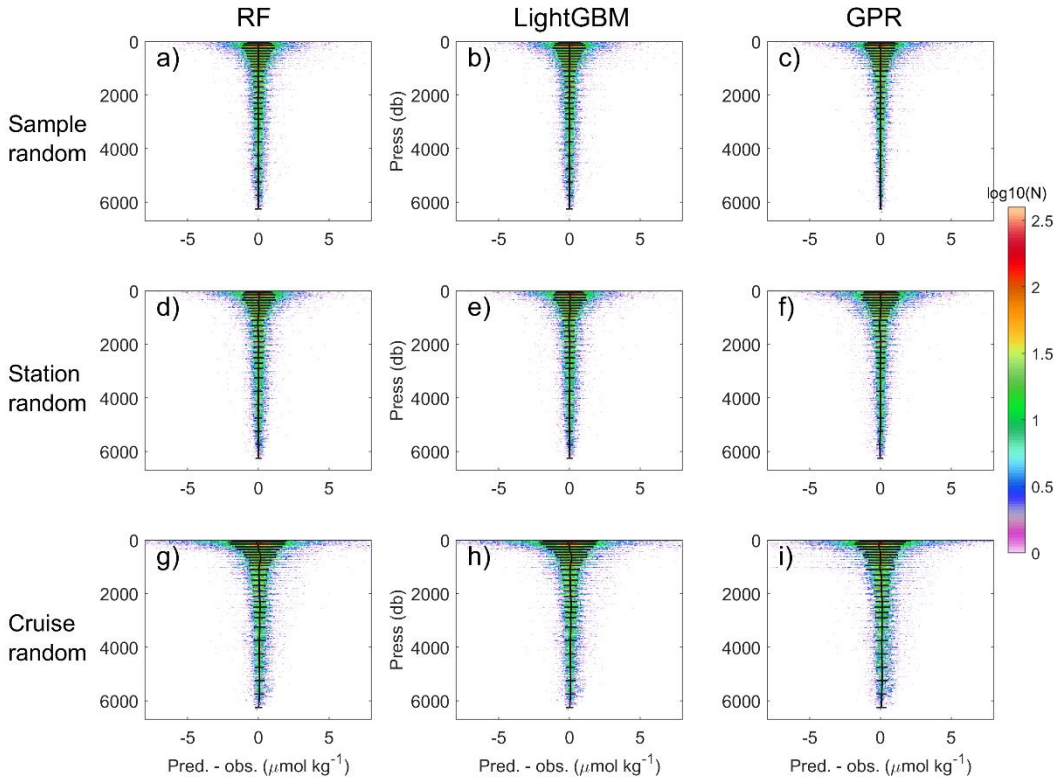


Figure S4. Vertical profiles of NO_x^- (NO_3^- + NO_2^-) reconstruction errors with different data selection strategies and machine learning methods. Colors indicate data density; the solid black vertical line represents the mean absolute error (μmol kg^{-1}); and the horizontal error bars denote the ±1 standard deviation range. Plots in row 1 correspond to the sample random strategy (a-c); plots in row 2 correspond to the station random strategy (d-f); and plots in row 3 correspond to the cruise random strategy (g-i). Plots in column 1 correspond to the Random Forest (RF; a, d, g); plots in column 2 correspond to LightGBM (b, e, h); and plots in column 3 correspond to Gaussian Process Regression (GPR; c, f, i).

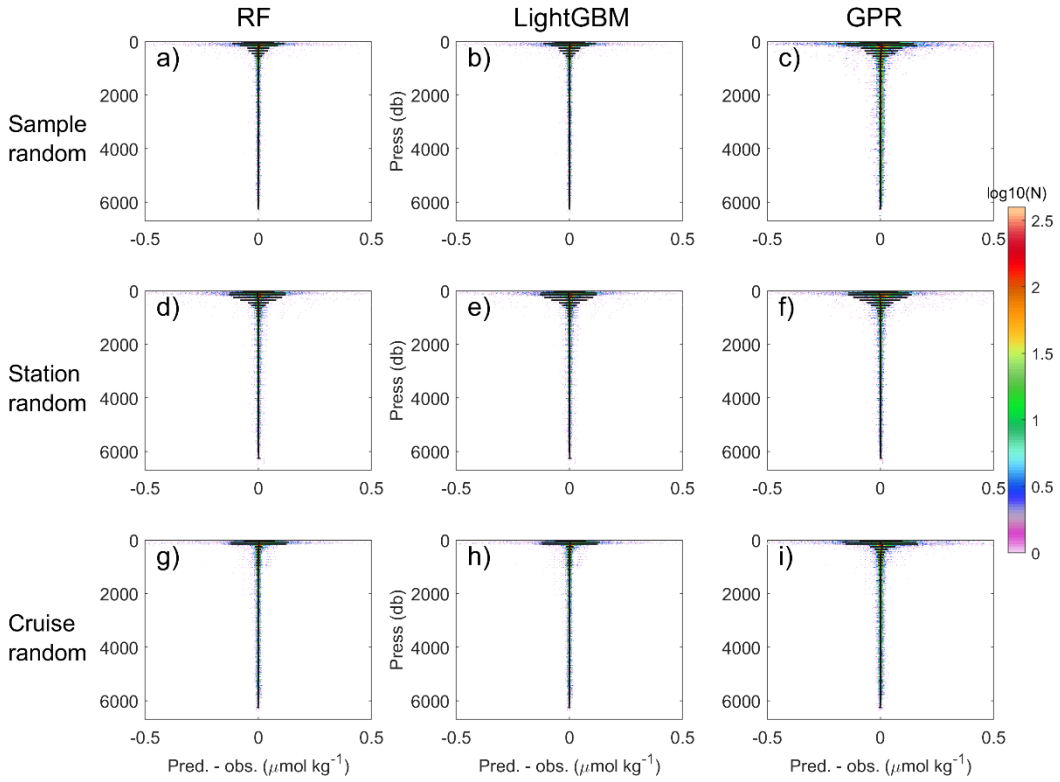


Figure S5. Same as Fig. S4, but for NO_2^-.

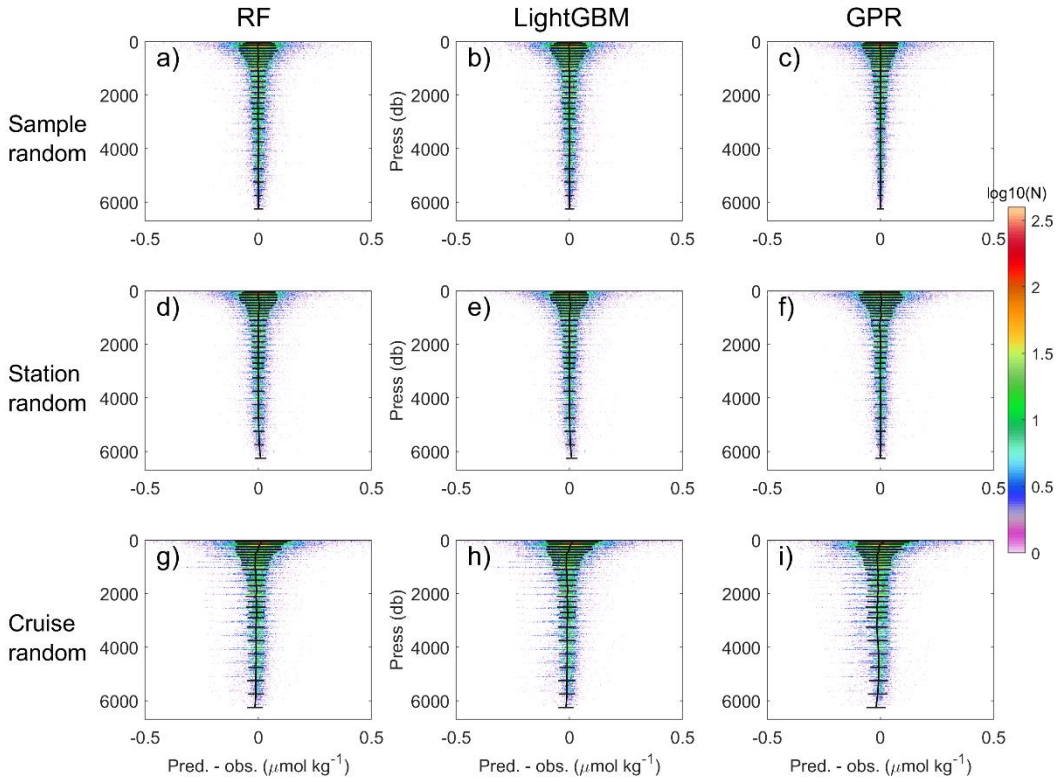


Figure S6. Same as Fig. S4, but for DIP.

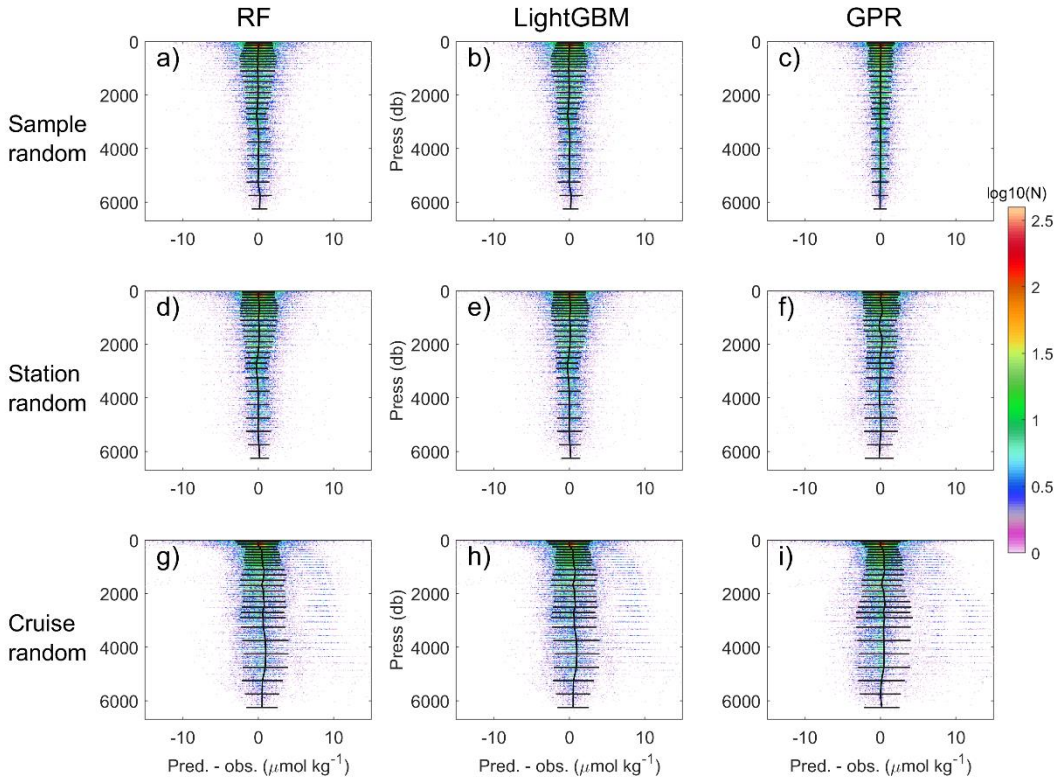


Figure S7. Same as Fig. S4, but for $Si(OH)_4$.

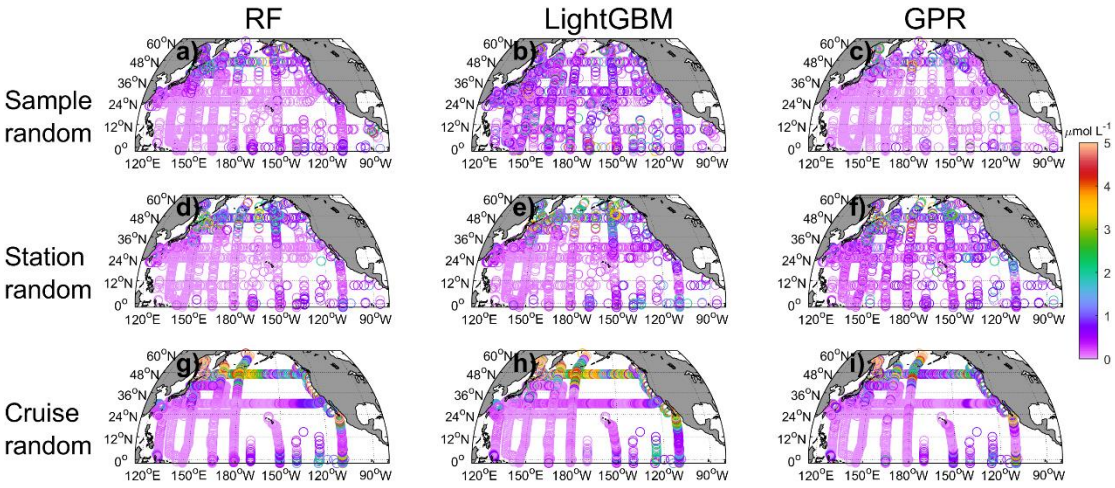


Figure S8. Spatial distribution of absolute errors (absolute value of predicted subtract observed values) for surface NO_x^- concentrations with different data selection strategies and machine learning methods. Plots shown in row 1 correspond to the sample random strategy (a-c), row 2 correspond to the 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).

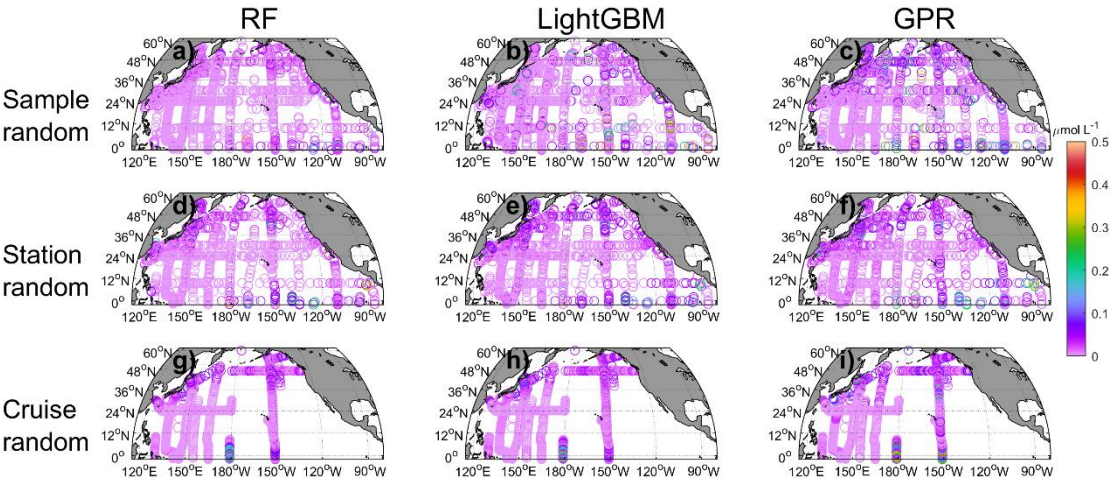


Figure S9. Same as Fig. S8, but for NO_2^-.

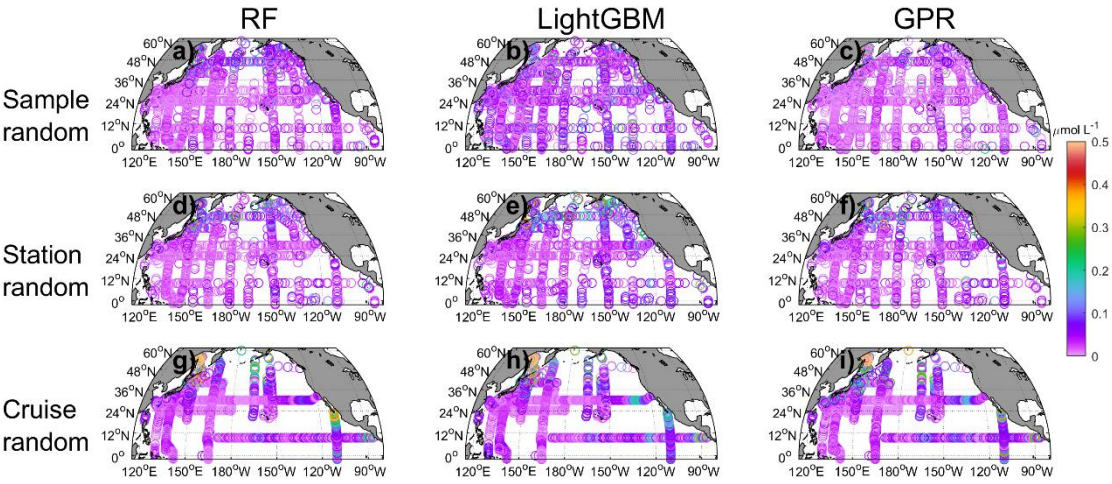


Figure S10. Same as Fig. S8, but for DIP.

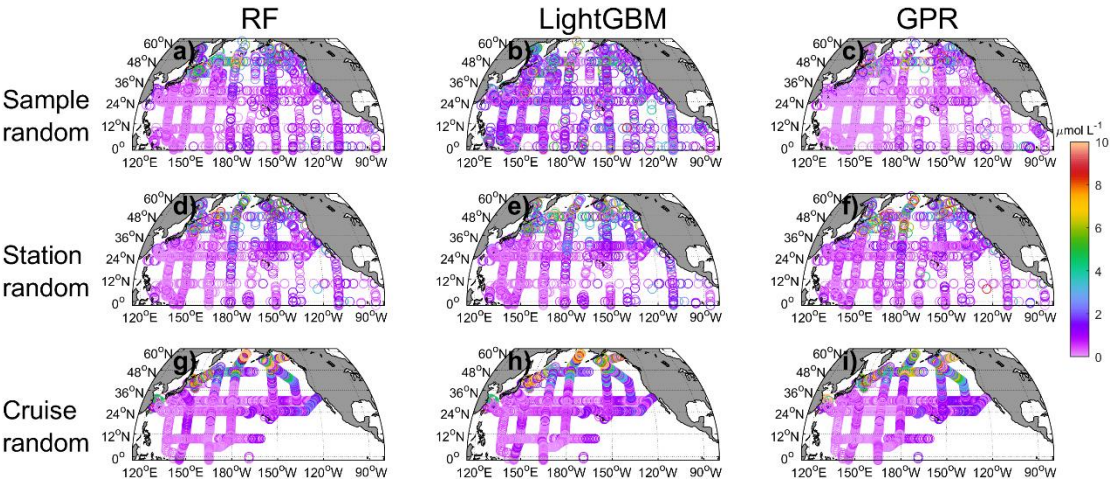


Figure S11. Same as Fig. S8, but for $Si(OH)_4$.

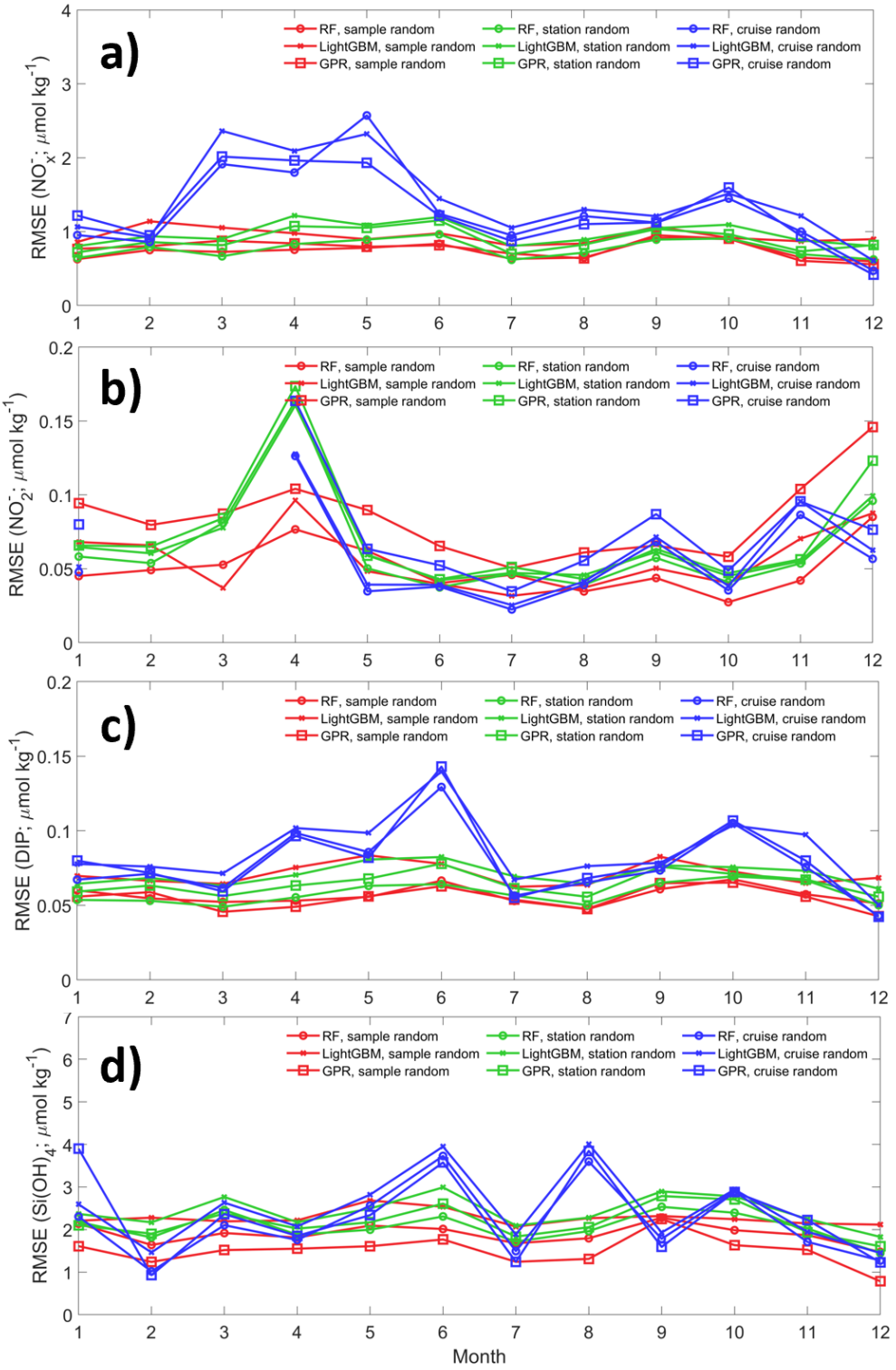


Figure S12. Monthly variations in reconstruction errors (root mean square error) for different parameters under various data selection strategies and machine learning methods.

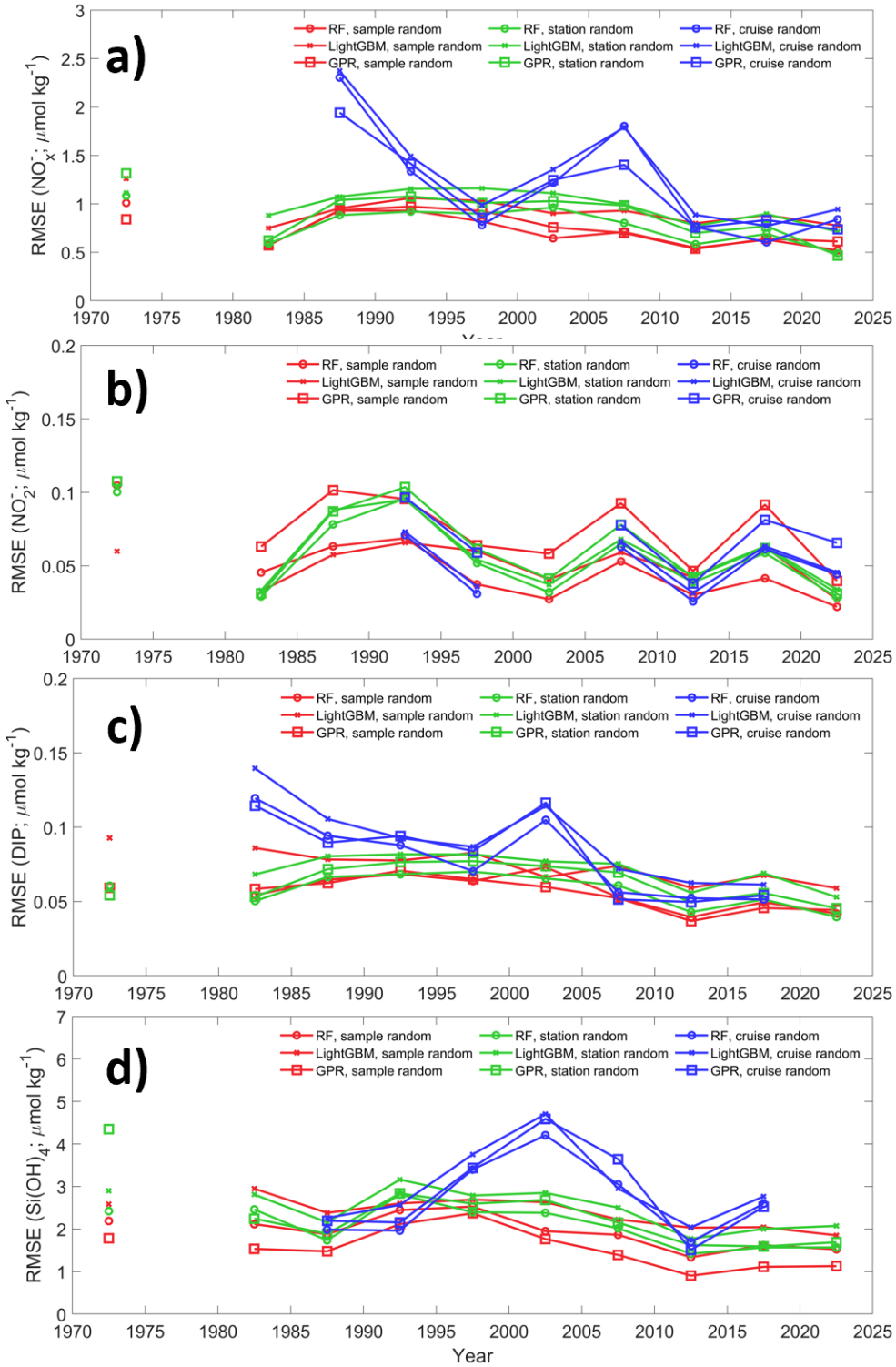


Figure S13. Long-term variations in reconstruction error (root mean square error) under different data selection strategies and machine learning methods.

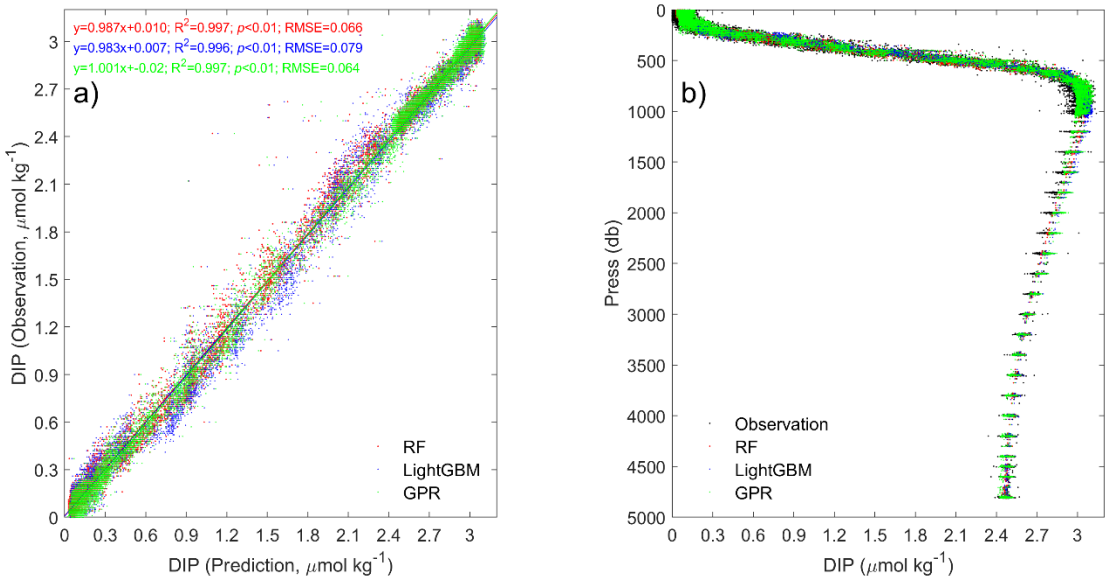


Figure S14. Validating the reconstructed DIP concentrations at Station ALOHA. a) Reconstructed DIP vs. observations: Random Forest (RF; red dots), LightGBM (blue dots), and Gaussian Process Regression (GPR; green dots); b) Profiles of observed (black dots) and reconstructed DIP from RF (red dots), LightGBM (blue dots), and GPR (green dots).

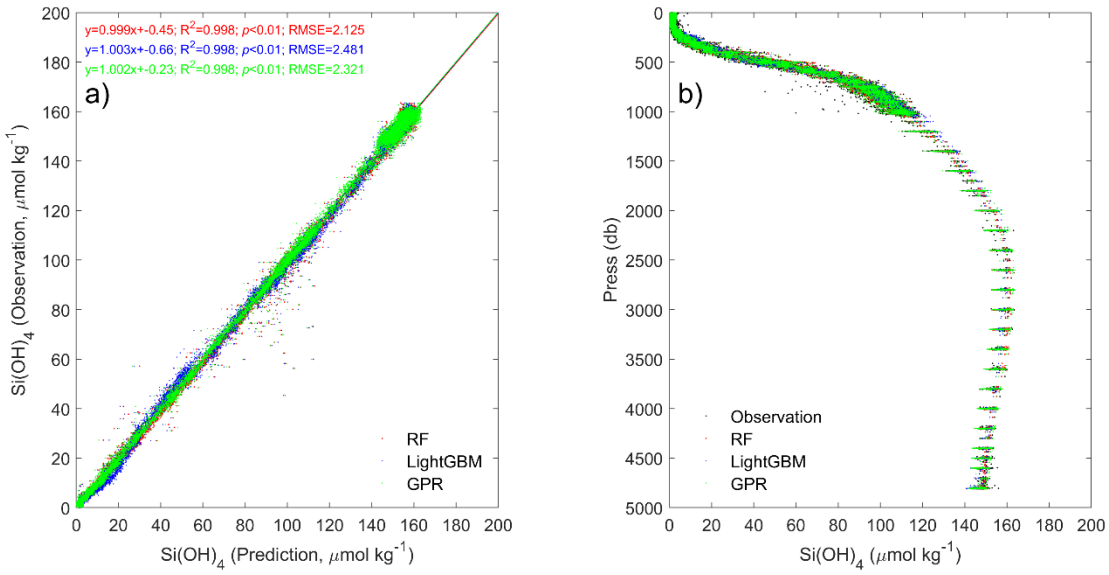
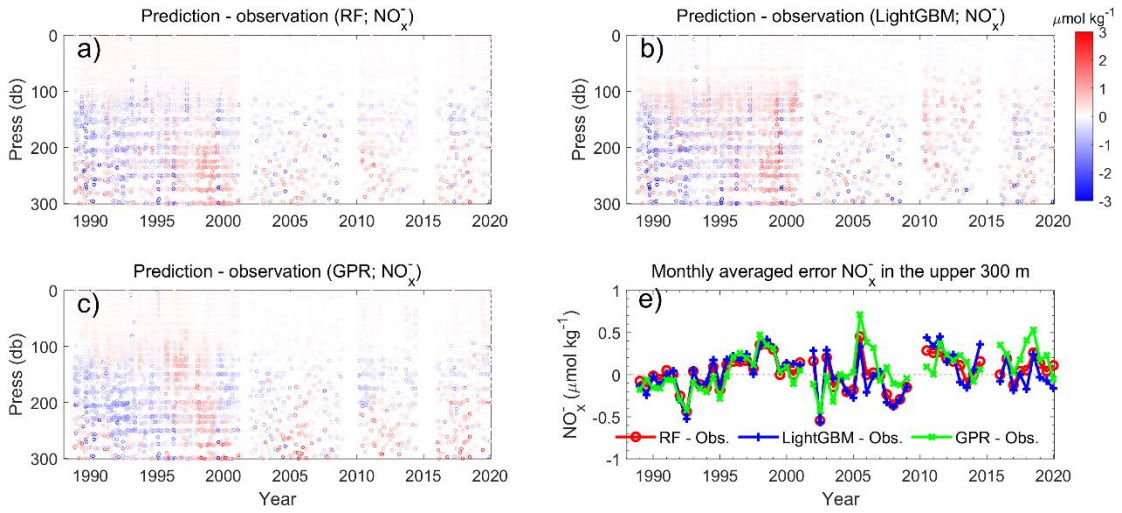


Figure S15. Similar to Fig. S14, but for $Si(OH)_4$.



92

93 **Figure S16.** Temporal variations of reconstruction error profiles of NO_x^- ($NO_3^- + NO_2^-$)
94 (prediction minus observations) in the upper 300 m at Station ALOHA from 1988 to
95 2021 with different machine learning models: Random Forest (RF; a), LightGBM (b),
96 and Gaussian Process Regression (GPR; c). Note that the colorbar plotted in (b) is also
97 applicable to (a) and (c). (e) Temporal variations of monthly averaged NO_x^- prediction
98 errors (prediction minus observations) in the upper 300 m from RF, LightGBM, and
99 GPR.

100

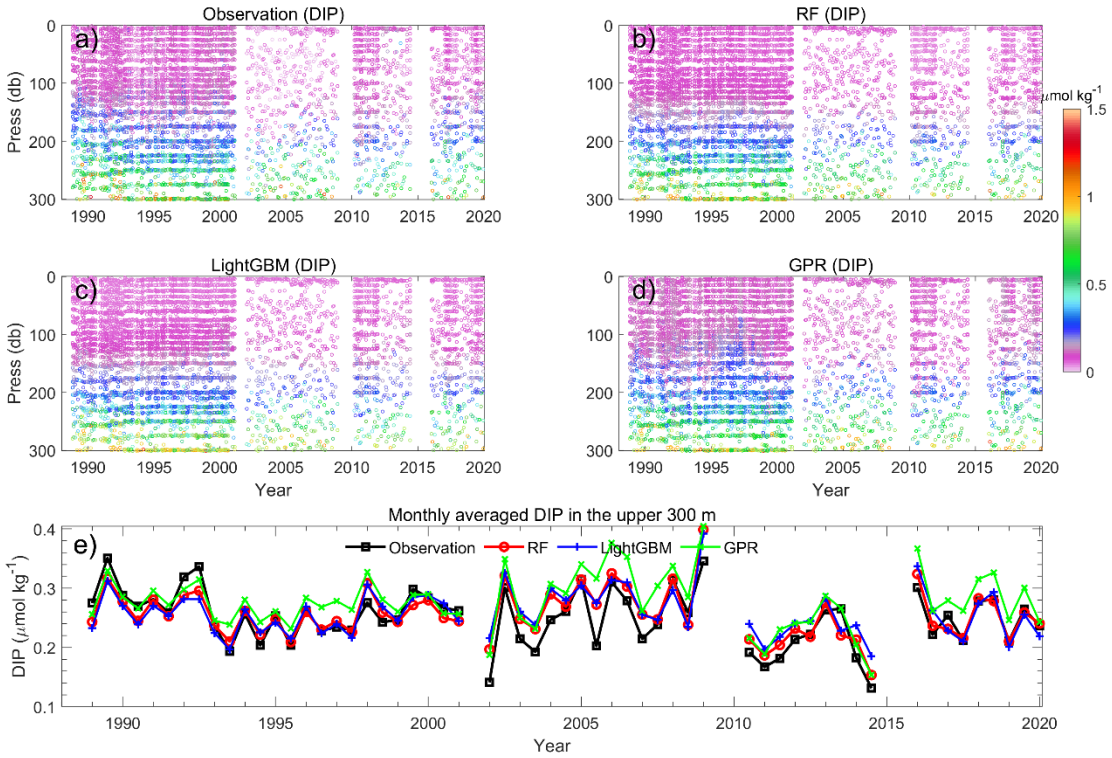


Figure S17. Temporal variations of DIP concentrations in the upper 300 m at Station ALOHA from 1988 to 2021 for observed (a) and reconstructed DIP by Random Forest (RF; b), LightGBM (c), and Gaussian Process Regression (GPR; d). (e) Time series of monthly averaged NO_x^- concentrations in the upper 300 m from observations, and reconstructions by RF, LightGBM, and GPR, respectively.

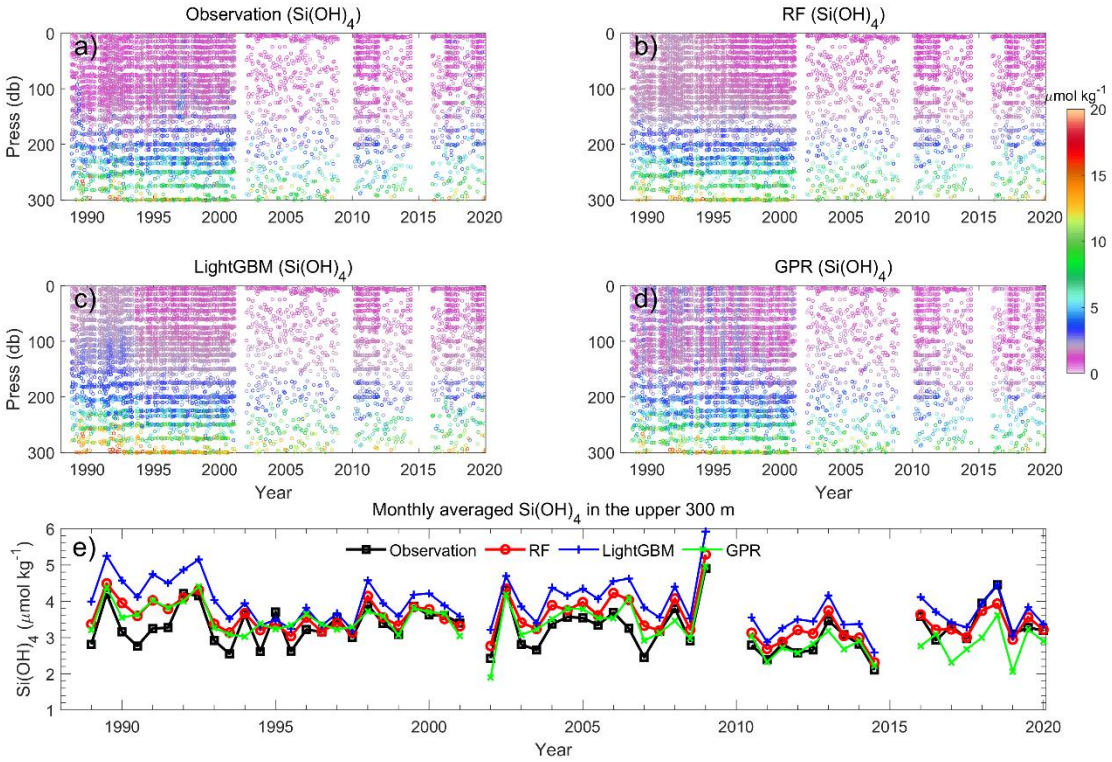


Figure S18. Similar to Fig. S17, but for $Si(OH)_4$.

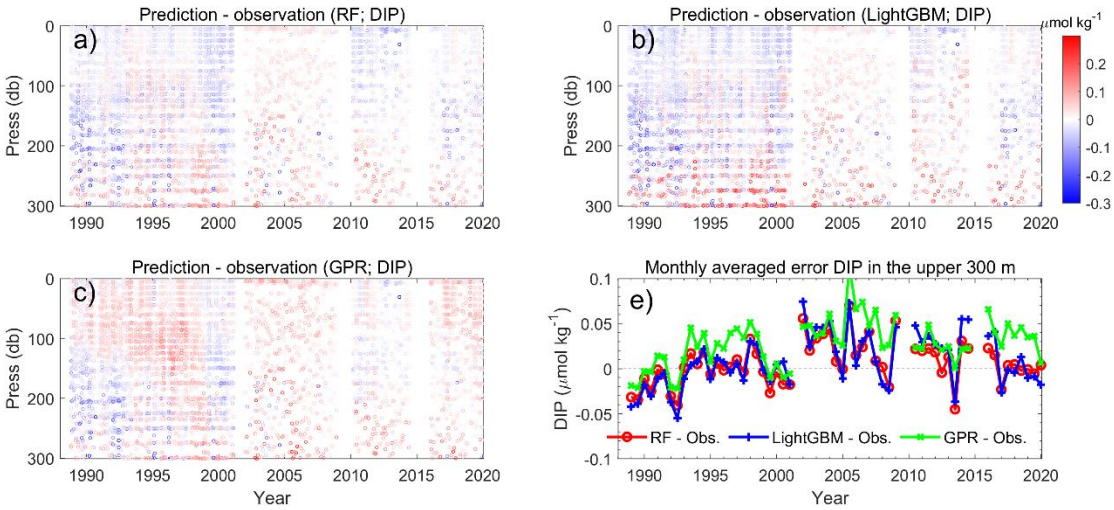


Figure S19. Same as Fig. S16, but for DIP.

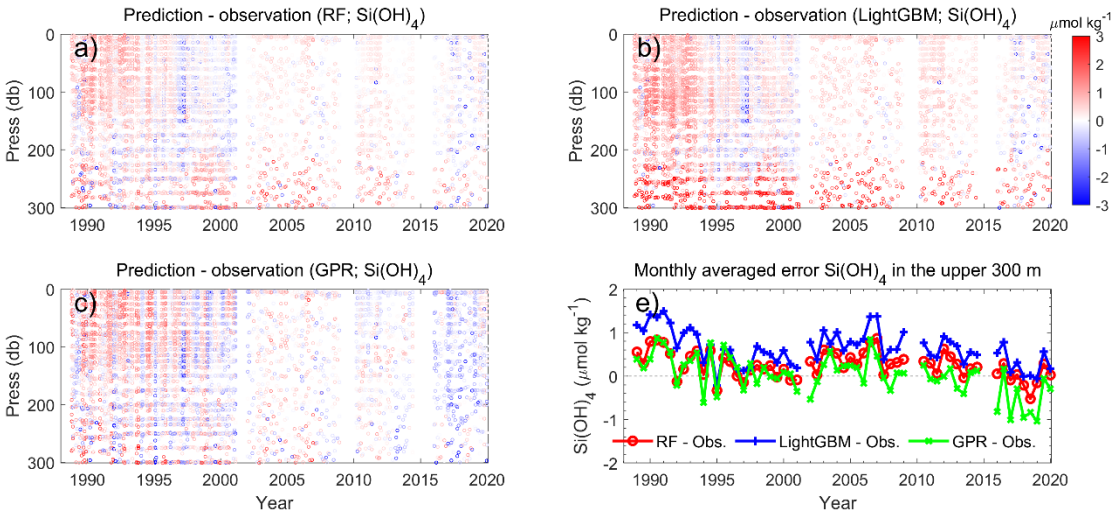


Figure S20. Similar to Fig. S16, but for $Si(OH)_4$.

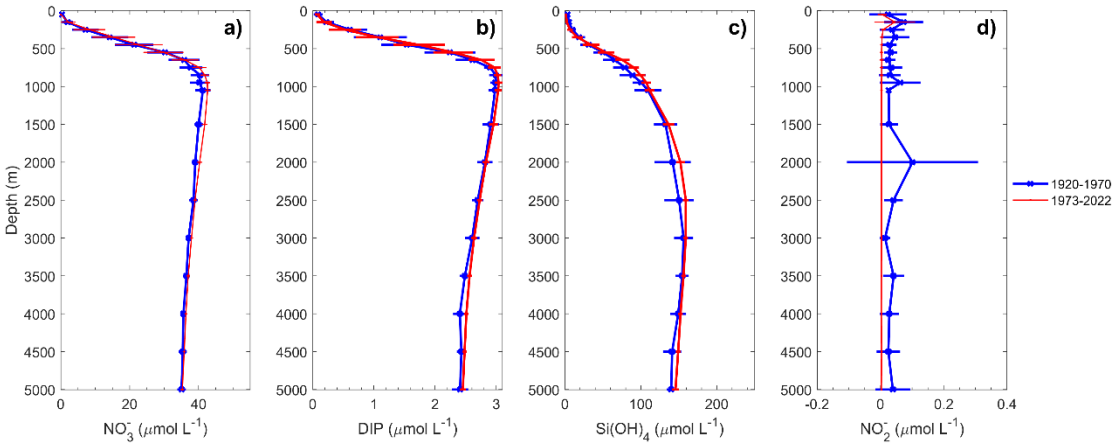


Figure S21. Mean profiles (vertical lines) with standard deviations (horizontal lines) of nutrients between 1920–1970 (Ocean Station Data in the World Ocean Database after quality control) and 1973–2022 (CCHDO data) in the North Pacific basin region (180°E–150°W, 15°N–30°N): (a) NO_3^-; (b) DIP; (c) $Si(OH)_4$; (d) NO_2^-.

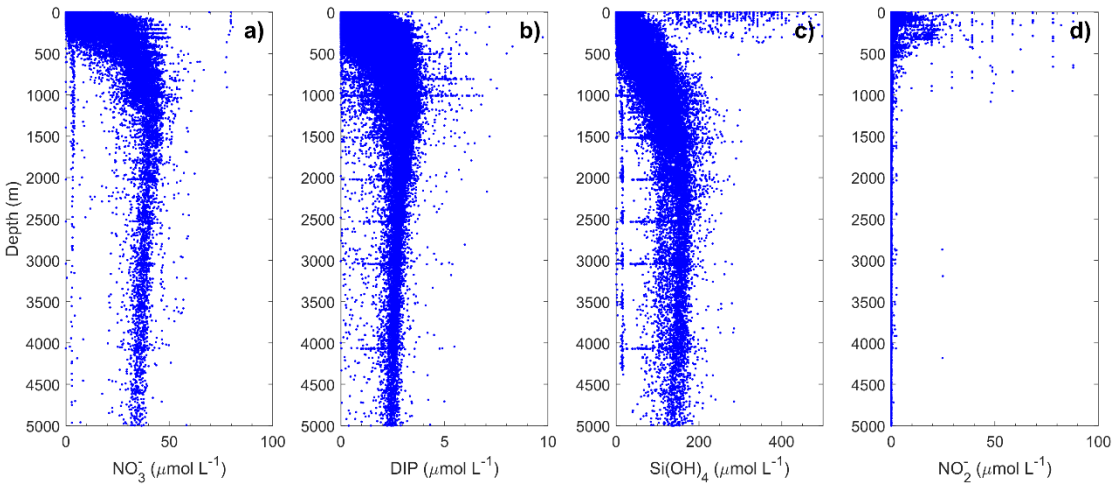


Figure S22. Profiles of nutrients collected from the Ocean Station Data in the World Ocean Database (before 1970) in the North Pacific basin: (a) NO_3^-; (b) DIP; (c) $Si(OH)_4$; (d) NO_2^-, illustrating the scattered vertical distribution.

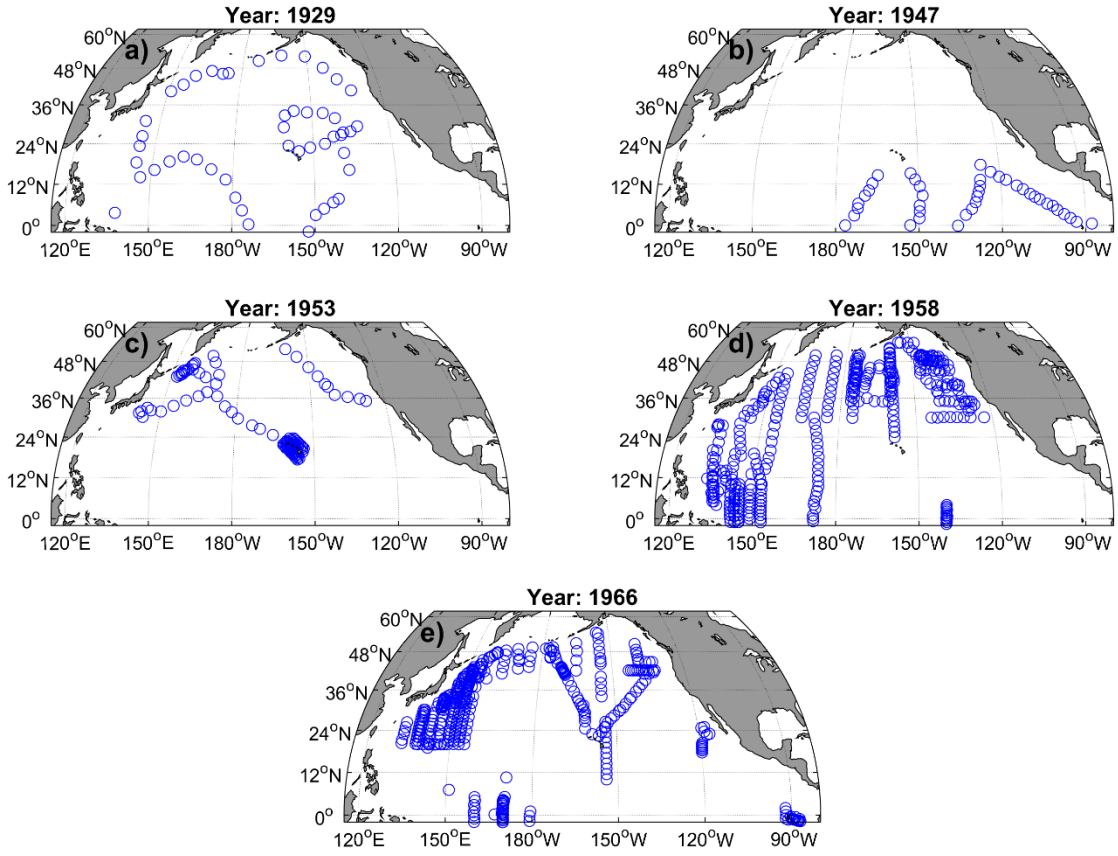


Figure S23. Hydrographic and nutrient sampling locations for selected years prior to 1970 (collected from Ocean Station Data in the World Ocean Database): (a) 1929, (b) 1947, (c) 1953, (d) 1958, and (e) 1966.

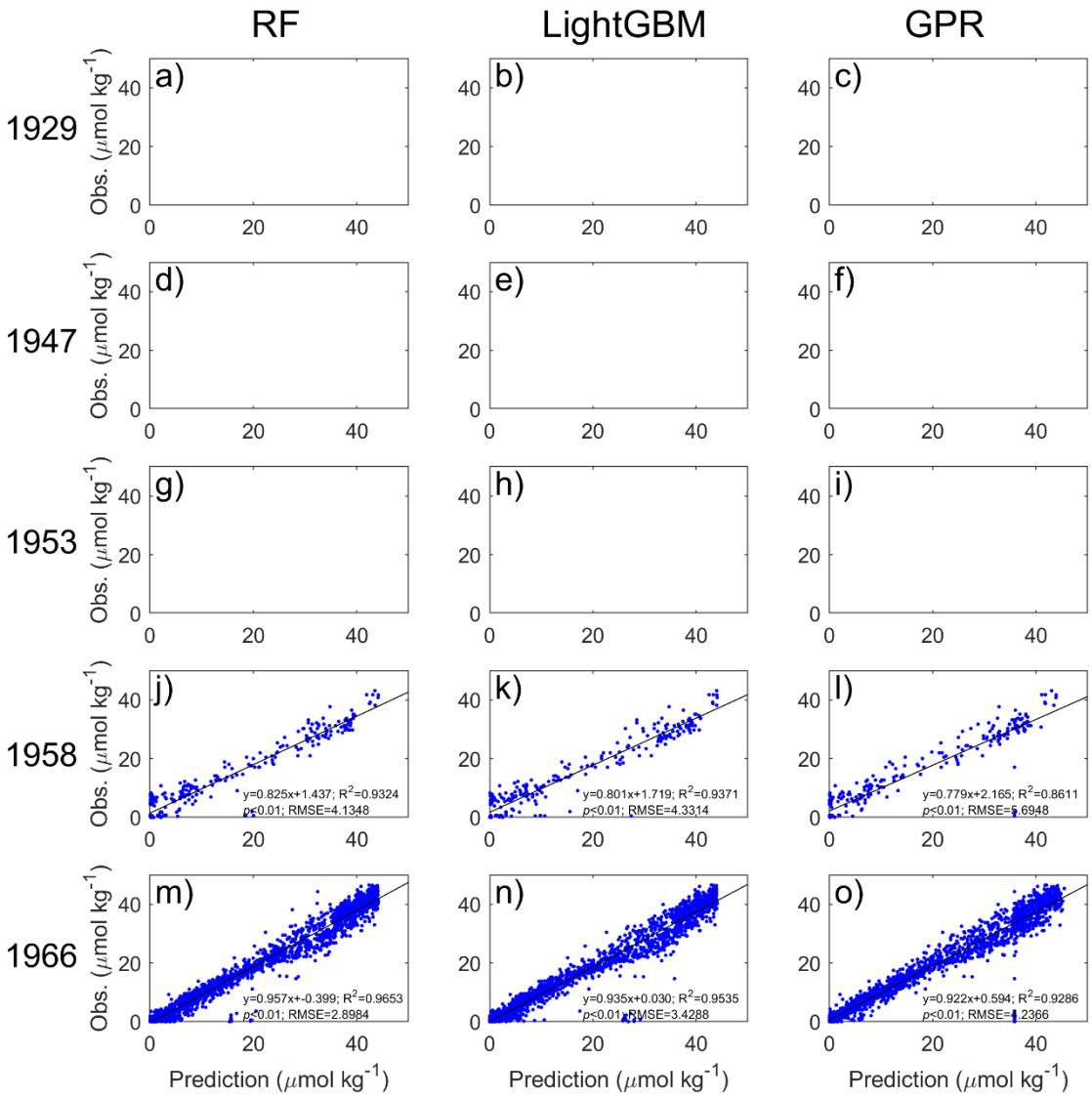


Figure S24. Validation of reconstructed NO_3^- concentrations from hydrography-based reconstructions for 1929 (row 1), 1947 (row 2), 1953 (row 3), 1958 (row 4), and 1966 (row 5). Station locations are shown in Fig. S23. Columns correspond to different reconstruction methods: Random Forest (RF; a, d, g, j, m), LightGBM (b, e, h, k, n), and Gaussian Process Regression (GPR; c, f, i, l, o). Each panel displays the linear regression fit (black line), regression equation, coefficient of determination (R^2), p-value, and root mean square error (RMSE). Note: no NO_3^- observations were available for 1929, 1947, or 1953.

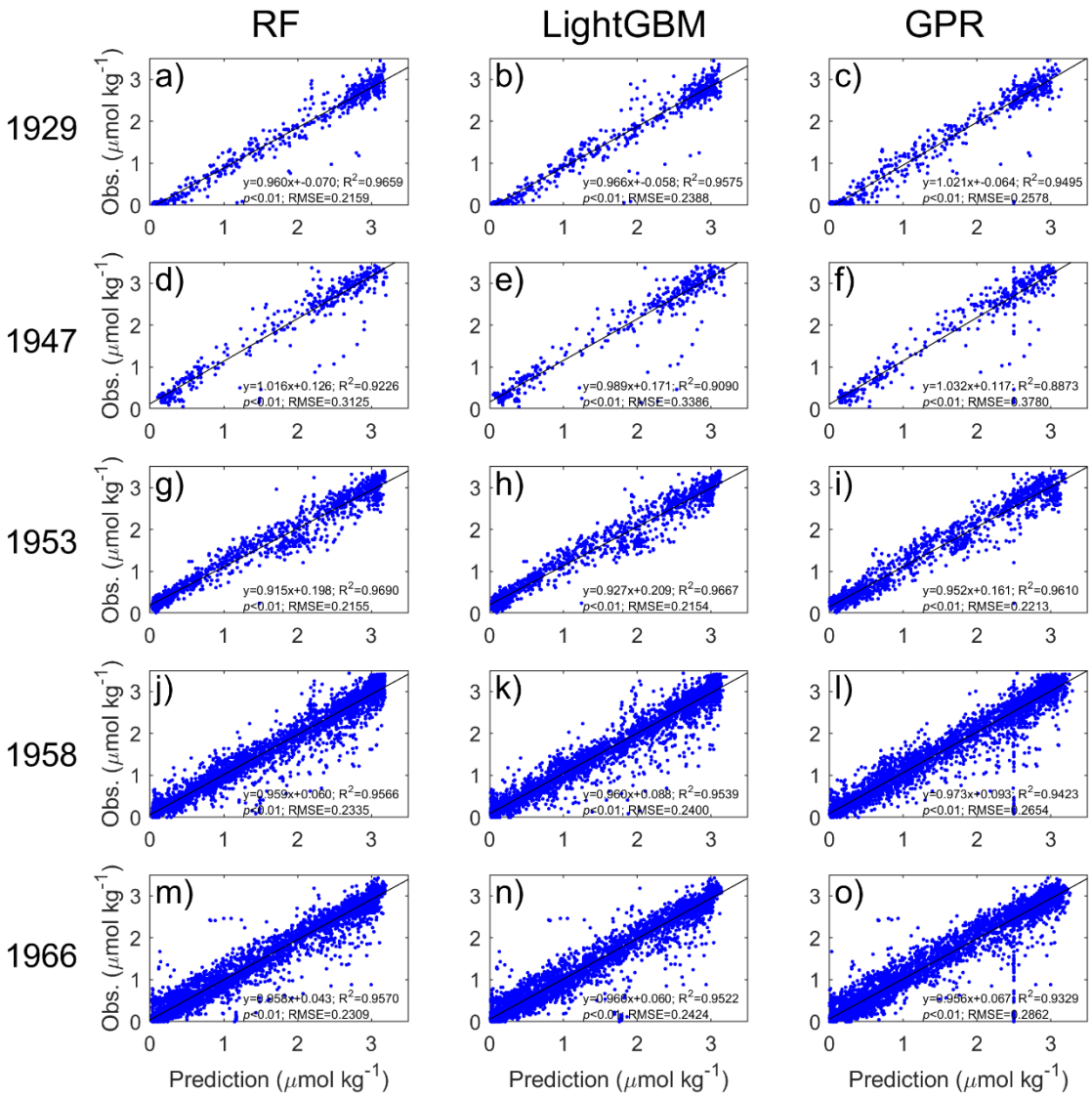


Figure S25. Same as Fig. S24, but for DIP.

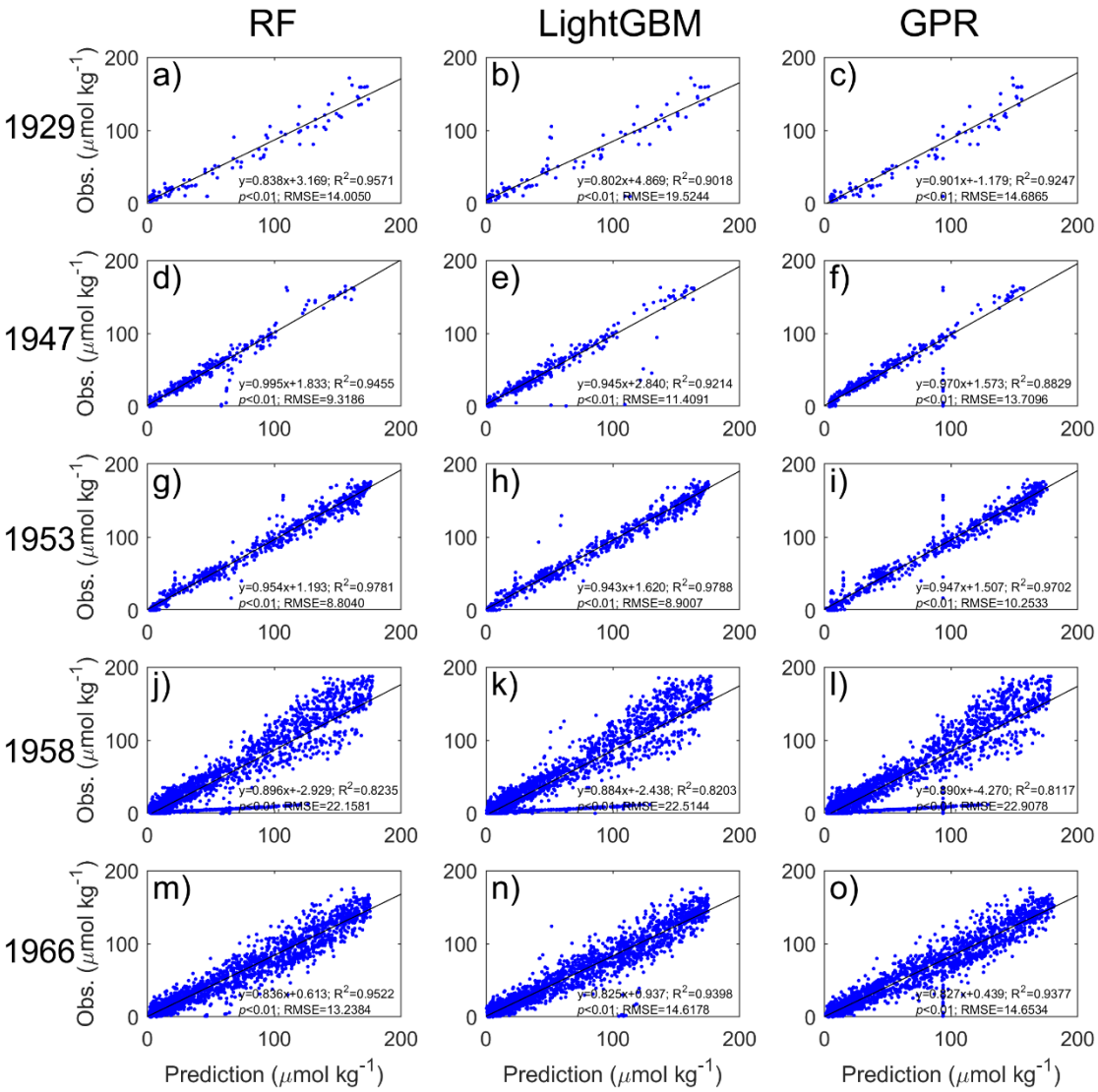


Figure S26. Same as Fig. S24, but for $Si(OH)_4$.

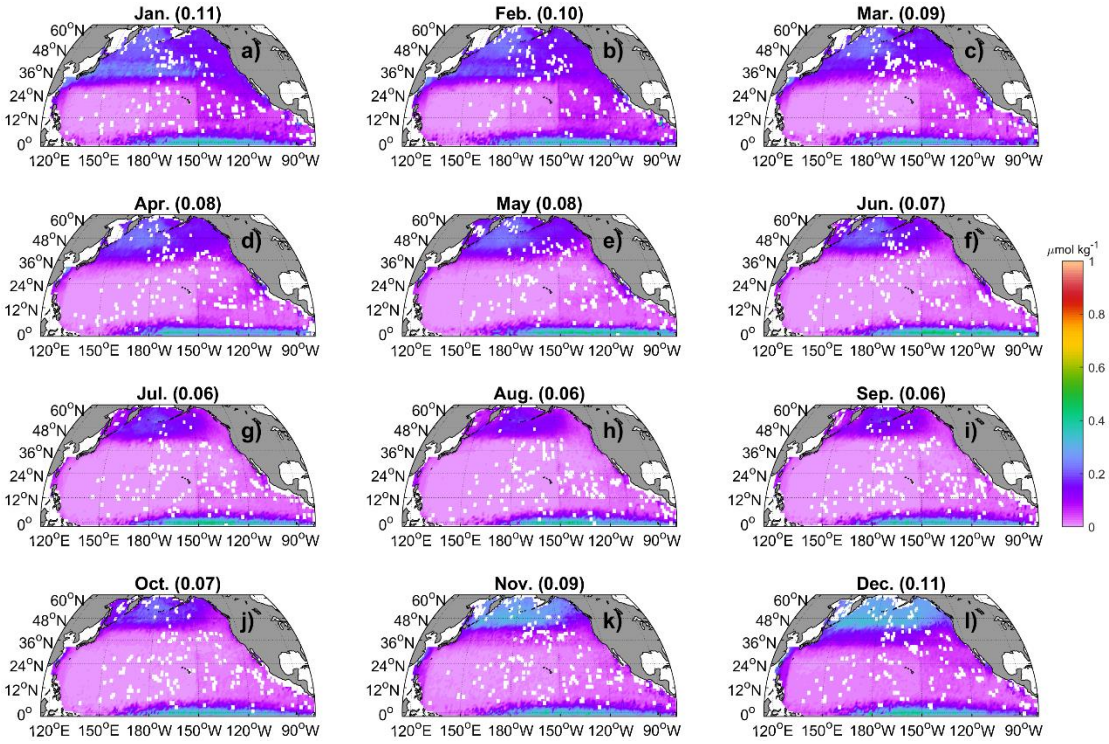


Figure S27. The monthly climatology of NO_2^- at 5 m in the North Pacific. Data are binned and averaged within 1×1º grid cell. The values in the title represent the spatial mean values.

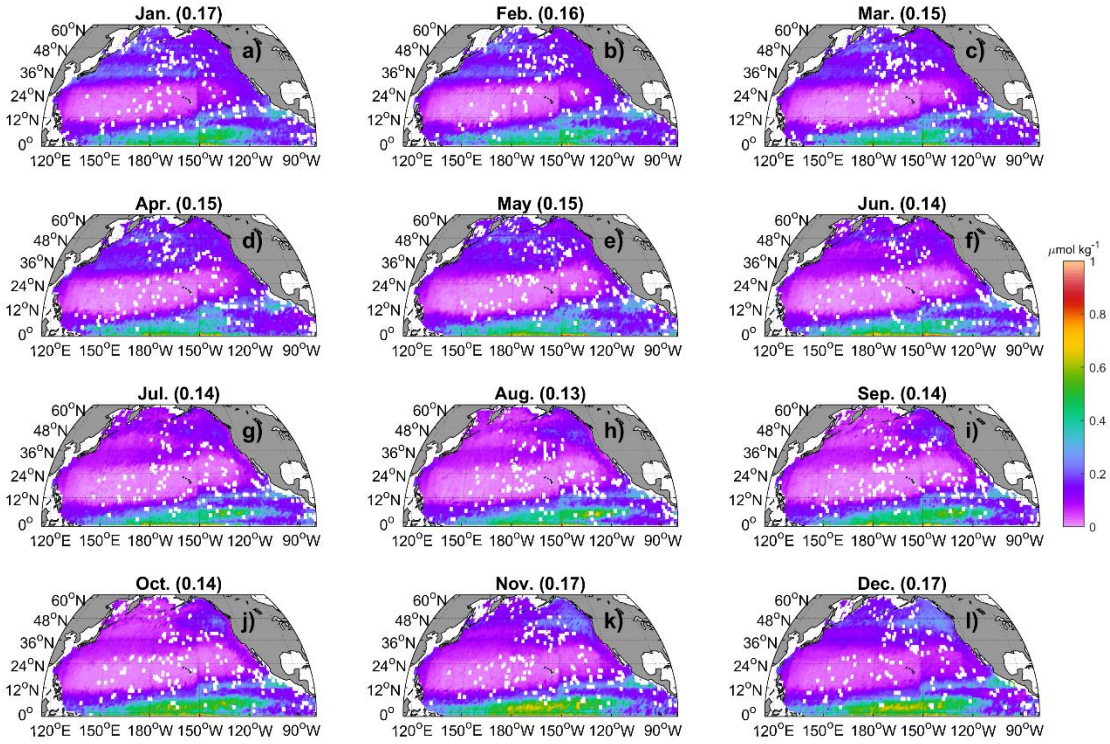


Figure S28. Similar to Fig. S27, but for a depth of 100 m.

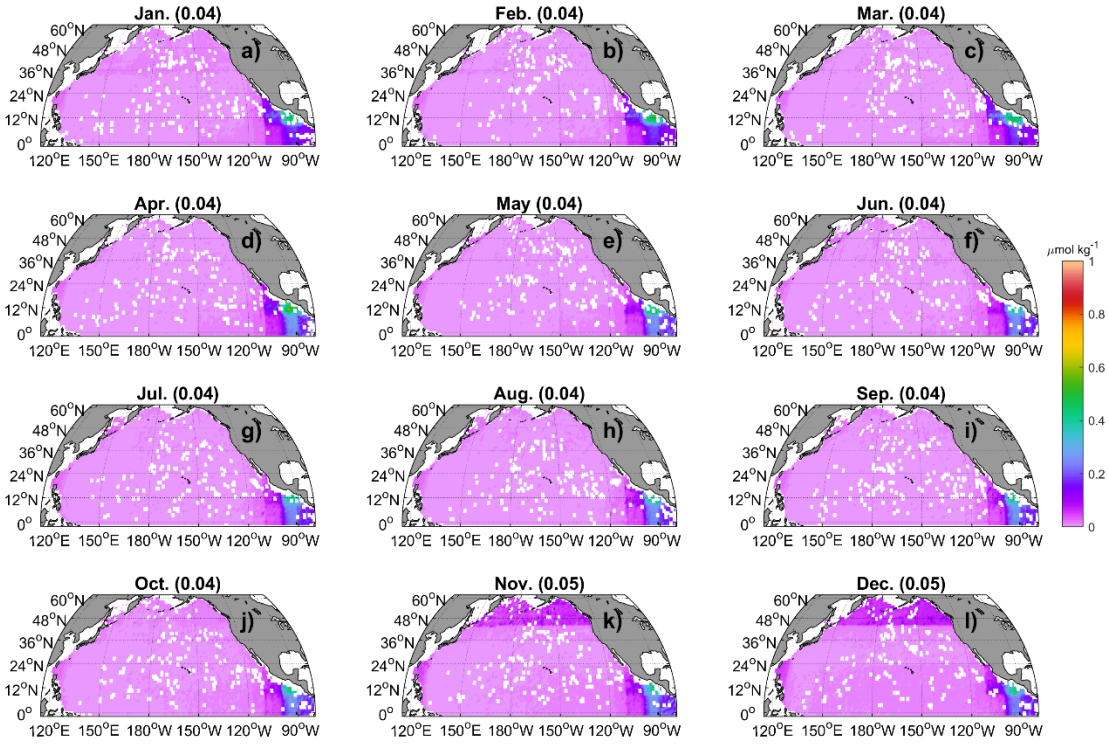


Figure S29. Similar to Fig. S27, but for a depth of 500 m.

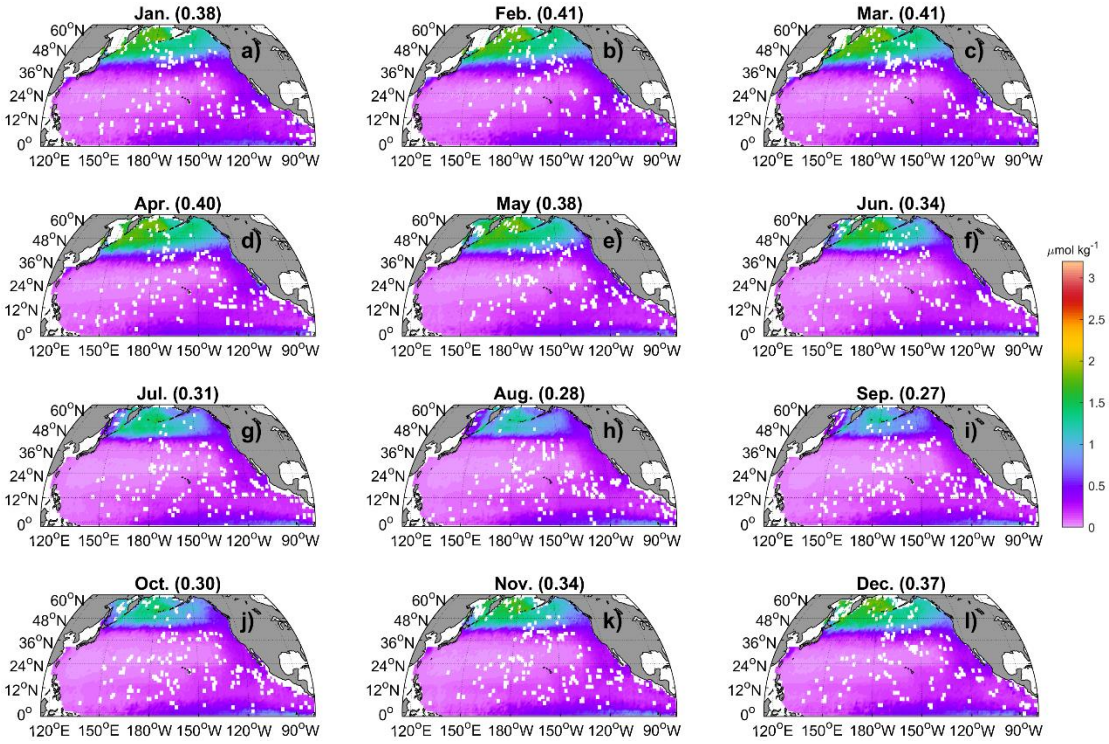


Figure S30. Similar to Fig. S27, but for DIP and at a depth of 5 m.

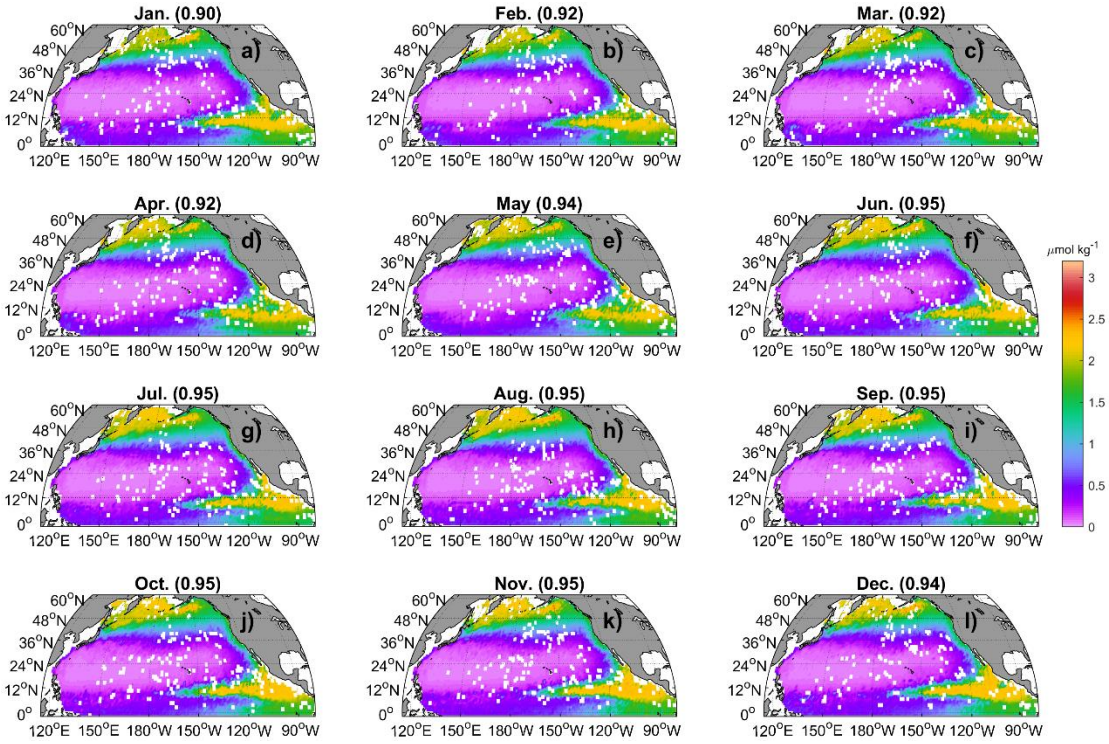


Figure S31. Similar to Fig. S27, but for DIP and at a depth of 100 m.

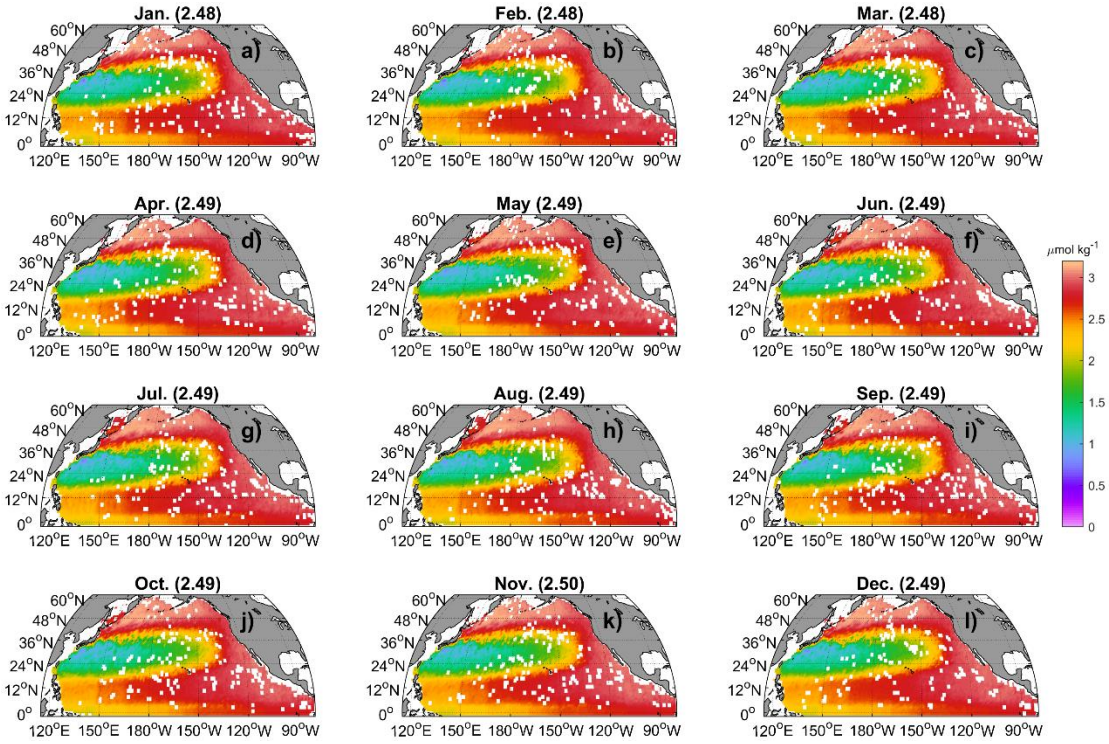


Figure S32. Similar to Fig. S27, but for DIP and at a depth of 500 m.

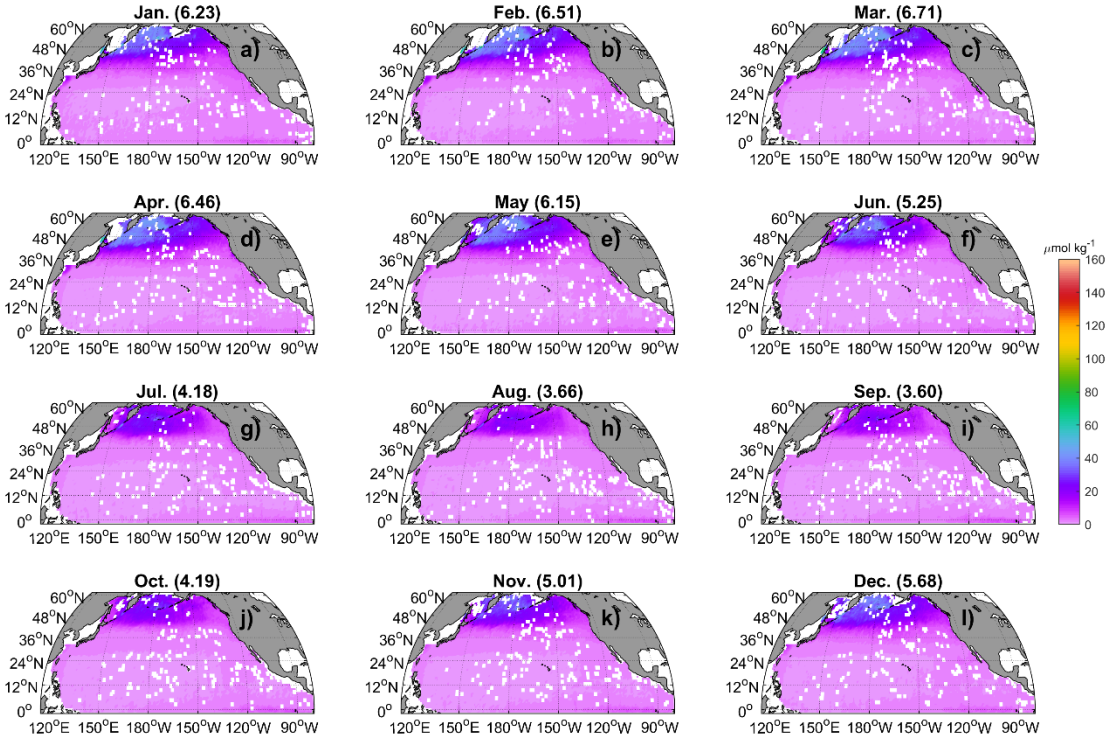


Figure S33. Similar to Fig. S27, but for $Si(OH)_4$ and at a depth of 5 m.

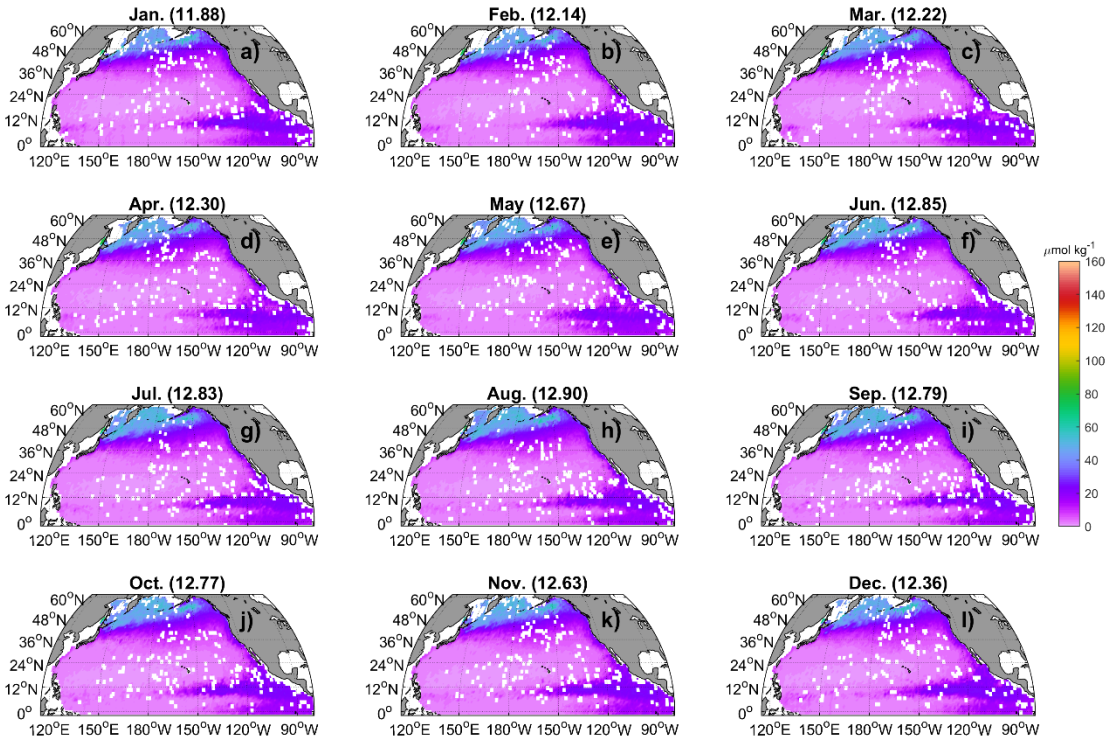


Figure S34. Similar to Fig. S27, but for $Si(OH)_4$ and at a depth of 100 m.

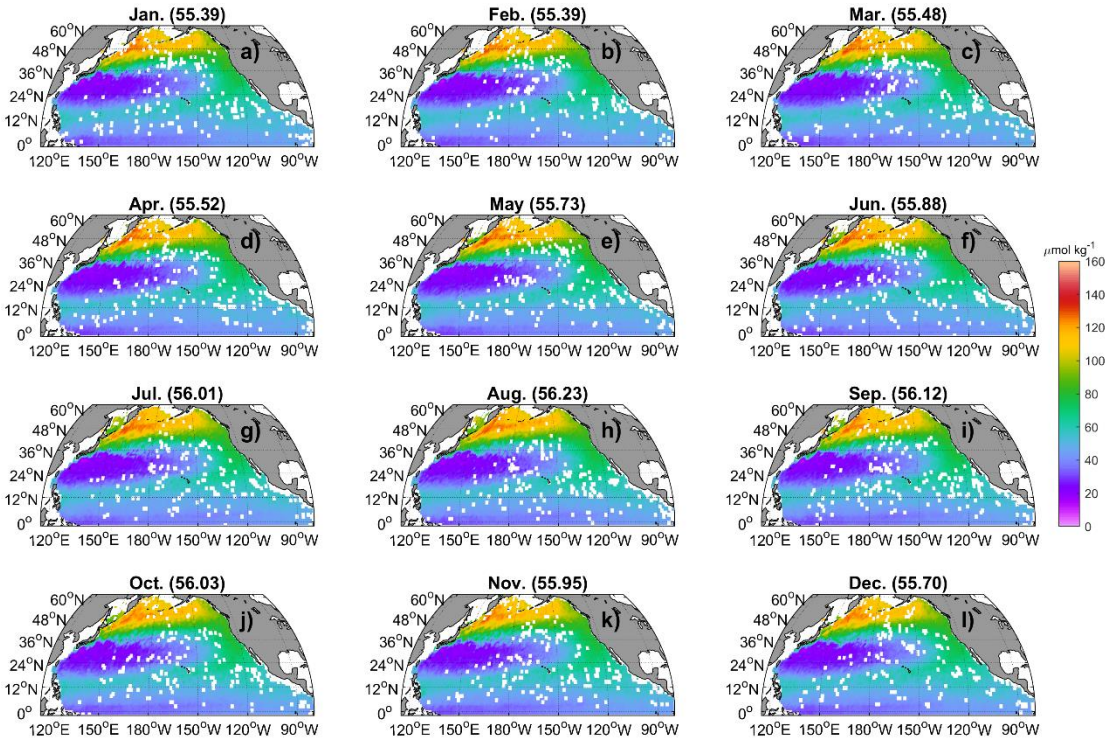


Figure S35. Similar to Fig. S27, but for $Si(OH)_4$ and at a depth of 500 m.

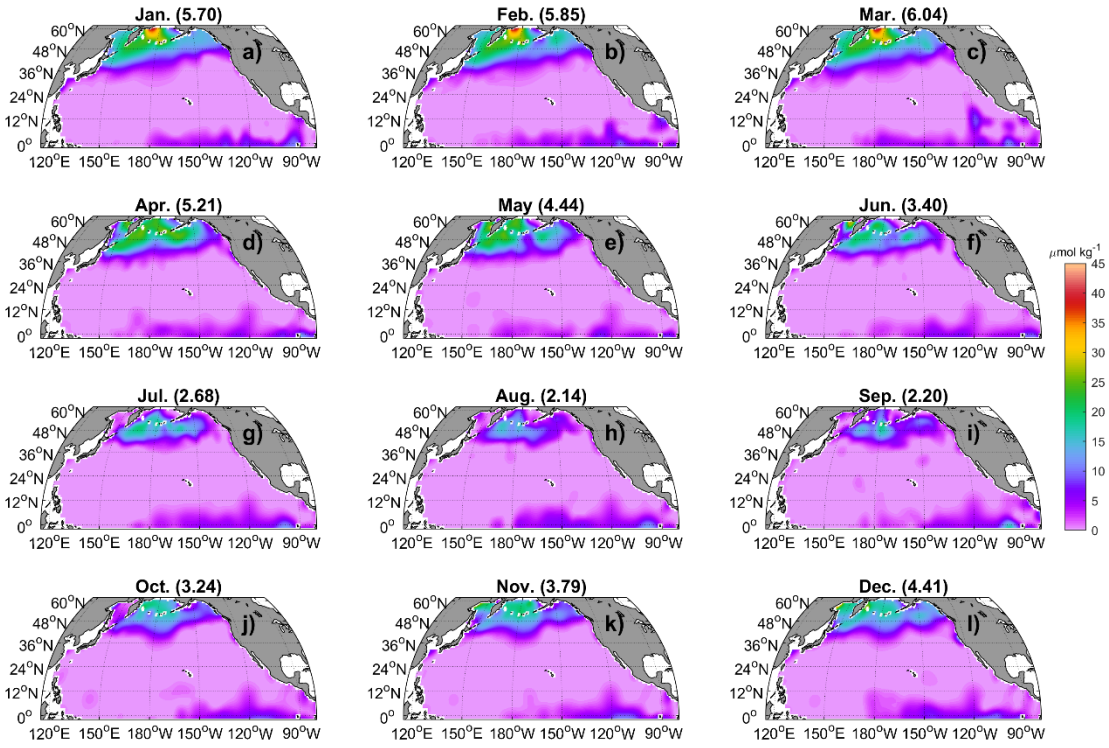


Figure S36. The WOA23 monthly climatology of NO_3^- at 5 m in the North Pacific. The values in the title represent the spatial mean values.

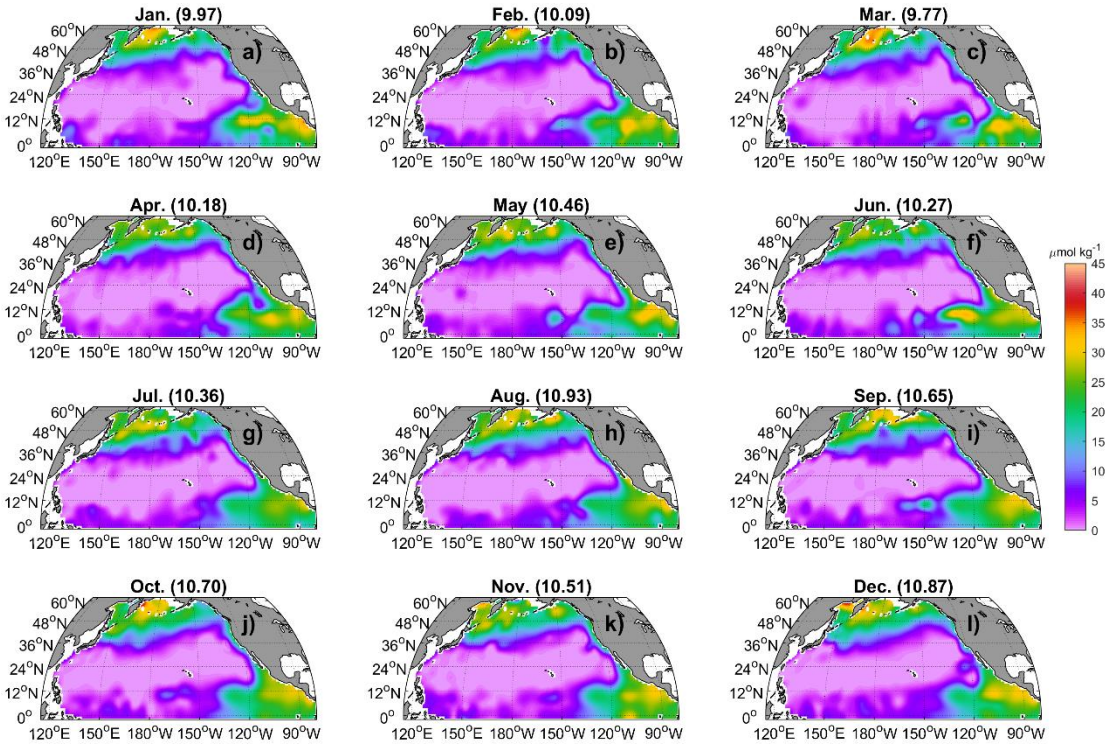


Figure S37. Similar to Fig. S36, but for a depth of 100 m.

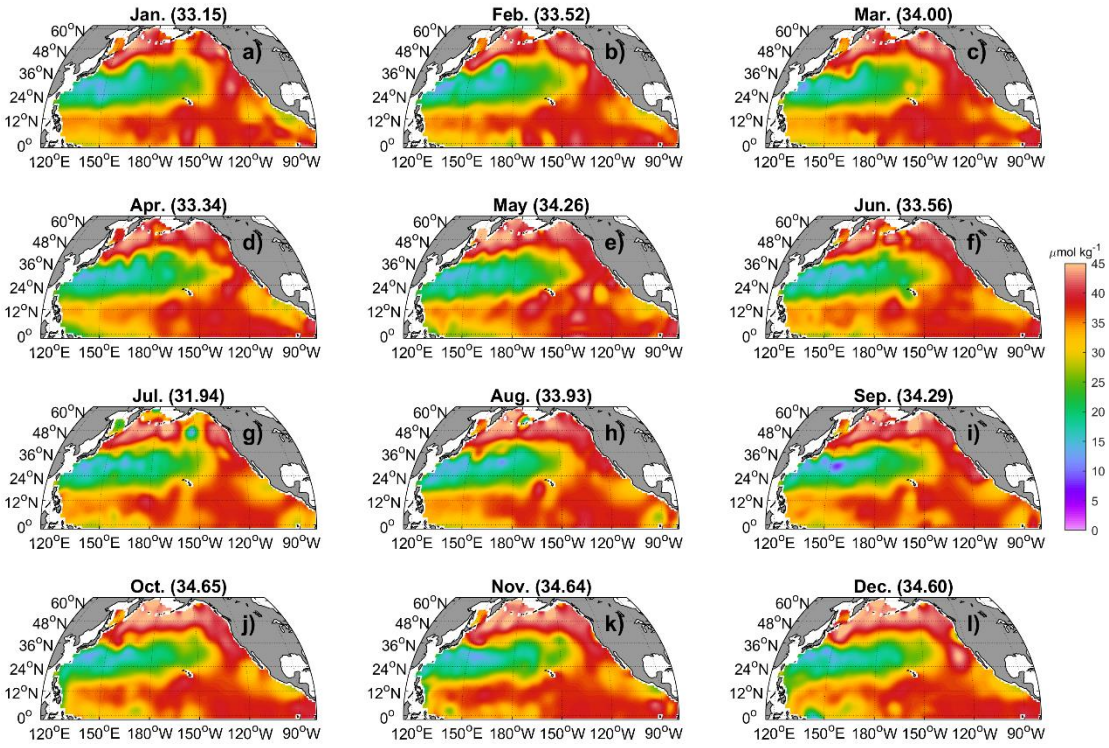


Figure S38. Similar to Fig. S36, but for a depth of 500 m.

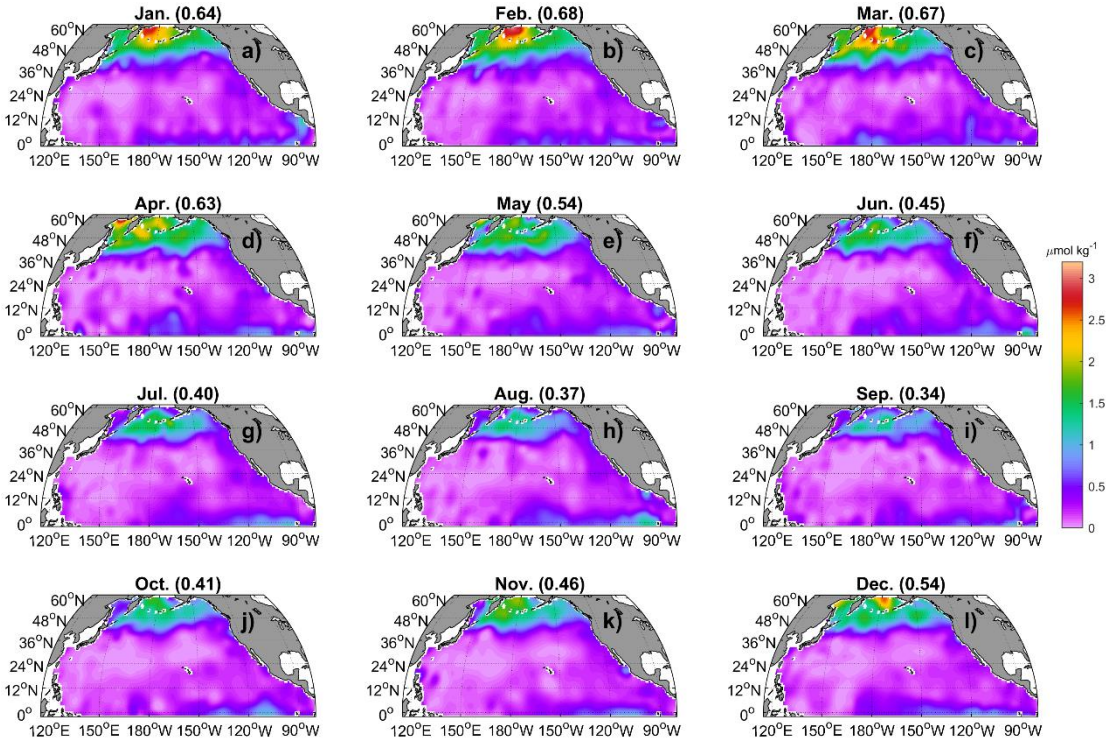


Figure S39. Similar to Fig. S36, but for DIP and at a depth of 5 m.

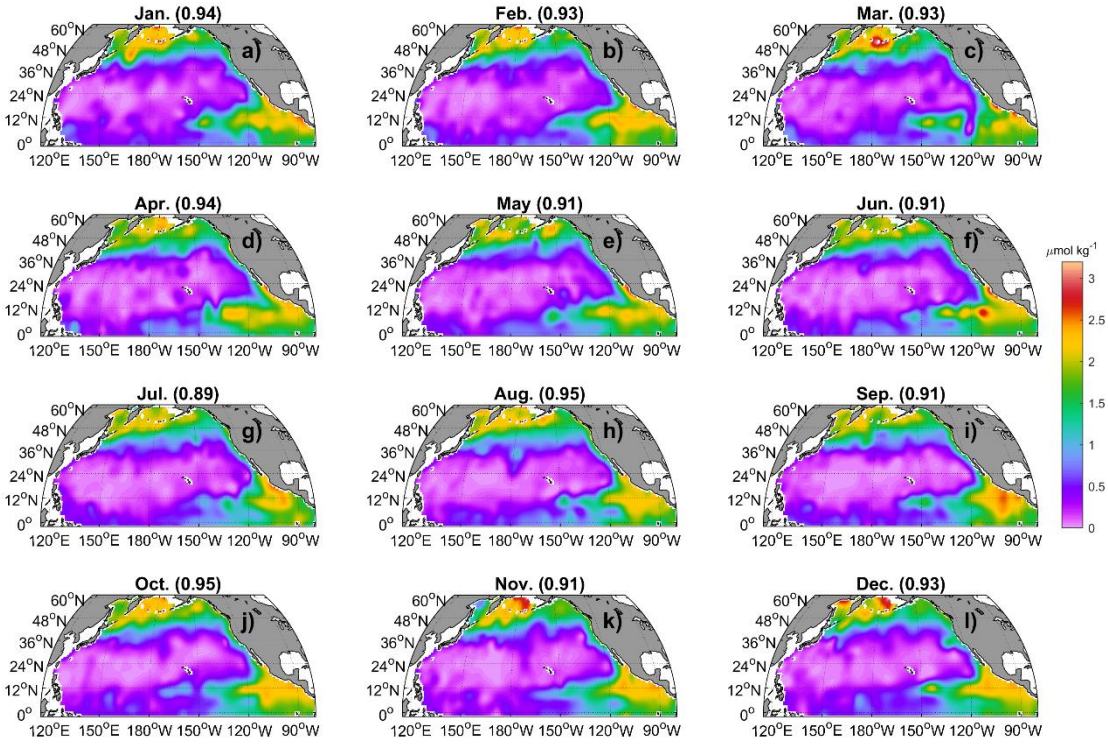


Figure S40. Similar to Fig. S36, but for DIP and at a depth of 100 m.

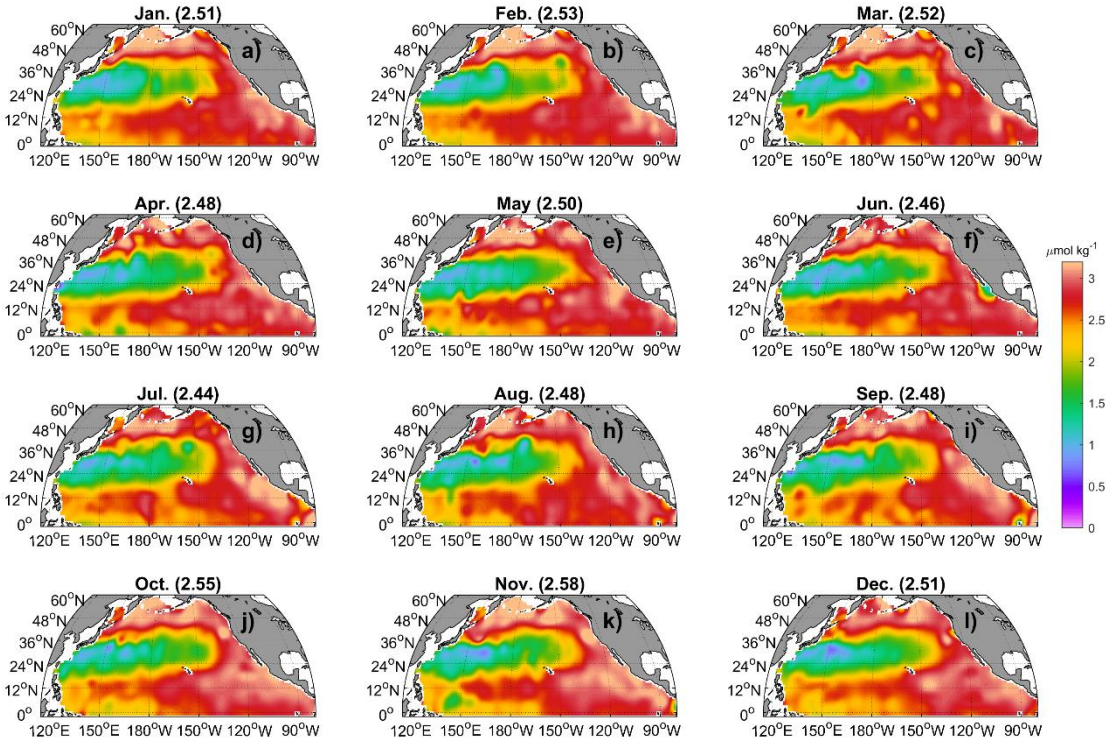


Figure S41. Similar to Fig. S36, but for DIP and at a depth of 500 m.

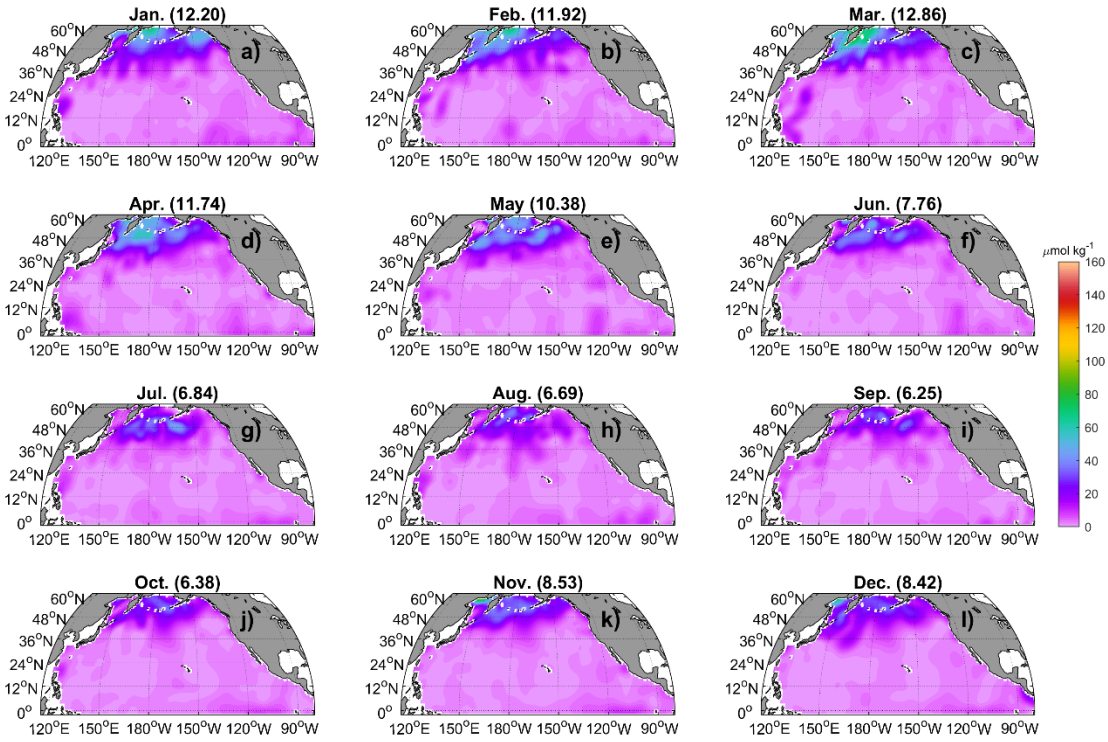


Figure S42. Similar to Fig. S36, but for $Si(OH)_4$ and at a depth of 5 m.

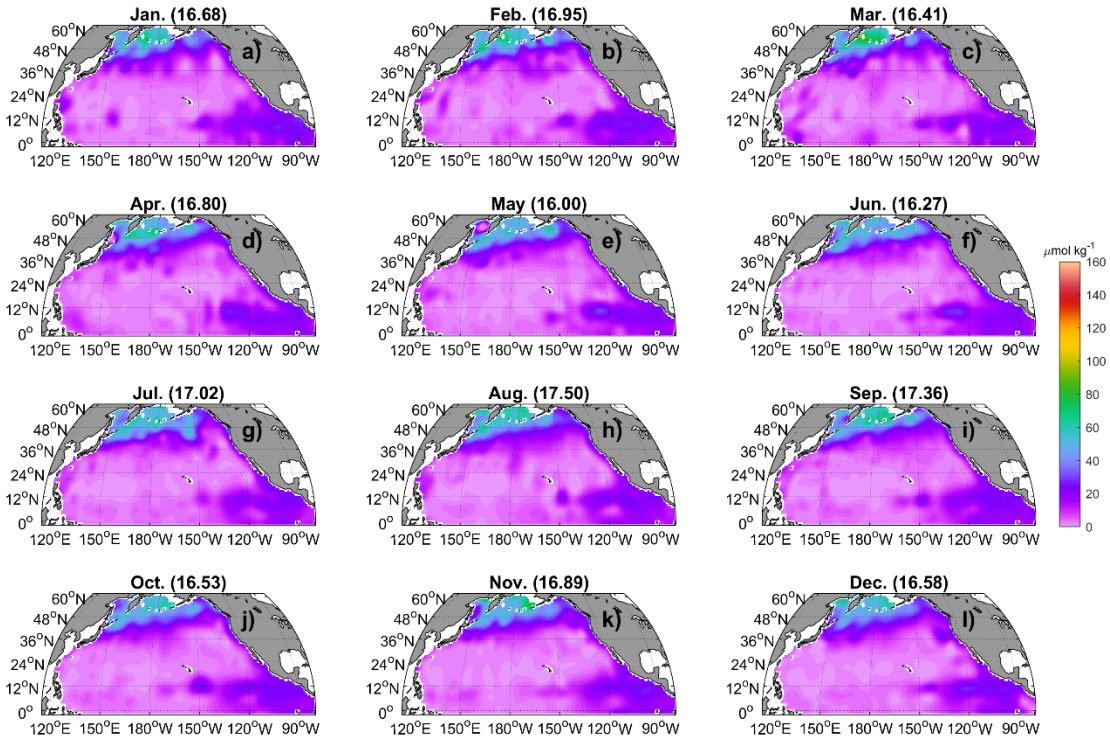


Figure S43. Similar to Fig. S36, but for $Si(OH)_4$ and at a depth of 100 m.

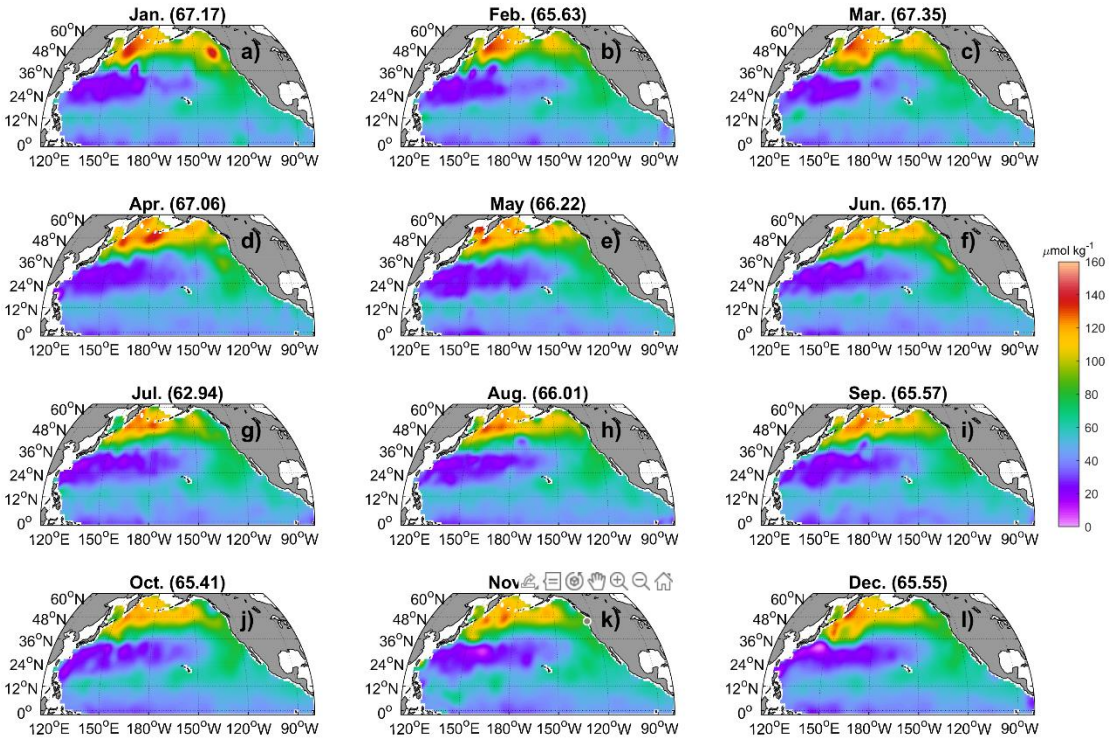


Figure S44. Similar to Fig. S36, but for $Si(OH)_4$ and at a depth of 500 m.

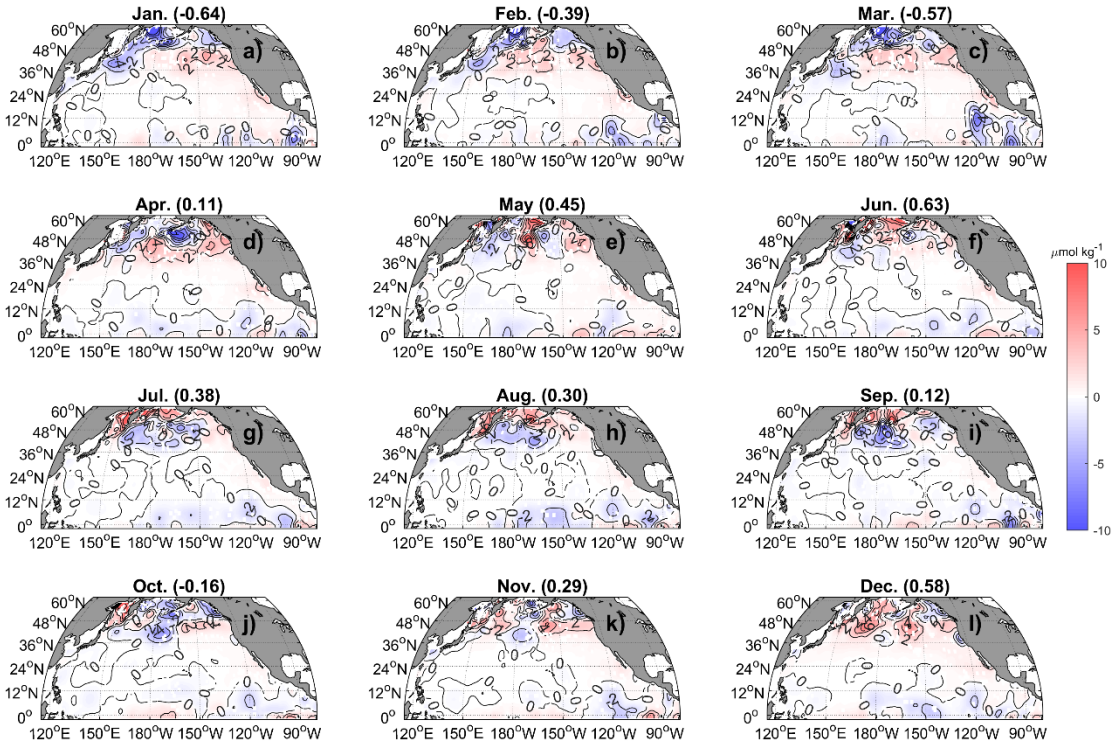


Figure S45. Difference between the predicted (Random Forest, RF) and World Ocean Atlas (WOA) climatologies of NO_x^- (NO_3^- + NO_2^-) at 5 m depth (RF minus WOA).

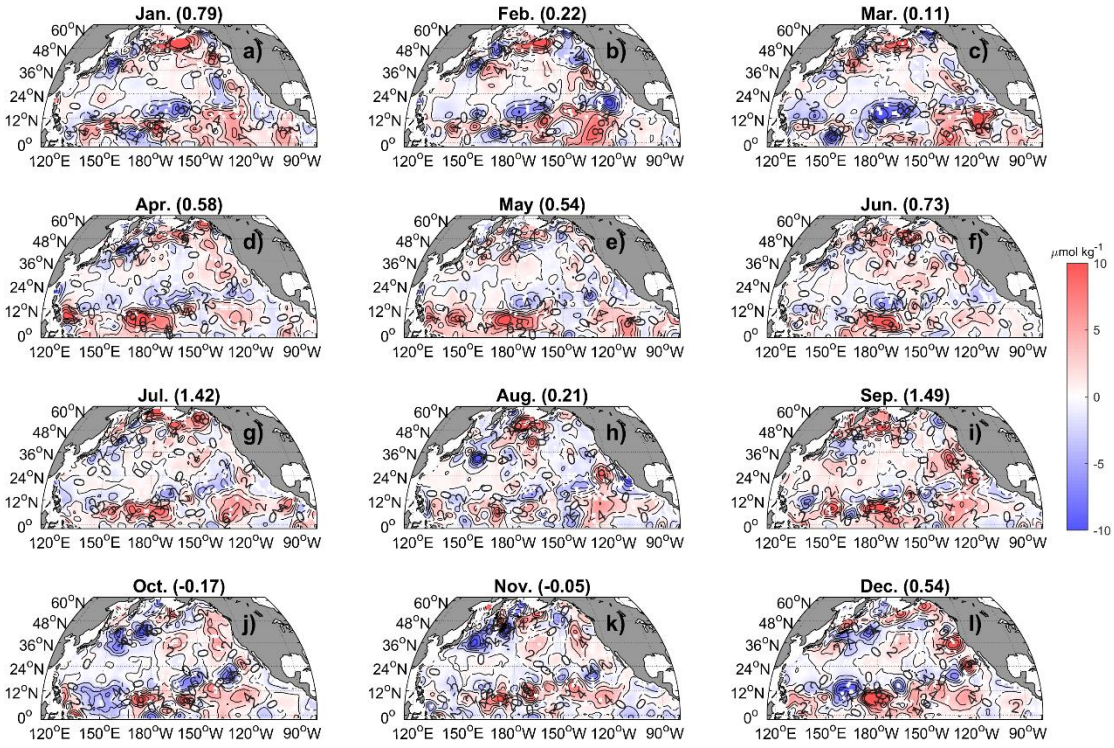


Figure S46. Same as Fig. S27, but for 100 m.

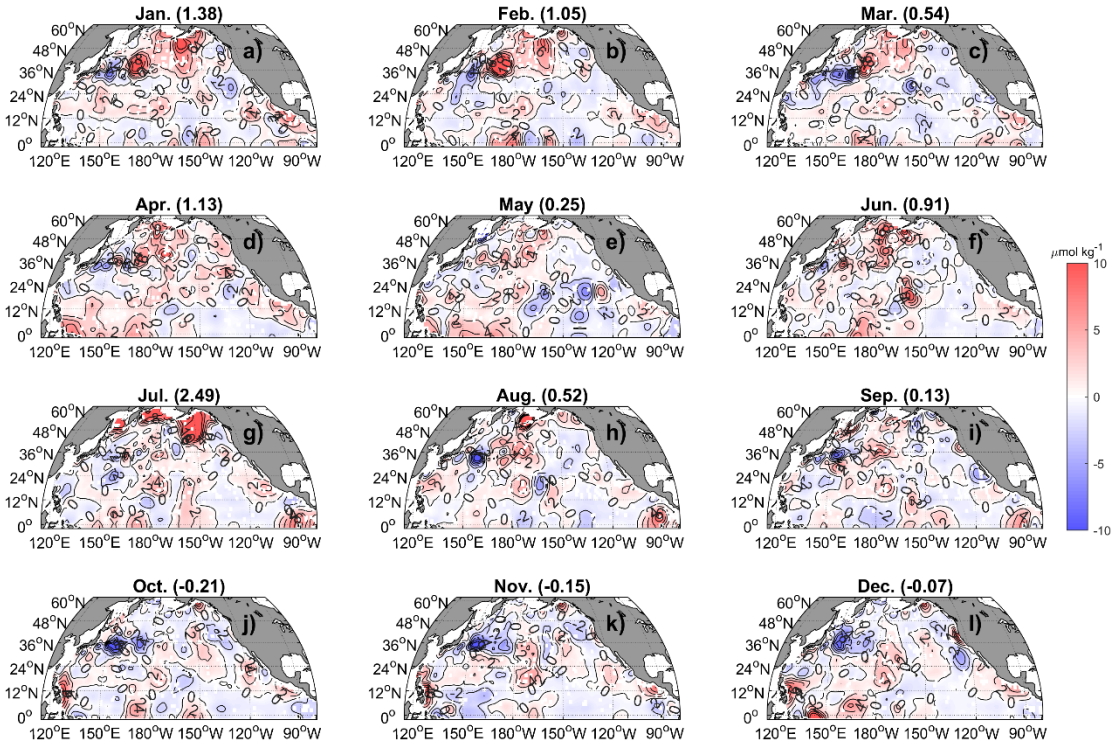
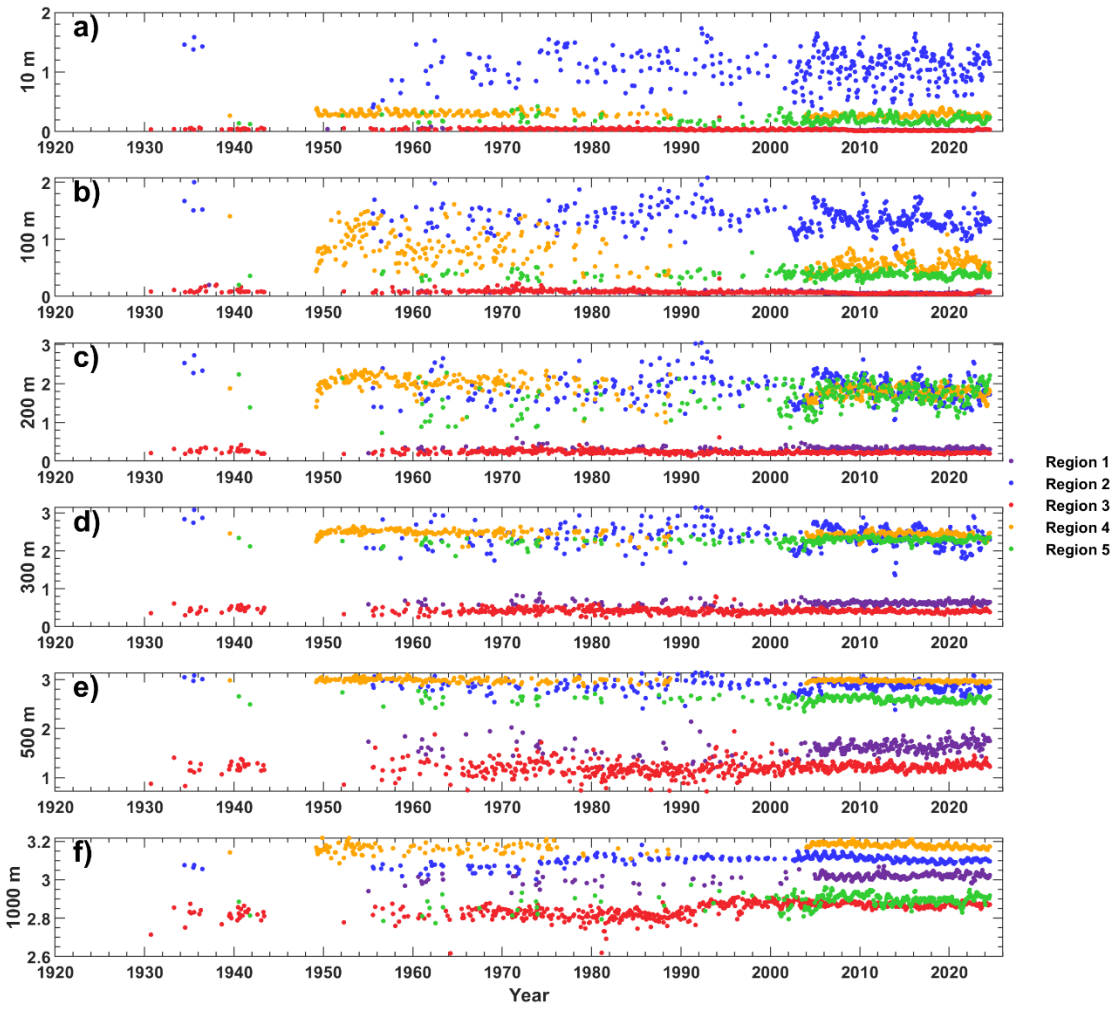


Figure S47. Same as Fig. S27, but for 500 m.



241

242 **Figure S48.** Time series of reconstructed DIP concentrations at 10 m (a), 100 m (b),
243 200 m (c), 300 m (d), 500 m (e), and 1000 m (f) for regions 1 - 5 (see Fig. 16). Data
244 were first binned by depth and region, then averaged by month.

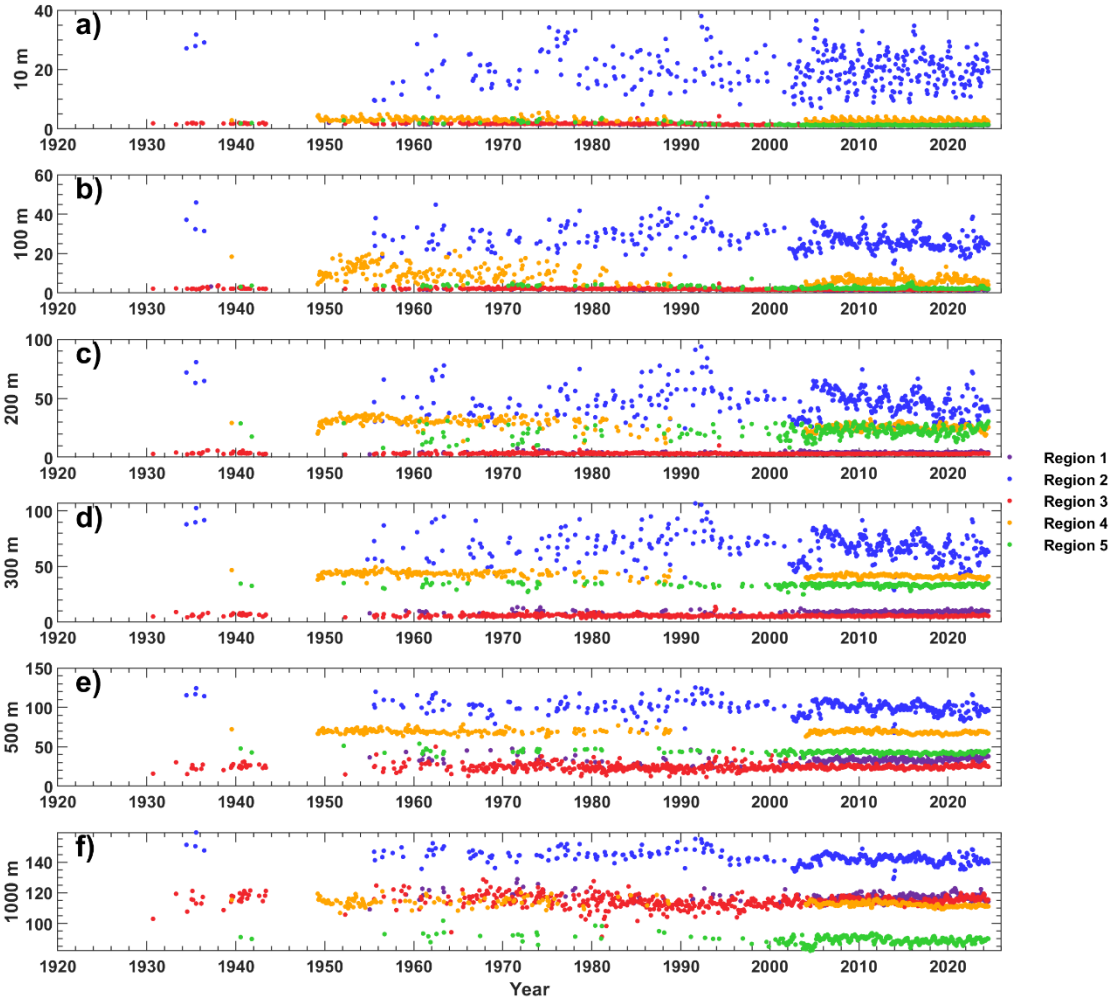


Figure S49. Same as Figure S48, but for $Si(OH)_4$.

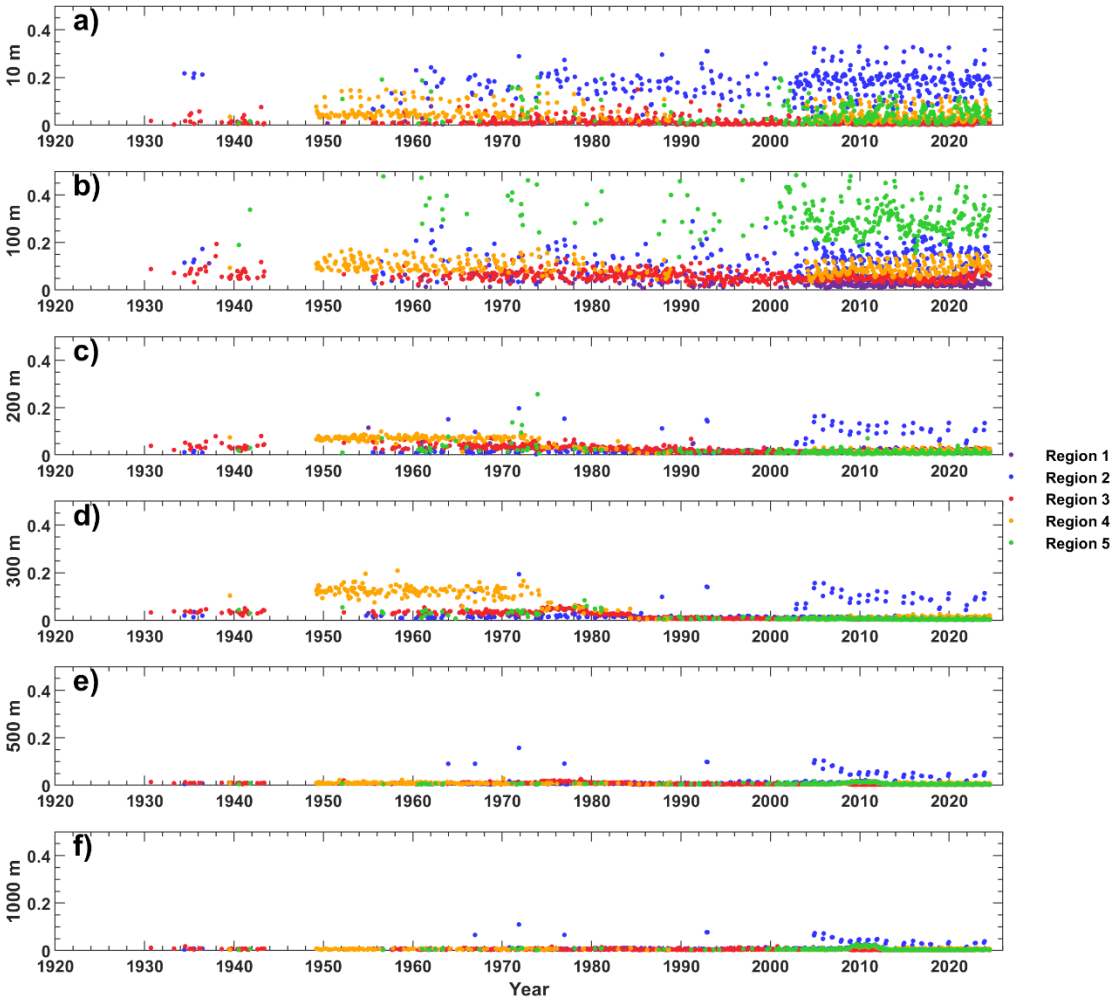


Figure S50. Same as Figure S48, but for NO_2^-.