Geospatial Machine Learning for the Earth Sciences

Warren T. Wood

U. S. Naval Research Laboratory

with help from: Taylor Lee, Benjamin Phrampus, Jeffrey Obelcz, Jordan Graw, Maureen Walton, Matthew Hornbach, Patrick Duff

Presented to the
Decadal Survey of Ocean Sciences for NSF Committee Meeting,
Topic: Opportunities for AI/ML to Advance Ocean Sciences

February 16th, 2024, University of Southern Mississippi Marine Research Center
1030 30th Ave, Gulfport, MS 39501

POC warren.wood@nrlssc.navy.mil

Can Machines do Earth Science?
They can certainly help

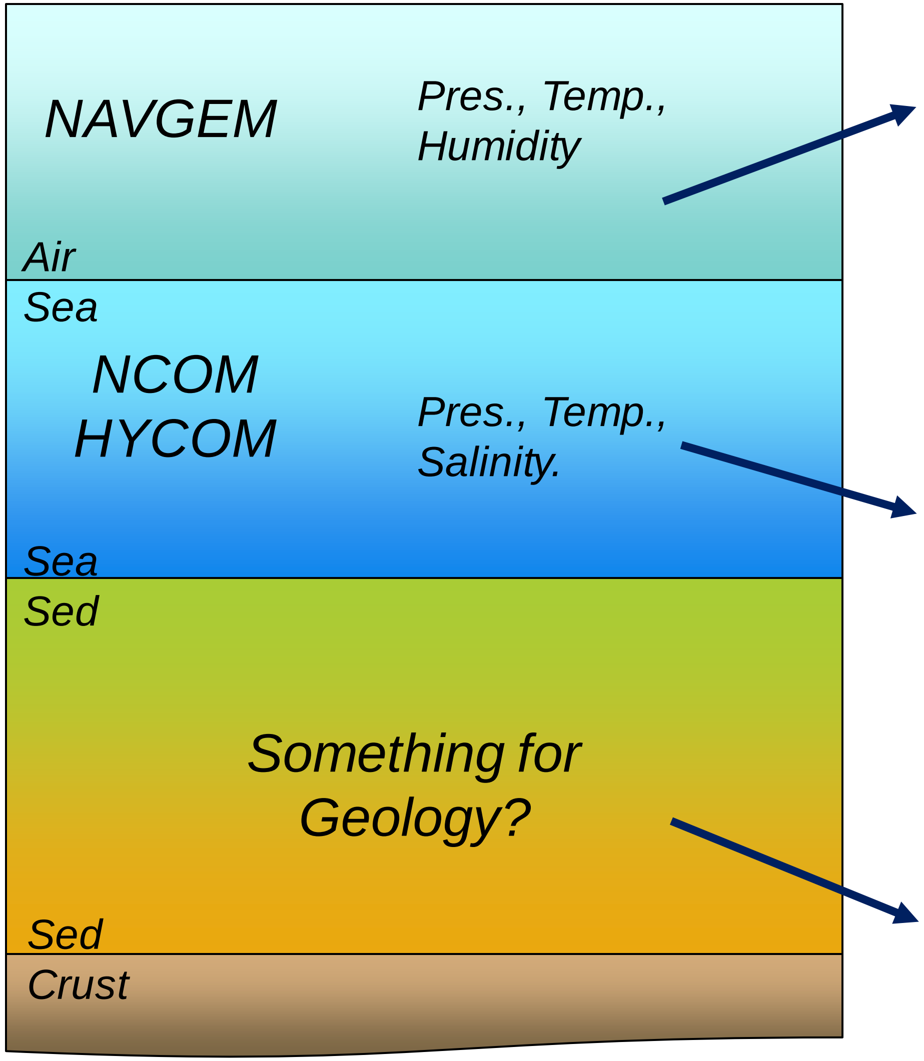
Past: Humans program machines to perform simple, repetitive tasks.
Present: Humans provide examples → machines perform increasingly sophisticated tasks.
Future: Self-teaching/learning? Can all humans even do this?

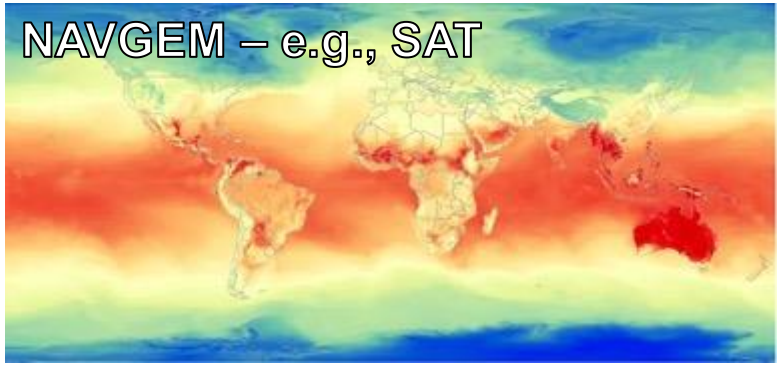
AI/ML (Artificial Intelligence / Machine learning) is a very powerful software tool.

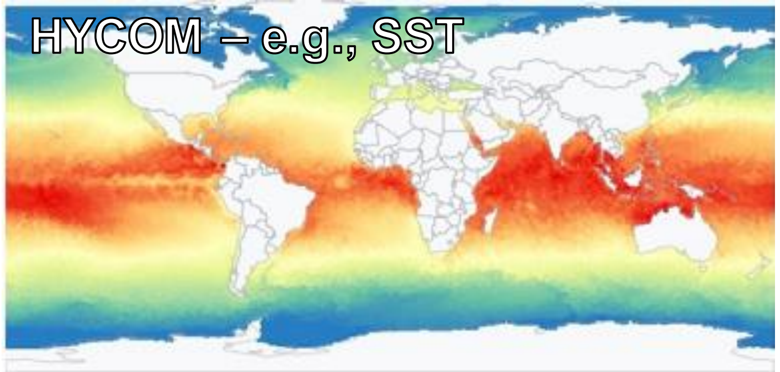
Which inventions are so ingrained in our way of life that we no longer think of life without them?

1850s – photography
1940s – Programable computers
1950s – Mass vaccinations
1980s – Personal computers & office software
2010s – Internet capable smart phones
2020s – AI/ML?, Generative AI? Deep fakes?
History will tell!

Motivation: Global Environmental Prediction







e.g., Fraction sand?

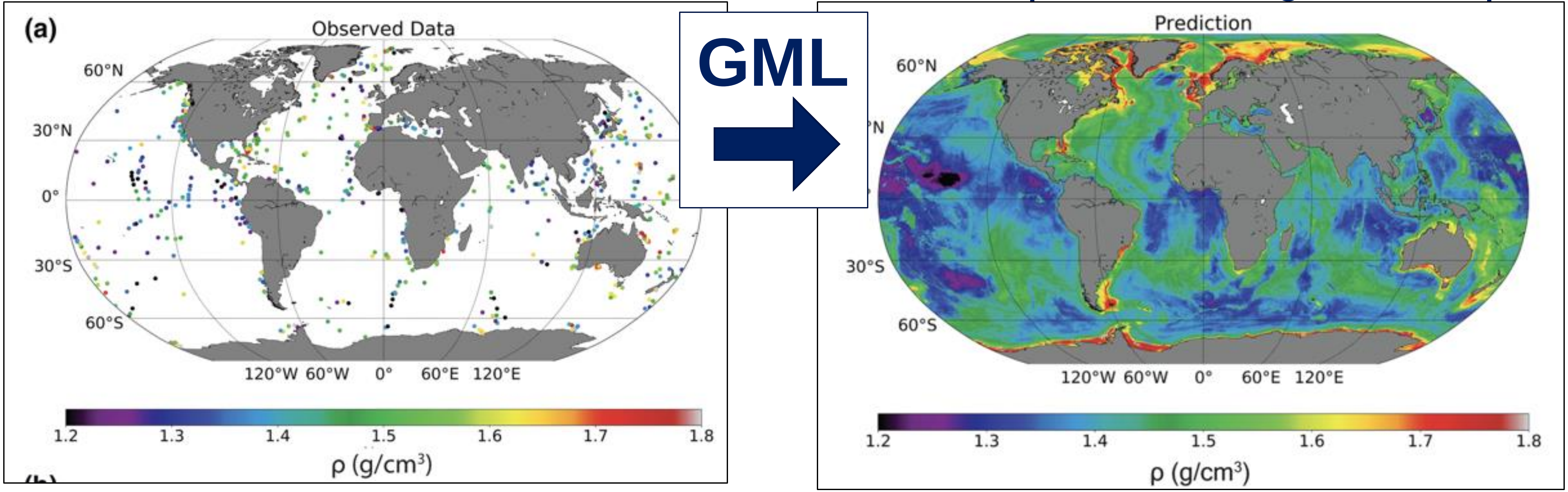
Geospatial Machine Learning

- We apply GML to quantify the properties of interest on a global scale
  - Predict properties where no observations exist (e.g., denied areas)
  - Quantify uncertainties based on predictive skill
  - Locate surrogate areas where future data collection can improve future predictions

One Example of AI/ML in Earth Science: Geospatial Machine Learning (GML)

Sparse (expensive) observations

Comprehensive global map



Seafloor density, Graw et al., 2020

But how does it actually work?

1) Data Curation: We must acquire examples of the quantities we wish to predict. Much of geology has historically been performed in "postage stamp" or small areas for purposes of resource extraction. The data can be in many different forms and formats.
We need: x,y,value

2) Predictor (Feature) Development: We Predictors are quantities that we know where we have observations, and where we want to predict observed values. We know very few quantities everywhere on earth. Global Bathymetry (and topography) maps contain significant information about the subsurface through their geospatial statistics.

3) Machine Learning:

a) Training searches through existing data and predictors to find the predictors that correlate best with the data. Internal (e.g. 10 fold) validation ensures that predictive skill is optimized.

b) Prediction uses the correlations between observations and predictors, to predict values where we have no data. The uncertainty in that prediction is based on the strength of the correlation between predictors and observations.

4) Conformal Uncertainty is based on the validation

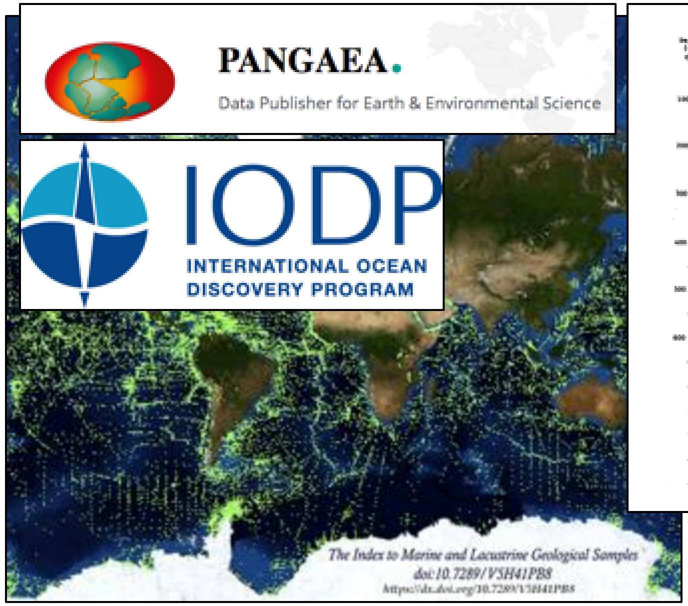
GML – Lot’s of data, poorly curated

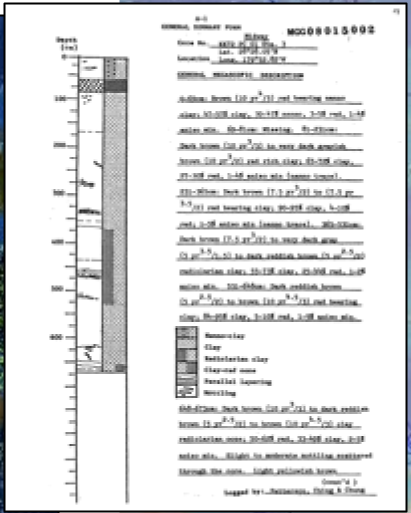
- Predictions require observed data that span range of expected values
- Exploit what currently exists
- Growing need to curate other data formats (e.g., pdfs)

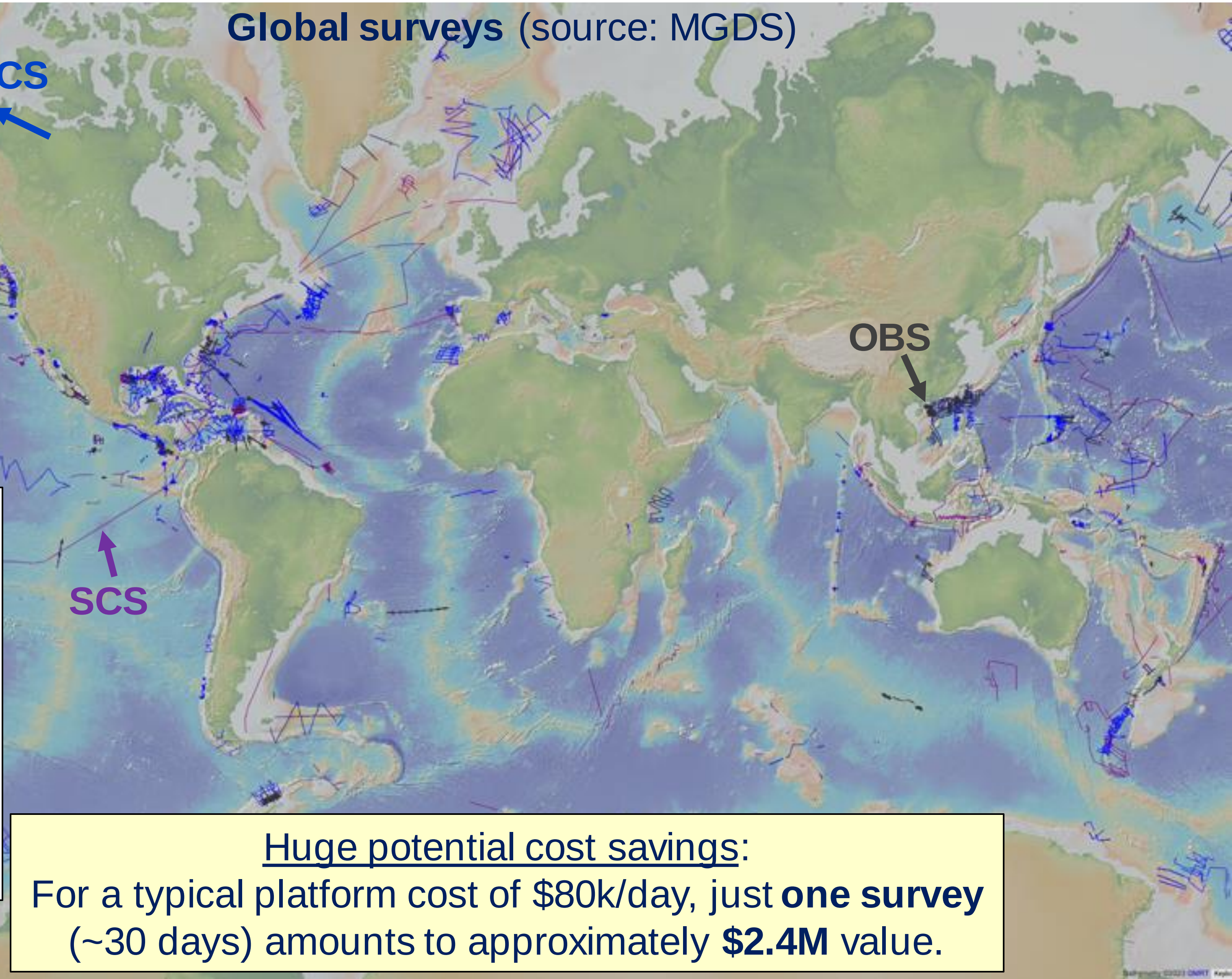
EarthCube

PANGAEA.
Data Publisher for Earth & Environmental Science

IODP
INTERNATIONAL OCEAN DISCOVERY PROGRAM







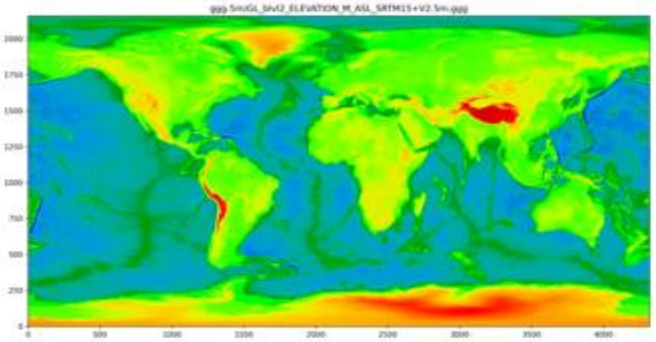
Huge potential cost savings:
For a typical platform cost of \$80k/day, just **one survey** (~30 days) amounts to approximately **\$2.4M** value.

GML – Predictors – (Features)

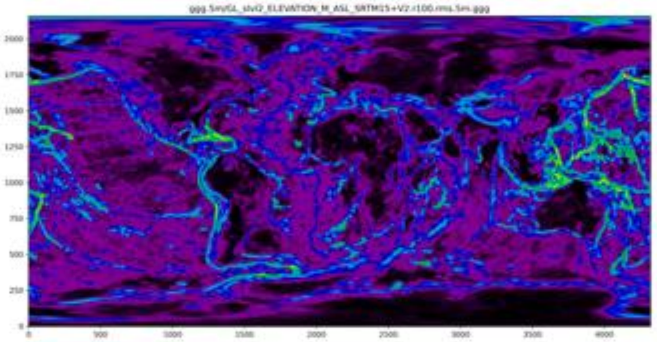
Predictors are values we know everywhere. Observations are correlated with predictors where we have observations, and we use those correlations to predict what we would observe in places where we have no direct observations.

We use quantities that are known or estimated globally, and apply spatial statistics over various radii to generate many thousands of predictors

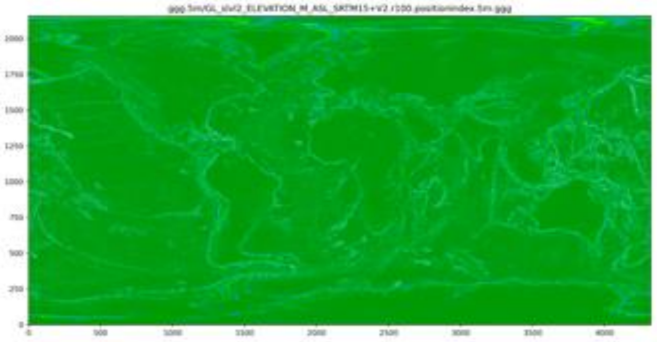
Elevation



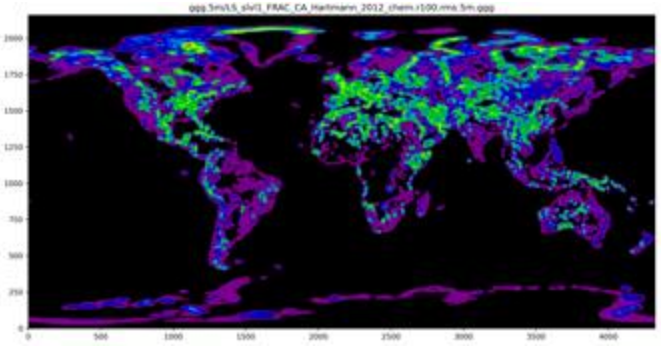
RMS roughness (100m)



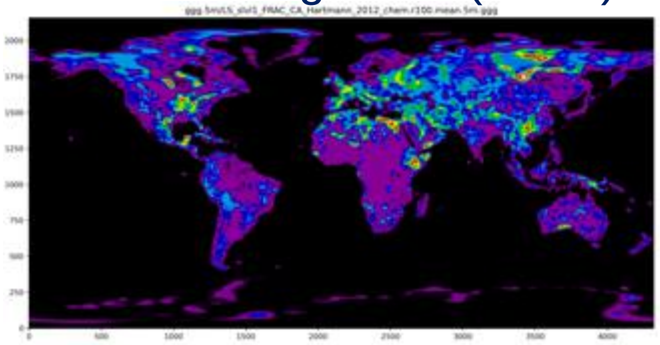
Position Index (100m)



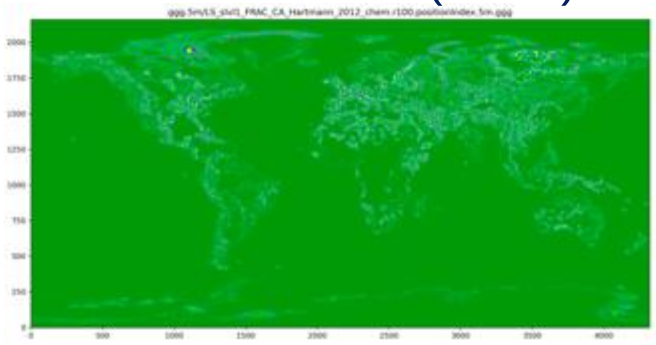
Fraction Calcite



RMS roughness (100m)



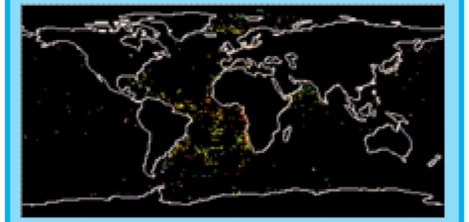
Position Index (100m)



GML: Machine Learning (KNN, Random Forest, SVM, etc.)

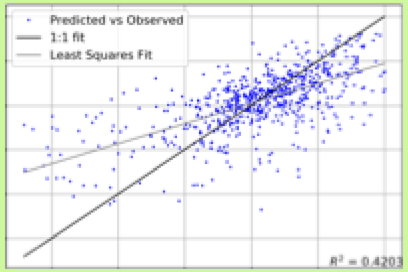
(blue – data)

Observed Data



Feature Selection

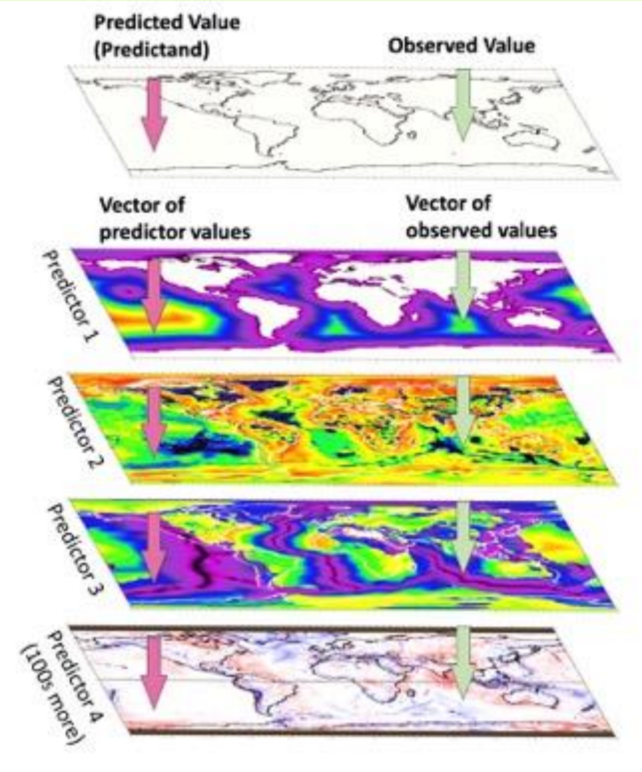
Use only best 20-50 predictors, based on individual predictive skill; 10-fold validation



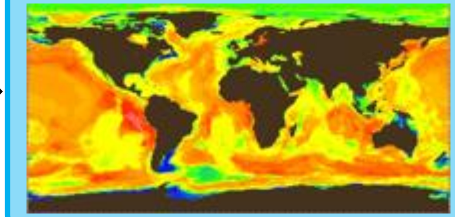
(Green - algorithms)

Machine Learning

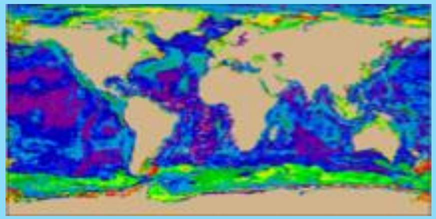
Find correlations



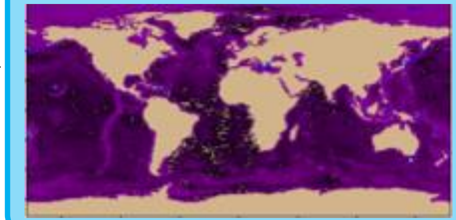
Nowcast



Uncertainty



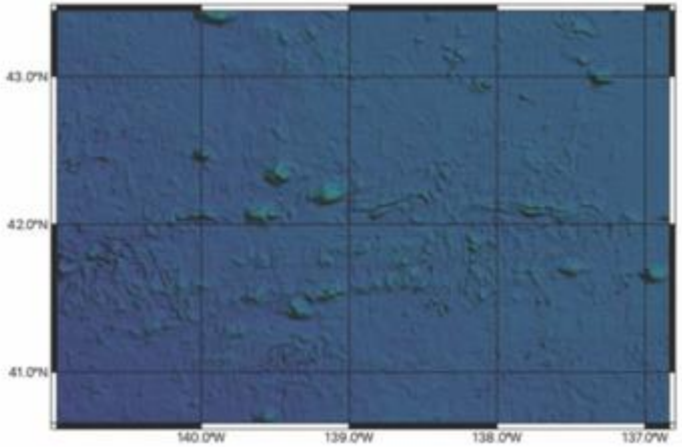
Guide to next obs.



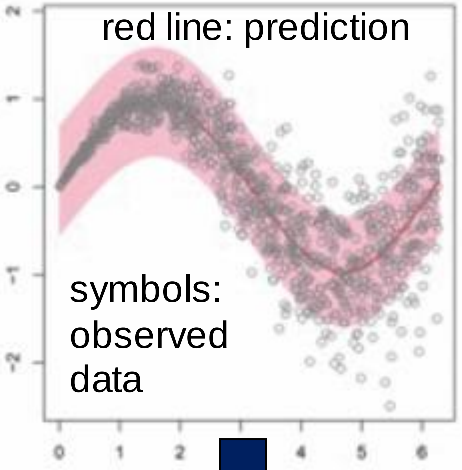
GML – Conformal Uncertainty

Conformal uncertainty is a powerful and easily accomplished method of estimating uncertainty, but it is ***NOT*** easily explained!

GML Predicted value (e.g. bathymetry)



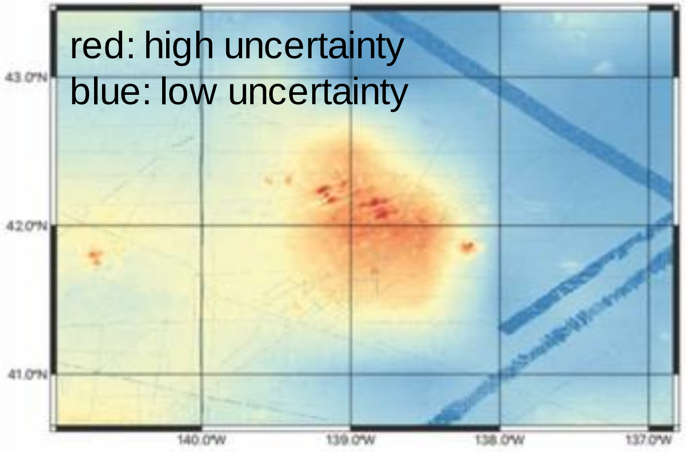
Validation (training) data

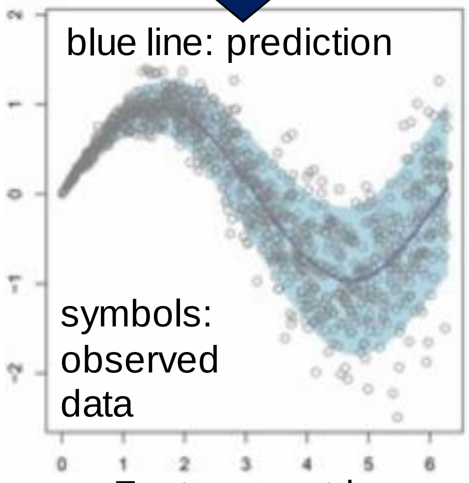


The prediction uncertainty is represented here by the difference between predicted (red line) and observed values (open circles), when a portion of the data has been withheld (e.g. 10 fold validation).

The pink shade represents the simplest way to represent uncertainty, a single number for the entire prediction. However this single number overpredicts the uncertainty in some places (left hand side), and underpredicts the uncertainty in other areas (right hand side). We can do better.

GML Predicted uncertainty





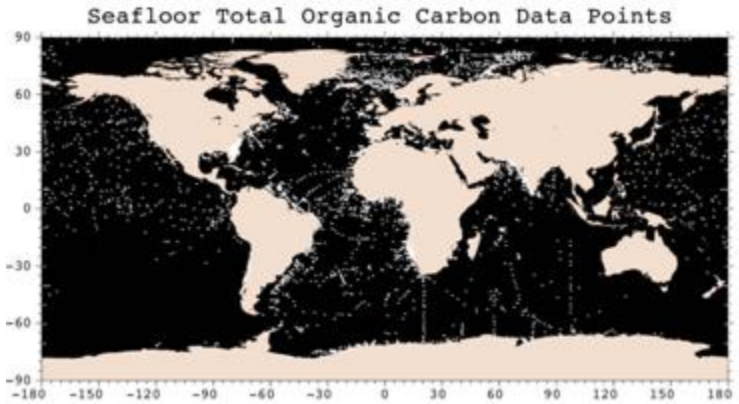
We perform a separate prediction, where the predictand is the difference between the predicted and observed values. This second prediction is assumed to be a measure of the non-conformity of the original prediction, and is used to normalize the uncertainty, while encompassing a user defined amount of the data (e.g. 1 sigma). This results in an uncertainty estimate indicated by the blue shaded area (left).

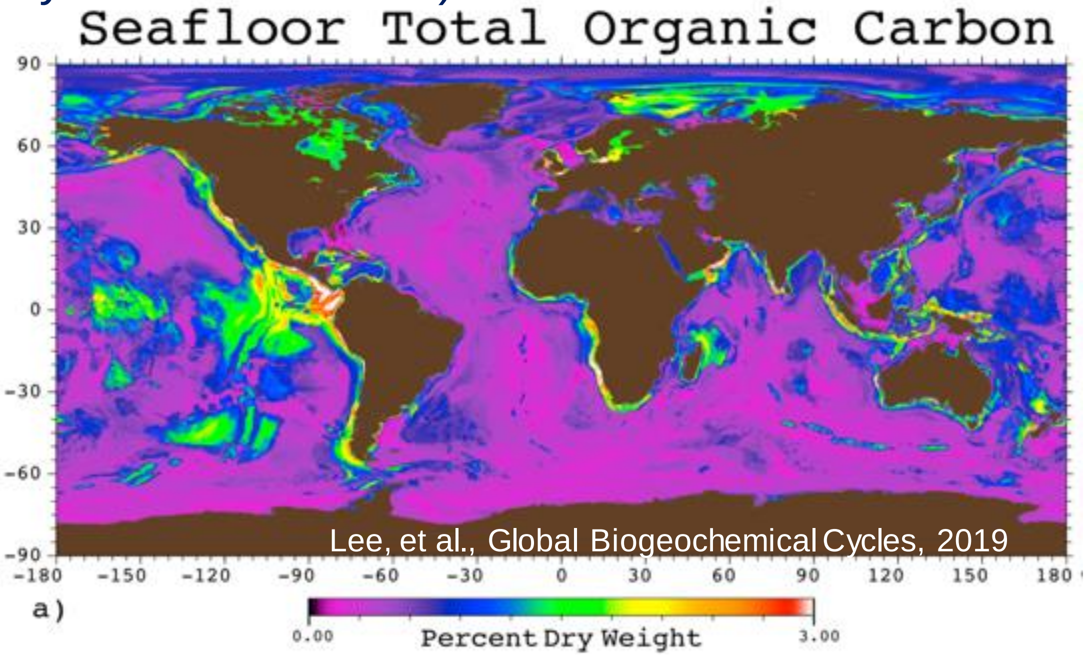
The final predicted uncertainty (far left) is small where the original validation is small, and large where it was large.

Published Examples

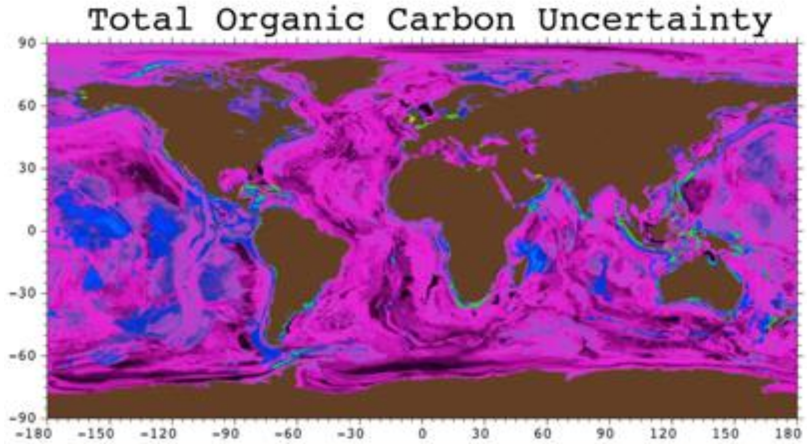
Global Seafloor Total Organic Carbon

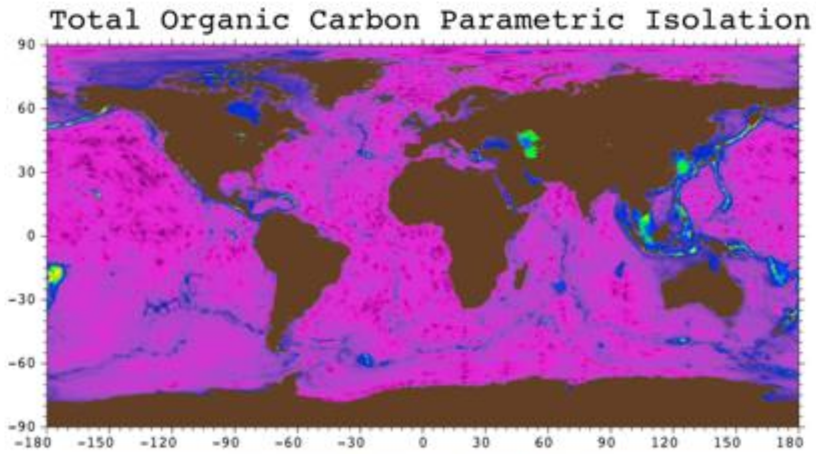
(POC Dr. Taylor Lee NRL 7352)





Sparse observations (above) are interpolated (upper, right) via machine-learned correlations with a variety of predictors. The analysis yields an uncertainty (below) and an estimate of where to sample next to best improve the prediction (lower, right).

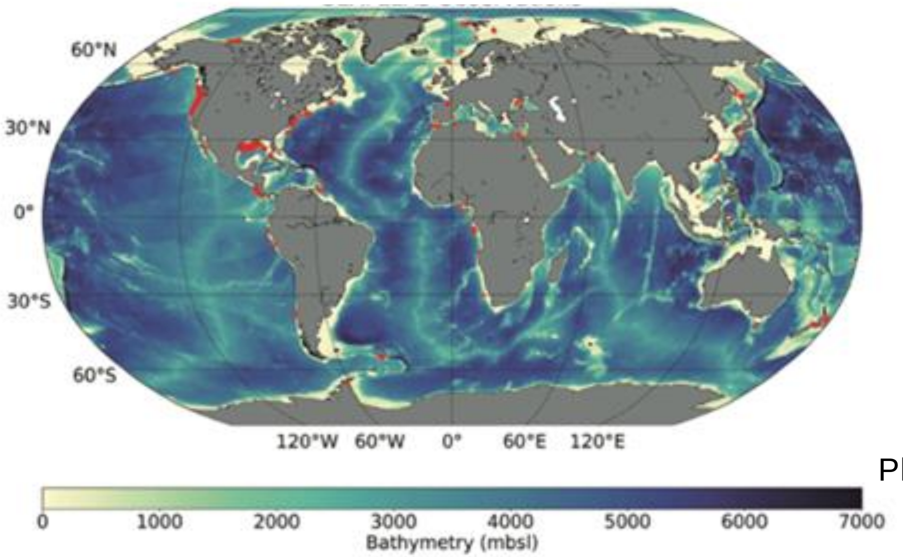




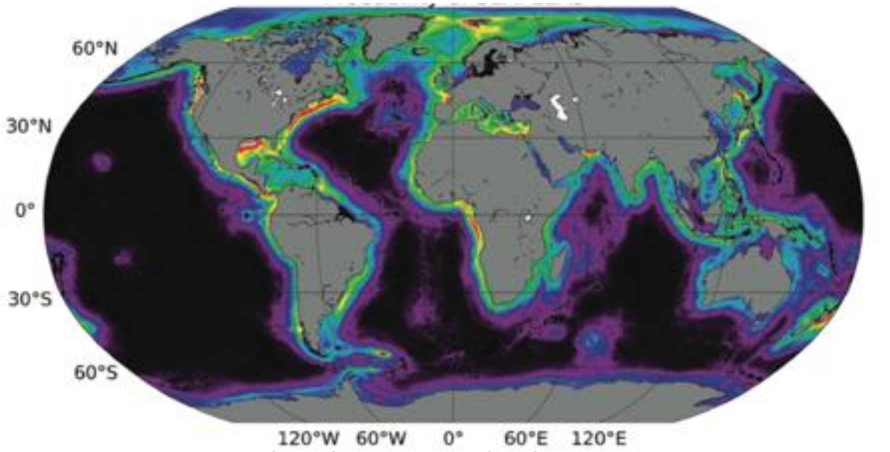
Global Prediction of *SEAfloor FLuid Expulsion Anomalies (SEAFLEAs)*

(POC Dr. Benjamin Phrampus NRL 7352)

Observations

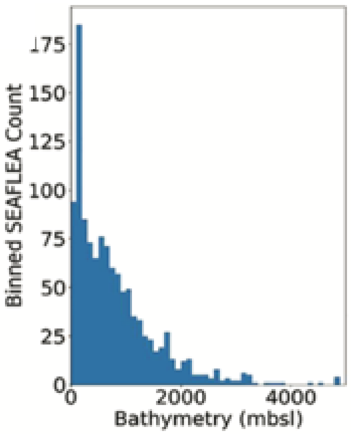


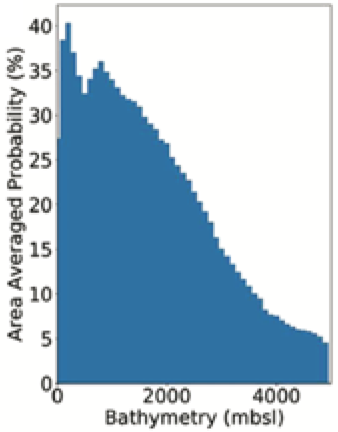
Probability of Occurrence



Phrampus, et al., Geochemistry, Geophysics, Geosystems, 2020

The SEAFLEA distribution is heavily biased toward North American continental margins (above, left), with most observations between 100- and 200-m water depth globally. Our final prediction (above right) reveals a more even distribution with depth (right) and roughly equal probability of SEAFLEAS occurring on passive and active margins.

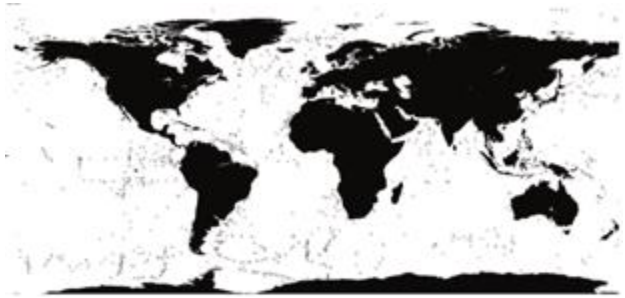


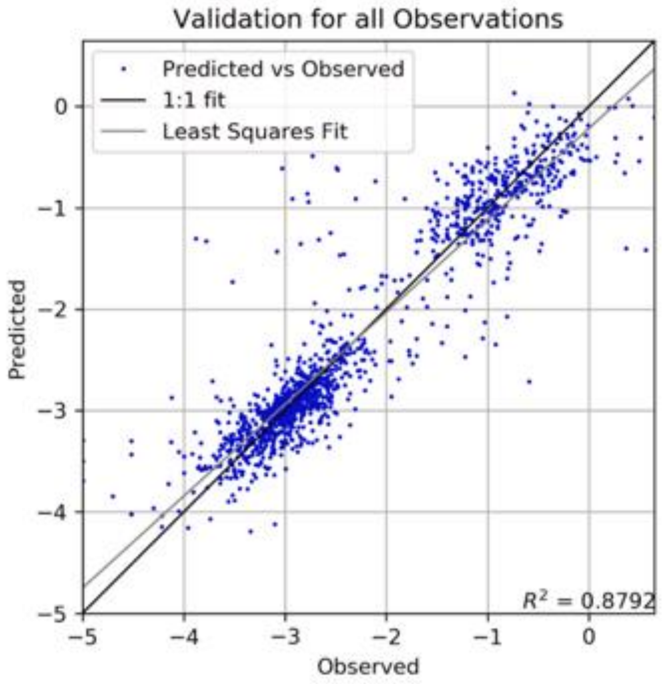


Global Mass Accumulation Rates (MAR)

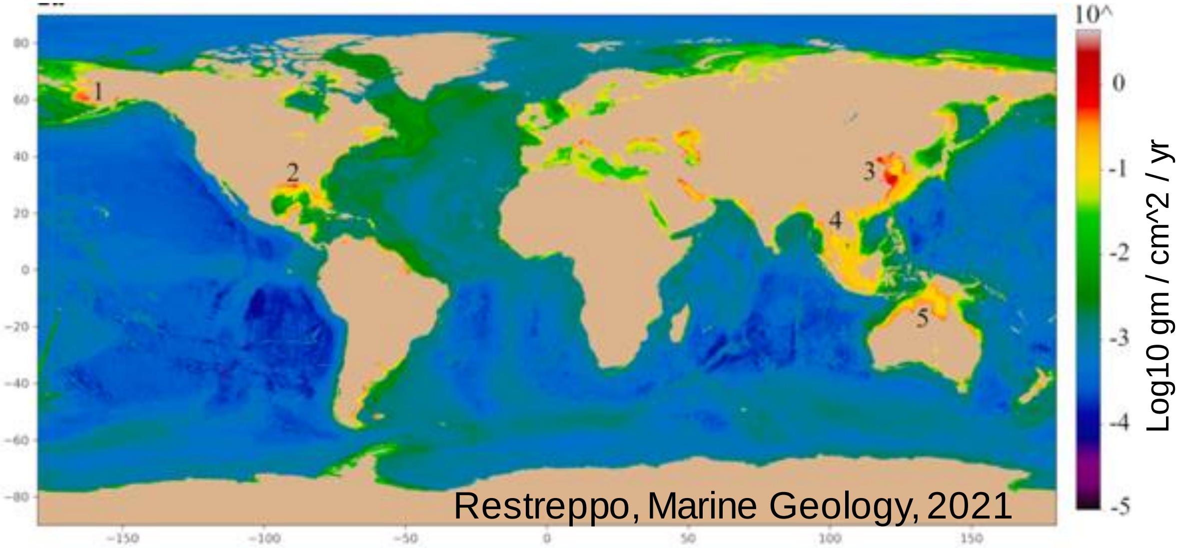
(POC Dr. Giancarlo Restreppo NRL 7352)

Observed MAR



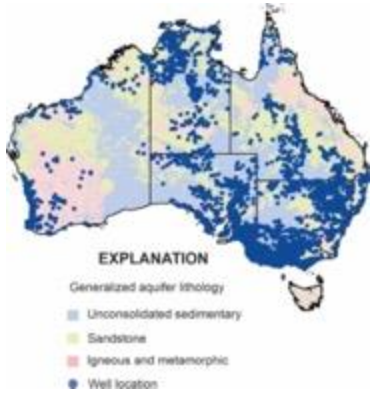


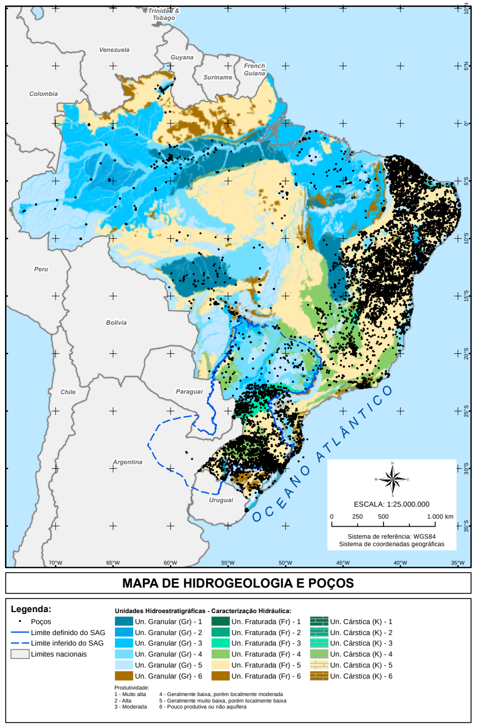
Predicted MAR

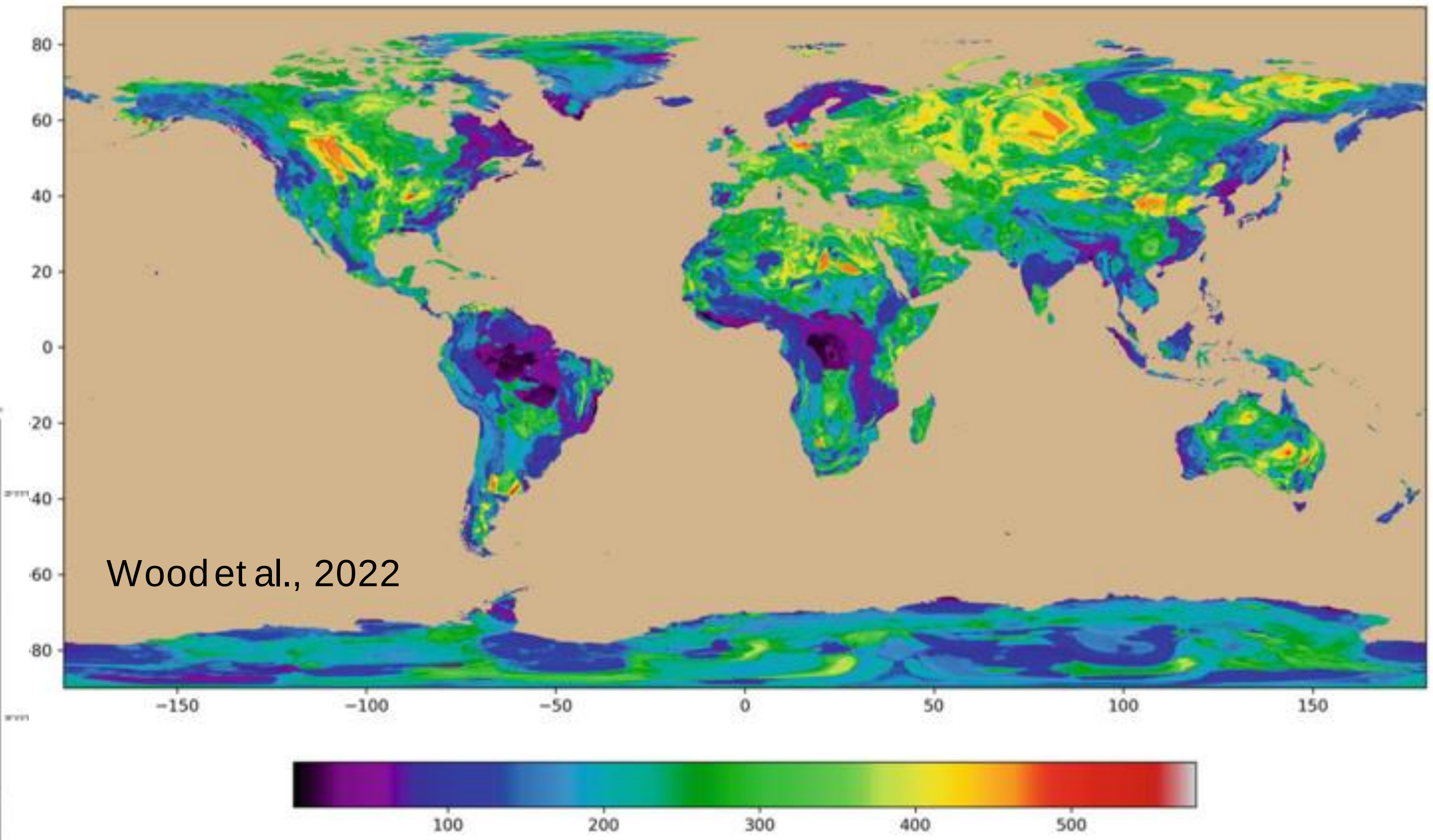


Predicted benthic MARs in log10-space, in gm / cm^2 / yr. Highest rates at; 1) Yukon–Kuskokwim Delta, 2) The Mississippi River Delta, 3) the Huang He and Yangtze River mouth area, 4) Southeastern Asia and associated islands, and 5) Oceania. We predict a total sediment load of ~3.3 × 10^4 Mt. / year on the global seafloor.

Global Bicarbonate (Land)







Map of calculated global bicarbonate concentration distribution in active aquifer systems. Vertical scale degrees latitude, horizonal scale degrees longitude, and concentrations in mg/L are given in color.

End