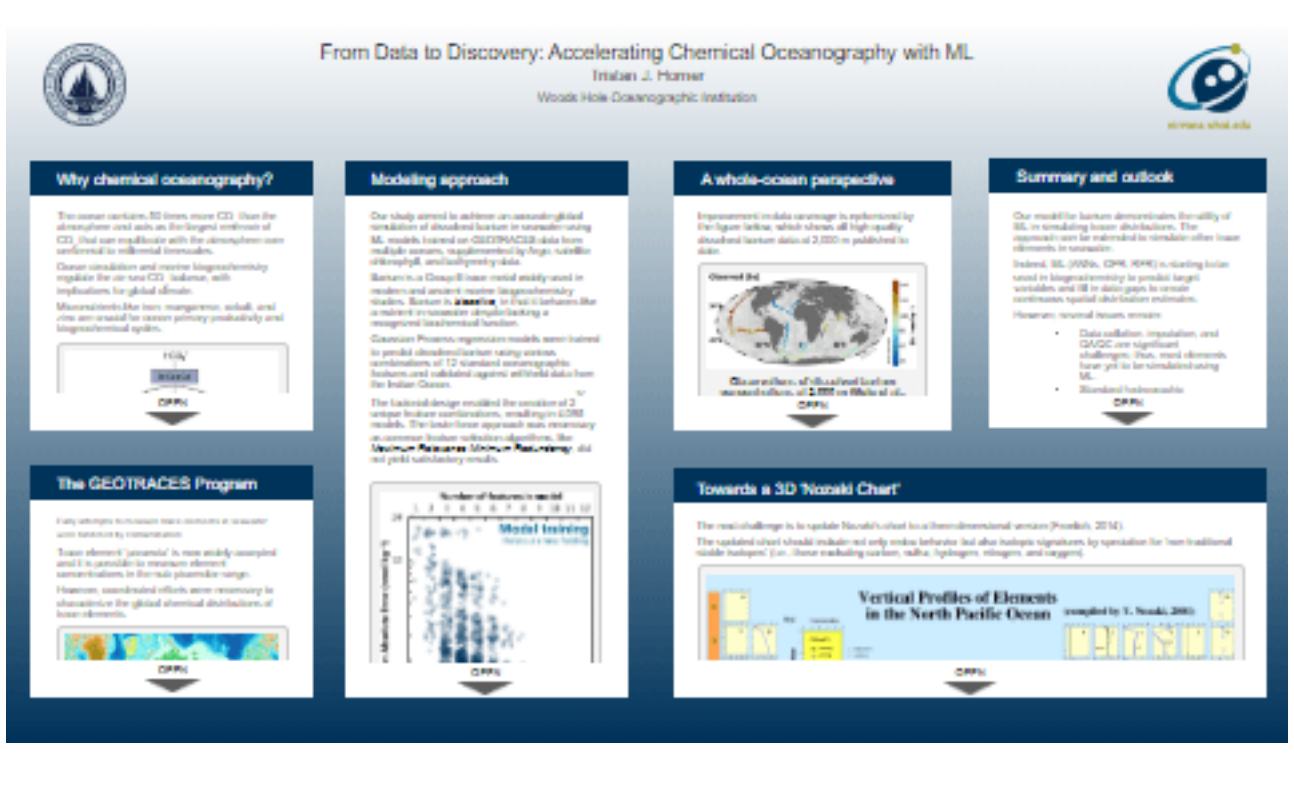
From Data to Discovery: Accelerating Chemical Oceanography with ML



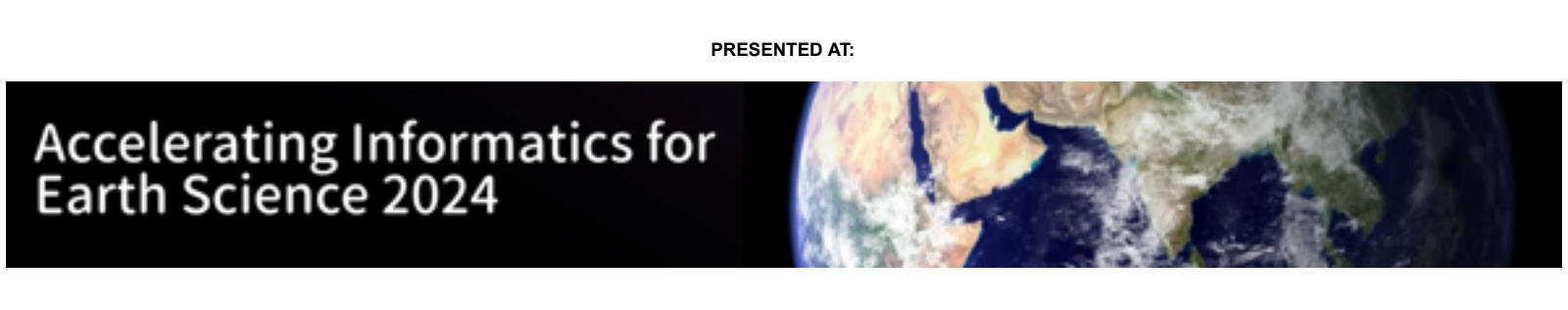
From Data to Discovery: Accelerating Chemical Oceanography with ML
Tristan J. Horner
Woods Hole Oceanographic Institution
Why chemical oceanography?
Modeling approach
A whole-ocean perspective
Summary and outlook
The GEOTRACES Program
Towards a 3D 'Nozaki Chart'

Tristan J. Horner

Woods Hole Oceanographic Institution

nirvana.whoi.edu

PRESENTED AT:

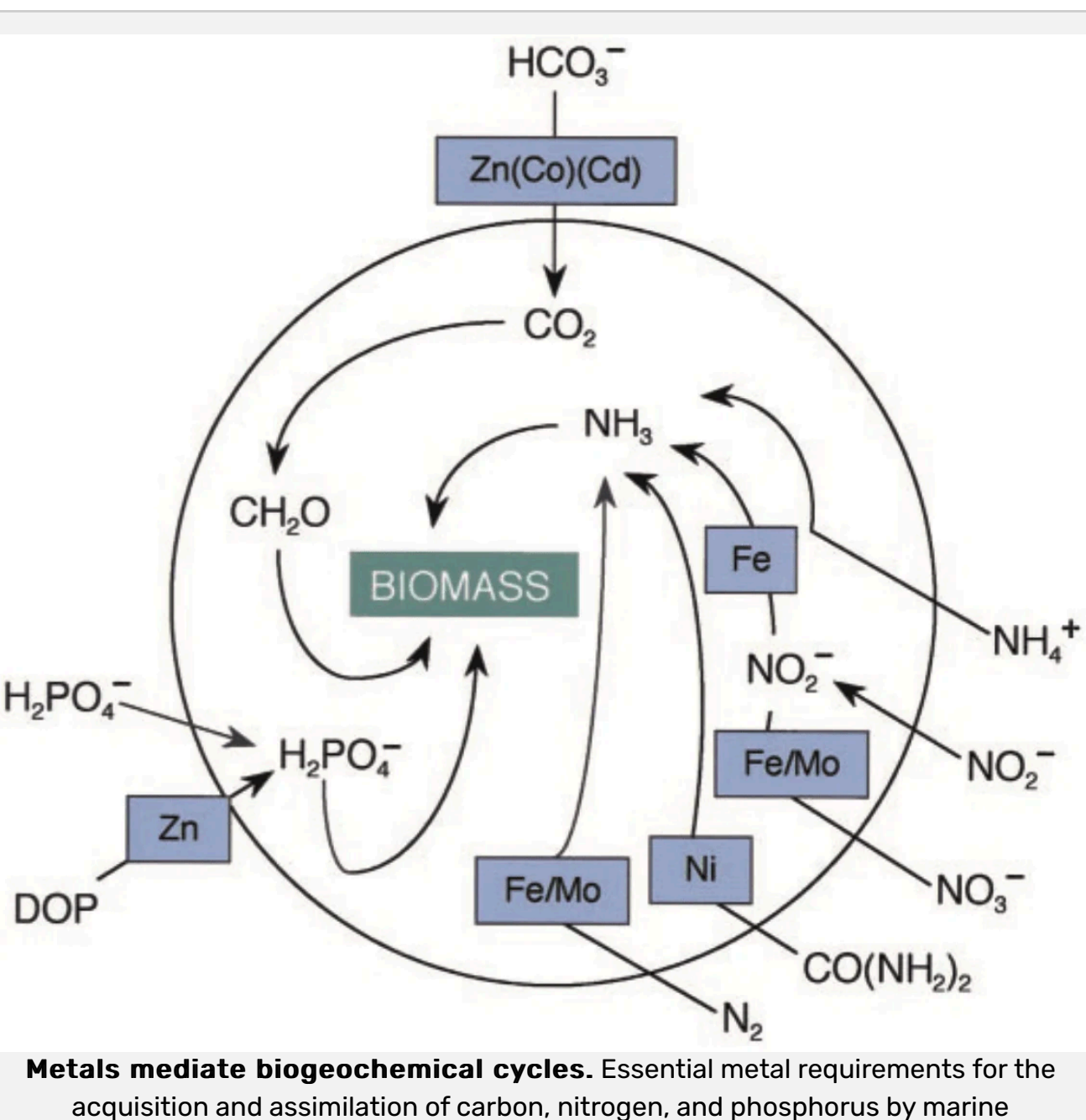


WHY CHEMICAL OCEANOGRAPHY?

The ocean contains 50 times more CO_2 than the atmosphere and acts as the largest reservoir of CO_2 that can equilibrate with the atmosphere over centennial to millennial timescales.

Ocean circulation and marine biogeochemistry regulate the air-sea CO_2 balance, with implications for global climate.

Micronutrients like iron, manganese, cobalt, and zinc are crucial for ocean primary productivity and biogeochemical cycles.



Metals mediate biogeochemical cycles. Essential metal requirements for the acquisition and assimilation of carbon, nitrogen, and phosphorus by marine phytoplankton (Morel & Price, 2003)

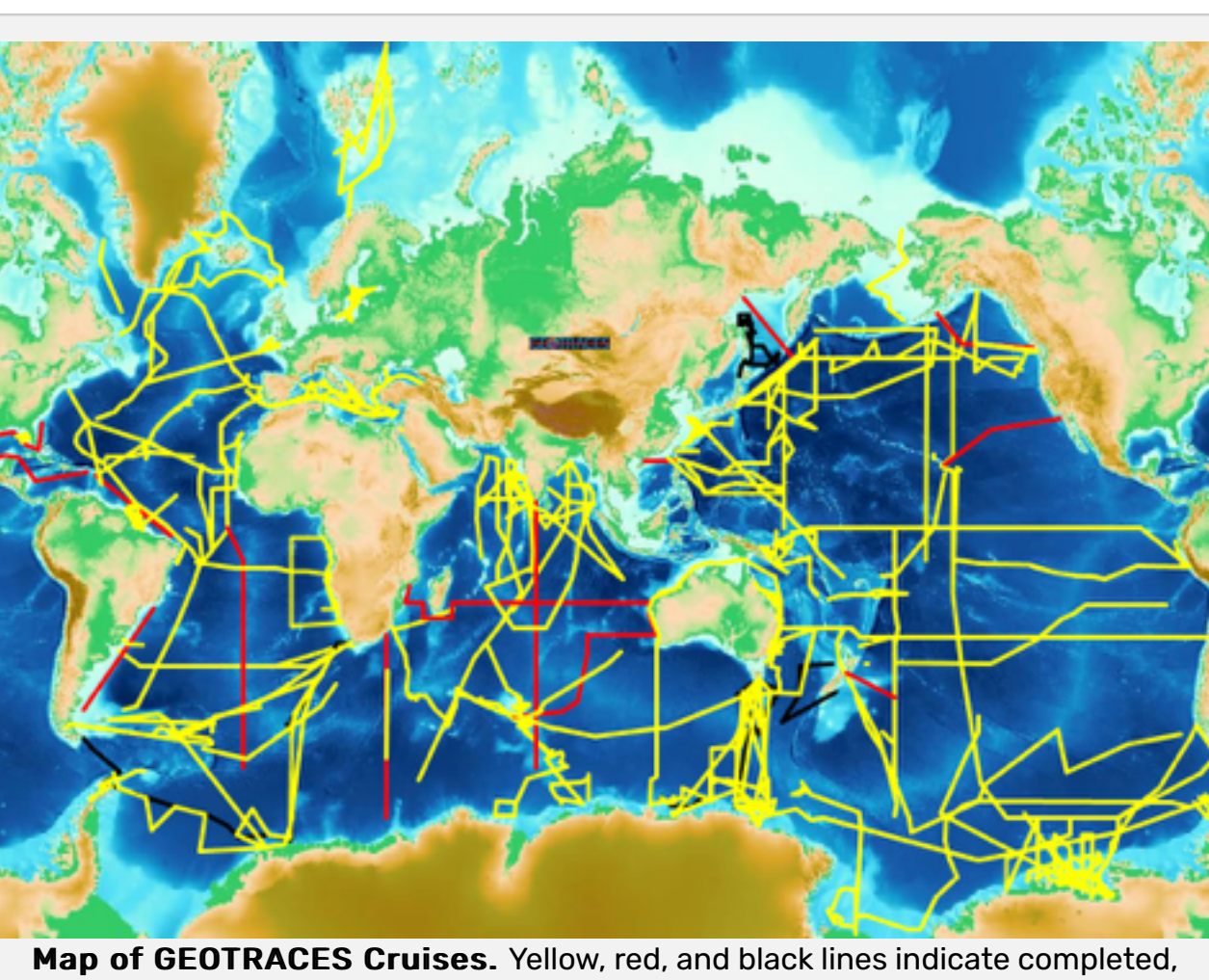
Chemical oceanographers study the marine distributions of these and other metals, analyzing their spatial, vertical, and temporal variations and the reasons behind these variations.

THE GEOTRACES PROGRAM

Early attempts to measure trace elements in seawater were hindered by contamination.

Trace element 'paranoia' is now widely accepted and it is possible to measure element concentrations in the sub-picomolar range.

However, coordinated efforts were necessary to characterize the global chemical distributions of trace elements.



Map of GEOTRACES Cruises. Yellow, red, and black lines indicate completed, planned, and IPY Sections, respectively (BODC, 2024).

This led to the creation of the GEOTRACES program, an *international study of the marine biogeochemistry of trace elements and their isotopes* that aims to:

1. Determine global distributions of trace elements and isotopes
2. Understand and model the processes controlling these distributions

The program has now sampled every major ocean basin through dozens of 'cruises.'

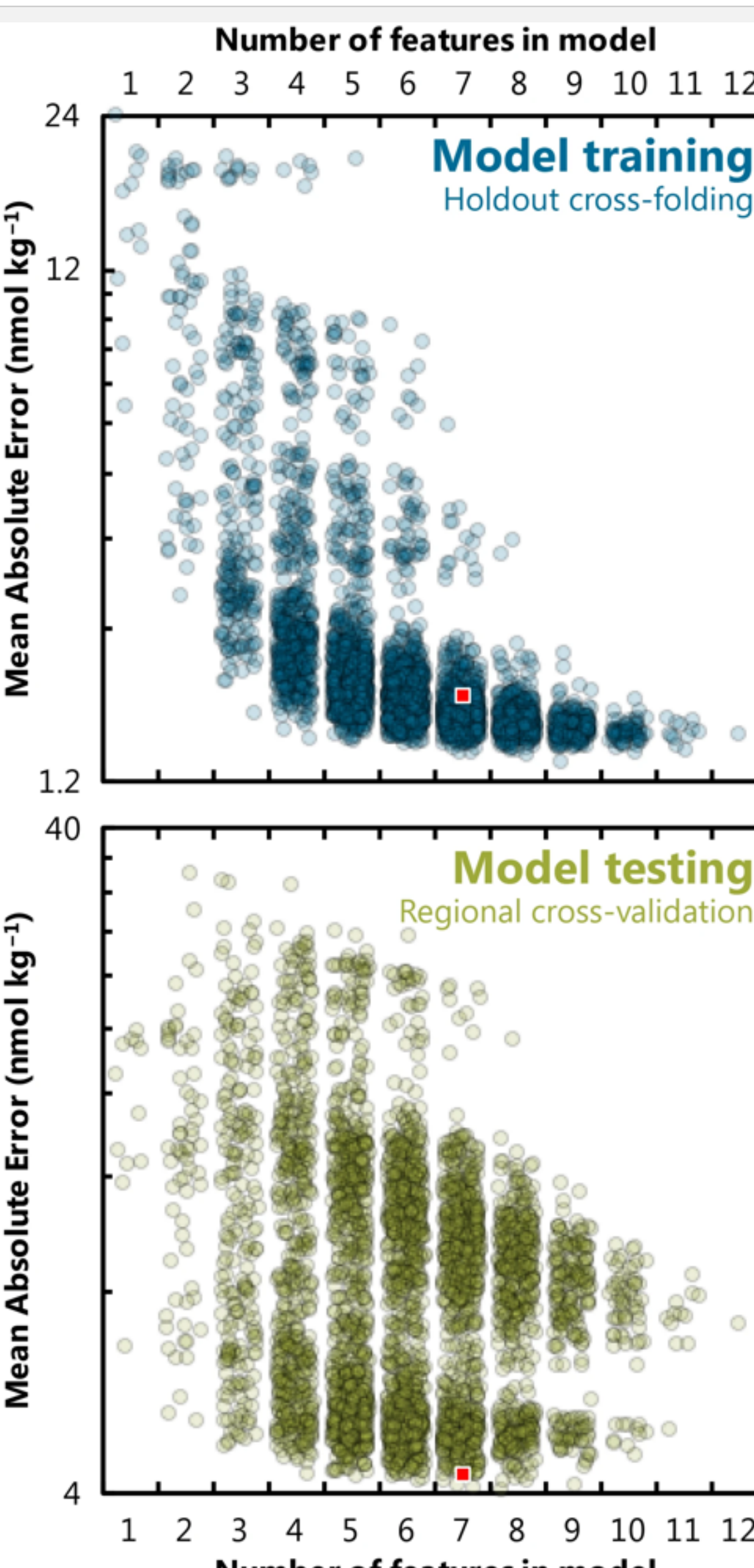
MODELING APPROACH

Our study aimed to achieve an accurate global simulation of dissolved barium in seawater using ML models trained on GEOTRACES data from multiple oceans, supplemented by Argo, satellite chlorophyll, and bathymetry data.

Barium is a Group II trace metal widely used in modern and ancient marine biogeochemistry studies. Barium is *bioactive,* in that it behaves like a nutrient in seawater despite lacking a recognized biochemical function.

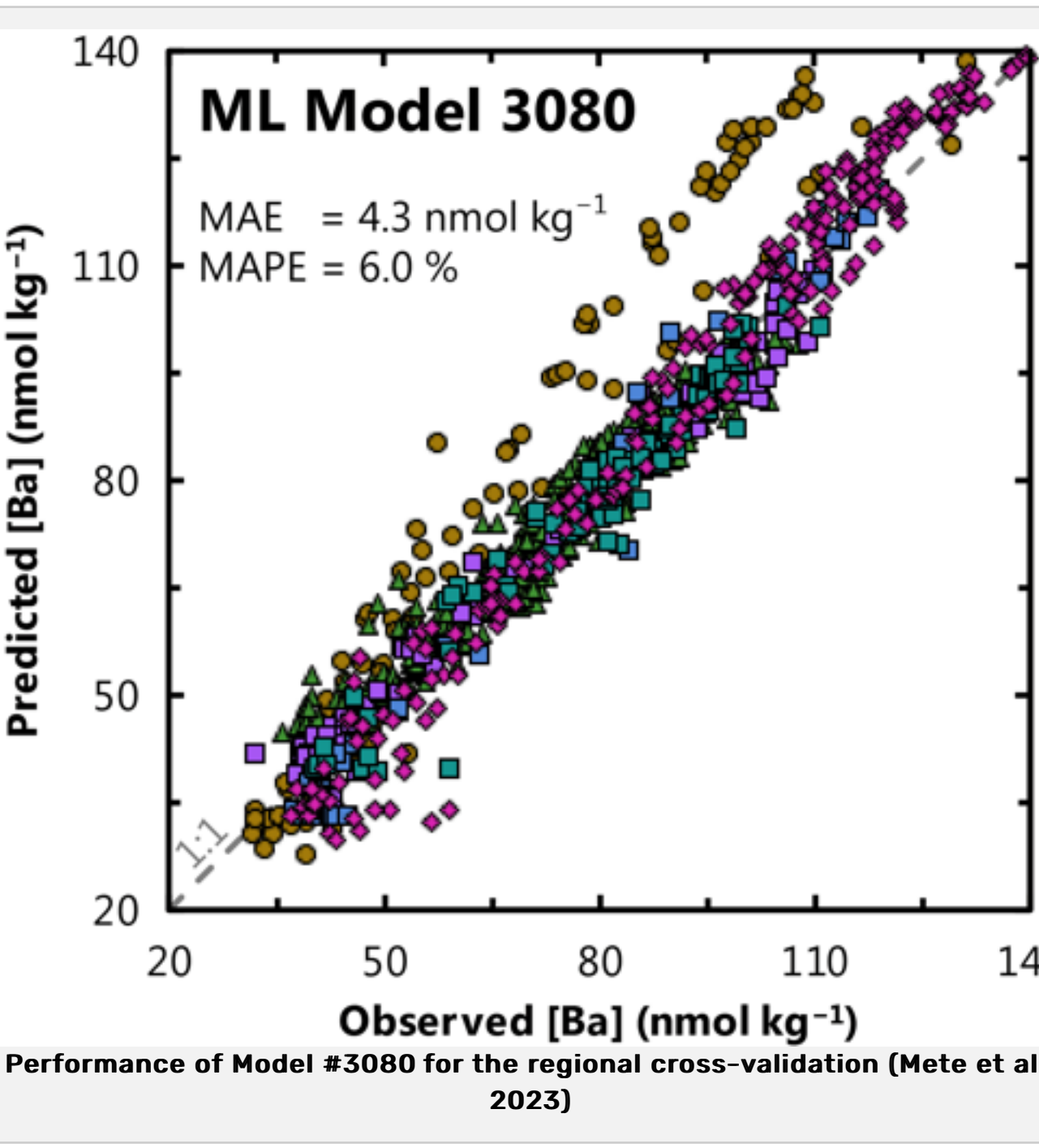
Gaussian Process regression models were trained to predict dissolved barium using various combinations of 12 standard oceanographic features and validated against withheld data from the Indian Ocean.

The factorial design enabled the creation of 2^{12} unique feature combinations, resulting in 4,095 models. The brute force approach was necessary as common feature selection algorithms, like *Maximum Relevance Minimum Redundancy*, did not yield satisfactory results.



Testing the accuracy of the 4,095 ML models. Accuracy, evaluted using Mean Absolute Error, plotted against the number of features in trained models for random holdout cross-validation (top) and regional cross-validation (bottom; Mete et al., 2023). Red square indicates Model #3080.

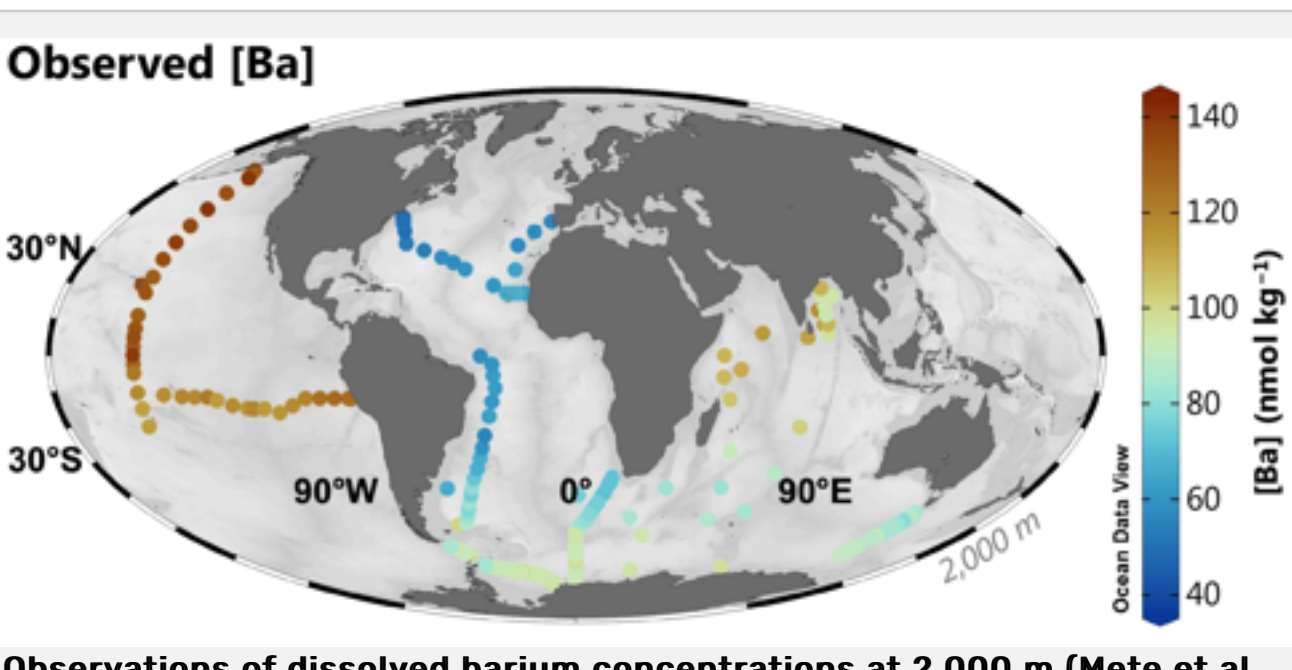
A model using depth, temperature, salinity, dissolved dioxygen, phosphate, nitrate, and silicate accurately predicted barium in the Indian Ocean with a mean absolute percentage deviation of 6.0%.



Performance of Model #3080 for the regional cross-validation (Mete et al., 2023)

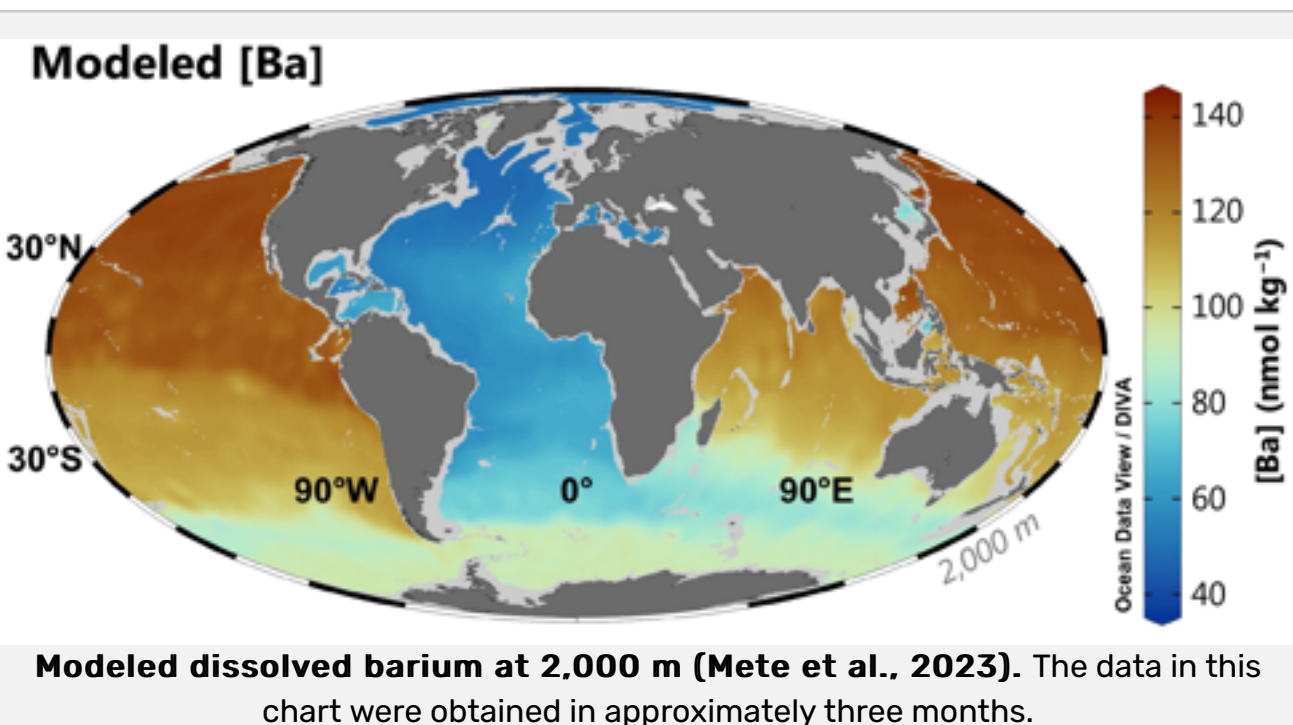
A WHOLE-OCEAN PERSPECTIVE

Improvement in data coverage is epitomized by the figure below, which shows all high-quality dissolved barium data at 2,000 m published to date.



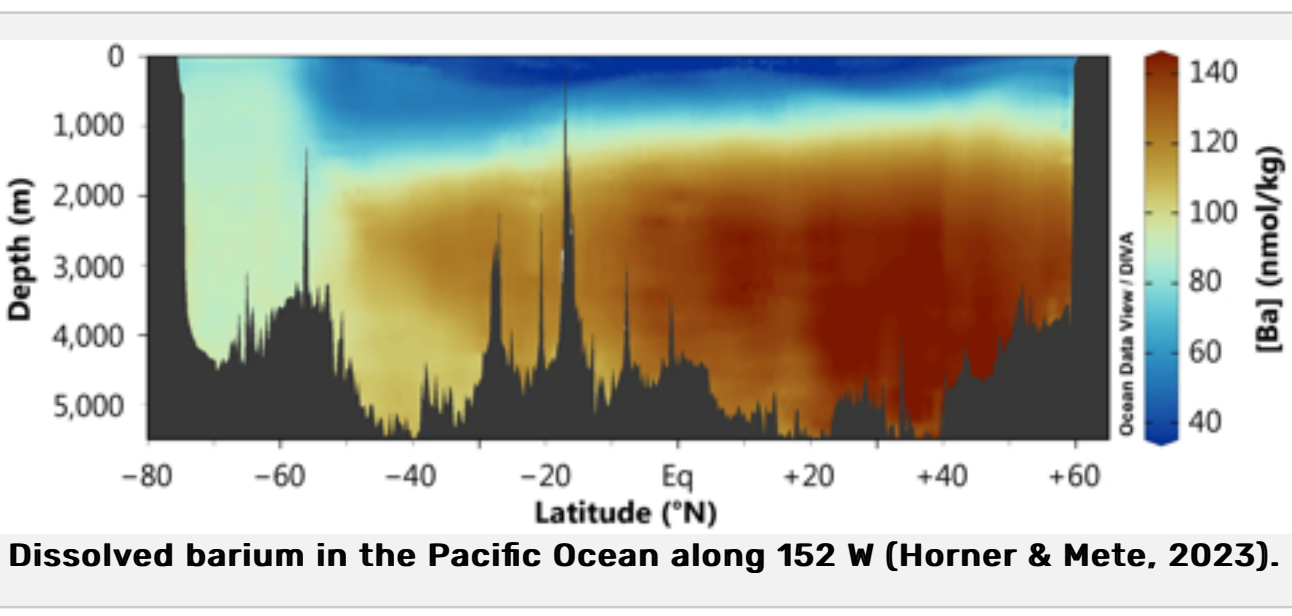
Observations of dissolved barium concentrations at 2,000 m (Mete et al., 2023). The data in this chart took over 40 years to obtain.

We then simulated global dissolved barium distributions by feeding seven predictors from the *World Ocean Atlas* into Model #3080.



Modeled dissolved barium at 2,000 m (Mete et al., 2023). The data in this chart were obtained in approximately three months.

Model output reveals oceanographically consistent variability in tracer distributions. For example, the model reproduces the nutrient-like distribution of barium and shows an increase in concentrations from the Atlantic to the Pacific Ocean.



Dissolved barium in the Pacific Ocean along 152 W (Horner & Mete, 2023).

Model outputs have applications in mechanistic biogeochemical models and paleoproxy calibration.

SUMMARY AND OUTLOOK

Our model for barium demonstrates the utility of ML in simulating tracer distributions. The approach can be extended to simulate other trace elements in seawater.

Indeed, ML (ANNs, GPR, RFR) is starting to be used in biogeochemistry to predict target variables and fill in data gaps to create continuous spatial distribution estimates.

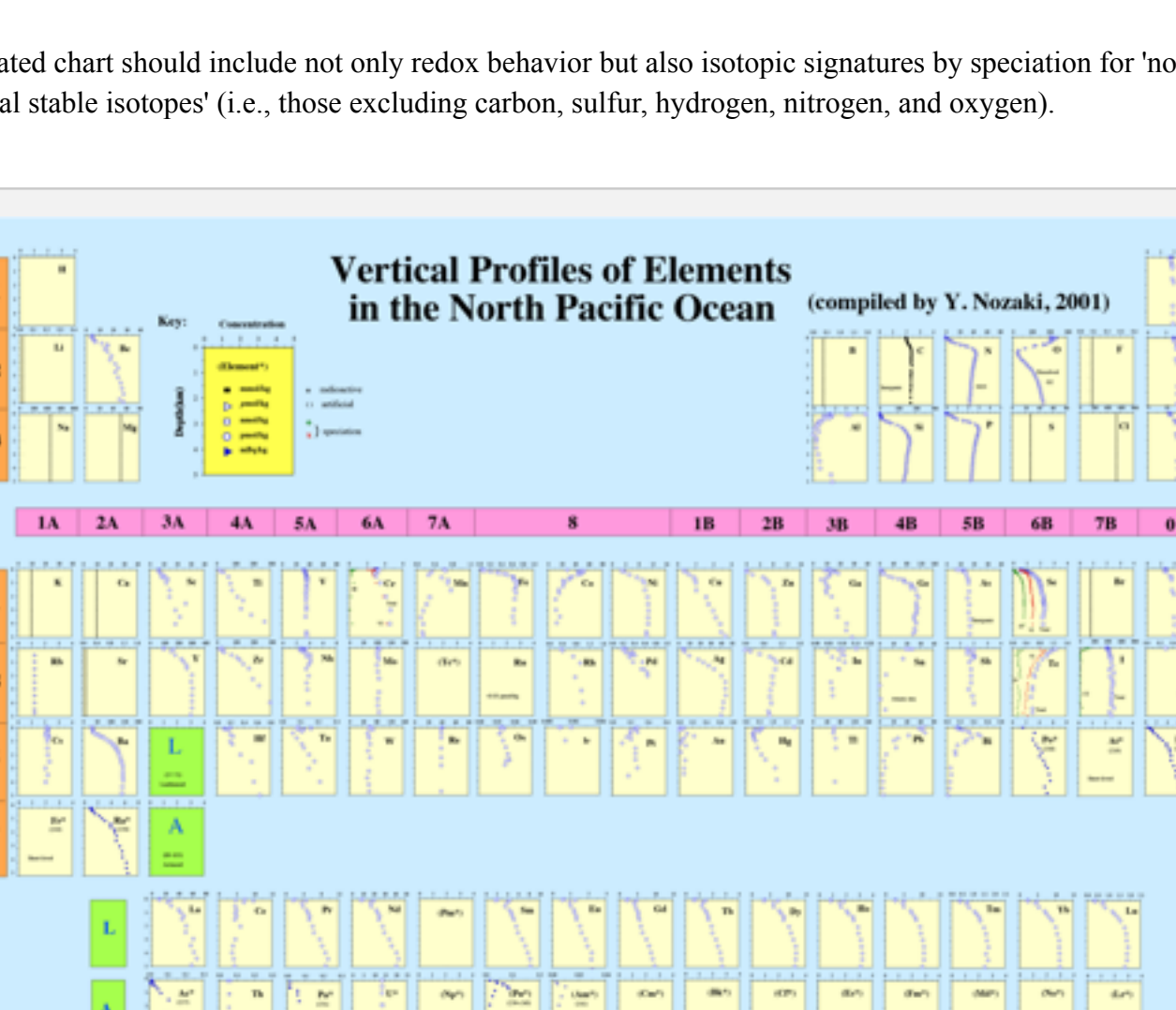
However, several issues remain:

- Data collation, imputation, and QA/QC are significant challenges; thus, most elements have yet to be simulated using ML.
- Standard hydrographic (temperature, salinity, pressure) and biogeochemical (nutrients, oxygen, etc.) predictors may not have sufficient information content to accurately simulate the target variable (e.g., iron, manganese).
- Lack of insight into underlying processes requires diagnostic modeling to infer biogeochemical fluxes.

TOWARDS A 3D 'NOZAKI CHART'

The next challenge is to update Nozaki's chart to a three-dimensional version (Froelich, 2014).

The updated chart should include not only redox behavior but also isotopic signatures by speciation for 'non-traditional stable isotopes' (i.e., those excluding carbon, sulfur, hydrogen, nitrogen, and oxygen).



The iconic 'Nozaki Chart,' named after its compiler and chemical oceanographer Yoshiyuki Nozaki (1947–2003).

AUTHOR INFORMATION

Tristan Horner is an Associate Scientist in the *Department of Marine Chemistry and Geochemistry* at the *Woods Hole Oceanographic Institution* where he is PI of the *NIRVANA Lab* (Non-traditional Isotope Research on Various Advanced Novel Applications). He is an isotope geochemist by training and has a keen interest in chemical oceanography and Earth history. His research focuses on understanding how biological, chemical, and physical processes influence the distribution of metals in the ocean. He also explores how the abundance and isotopic composition of metals in marine sediments can be used to provide insights into the evolution of seawater chemistry.

Lab website: nirvana.whoi.edu

ORCID: 0000-0003-1784-0391

TRANSCRIPT

ABSTRACT

The ocean plays a crucial role in regulating the air-sea balance of CO_2 through ocean circulation and marine biogeochemistry. Micronutrients such as iron, manganese, cobalt, and zinc are essential for ocean primary productivity and biogeochemical cycles, while others, such as barium, serve as valuable proxies of past conditions. Chemical oceanographers focus on studying the spatial, vertical, and temporal distributions of these metals to understand marine biogeochemistry. However, measuring metals in seawater is challenging due to contamination and the vast spatial scales that need to be sampled. To address these challenges, the GEOTRACES program was established 20 years ago to facilitate coordinated international efforts to determine the global distributions of trace elements and their isotopes in the ocean. Here, leveraging GEOTRACES data, we aimed to achieve an accurate global simulation of dissolved barium in seawater using ML. Gaussian Process regression models, trained on a combination of 12 standard oceanographic features and cross-validated using a regional holdout, demonstrated high predictive accuracy with a mean absolute percentage deviation of 6.0 %. Our model outputs reveal oceanographically consistent variability in tracer distributions and will be valuable in mechanistic biogeochemical models and for paleoproxy calibration. Given our results, we believe it is now possible to update Nozaki's chart to a three-dimensional version incorporating more elements and isotopic signatures. While ML shows promise in simulating tracer distributions and filling data gaps, significant challenges remain in data collation, imputation, and quality control. Additionally, standard predictors may lack sufficient information content for accurately simulating some elements. Overall, our study underscores the potential of ML in marine biogeochemistry and sets a foundation for future research on trace element distributions in seawater.

REFERENCES

References cited in this poster

BODC (2024) GEOTRACES Interactive Map. https://www.bodc.ac.uk/geotraces/cruises/section_maps/interactive_map/

Froelich, F. (2014). In praise of marine chemists. *Oceanography*, 27(1), 88-91, doi:10.5670/oceanog.2014.12.

Horner, T. J., Mete, O. Z. (2023) A spatially and vertically resolved global grid of dissolved barium concentrations in seawater determined using Gaussian Process Regression machine learning. *Biological and Chemical Oceanography Data Management Office* (BCO-DMO). (Version 2) Version Date 2023-07-11, doi:10.26008/1912/bco-dmo.885506.2.

Mete, Ö. Z., Subhas, A. V., Kim, H. H., Dunlea, A. G., Whitmore, L. M., Shiller, A. M., Gilbert, M., Leavitt, W. D., and Horner, T.J. (2023). Barium in seawater: dissolved distribution, relationship to silicon, and barite saturation state determined using machine learning, *Earth System Science Data*, 15, 4023–4045, doi:10.5194/essd-15-4023-2023.

Morel, F. M., & Price, N. M. (2003). The biogeochemical cycles of trace metals in the oceans. *Science*, 300(5621), 944-947, doi:10.1126/science.1083545.

Nozaki, Y. (2001). Elemental distribution: Overview. *Encyclopedia of Ocean Sciences*, 840-845, doi:10.1006/rwos.2001.0402.

Other applications of ML to chemical oceanography

Huang, Y., Tagliabue, A., & Cassar, N. (2022). Data-driven modeling of dissolved iron in the global ocean. *Frontiers in Marine Science*, 9, 837183, doi:10.3389/fmars.2022.837183.

John, S. G., Liang, H., Pasquier, B., Holzer, M., & Silva, S. (2024). Biogeochemical fluxes of nickel in the global oceans inferred from a diagnostic model. *Global Biogeochemical Cycles*, 38(5), e2023GB008018, doi:10.1029/2023GB008018.

Roshan, S., DeVries, T., Wu, J., & Chen, G. (2018). The internal cycling of zinc in the ocean. *Global Biogeochemical Cycles*, 32(12), 1833-1849, doi:10.1029/2018GB006045.

Roshan, S., DeVries, T., & Wu, J. (2020). Constraining the global ocean Cu cycle with a data-assimilated diagnostic model. *Global Biogeochemical Cycles,* 34(11), e2020GB006741, doi:10.1029/2020GB006741.

Roshan, S., & DeVries, T. (2021). Global contrasts between oceanic cycling of cadmium and phosphate. *Global Biogeochemical Cycles*, 35(6), e2021GB006952, doi:10.1029/2021GB006952.