Training Datasets for Modeling with AI across the Deep-Ocean Seafloor

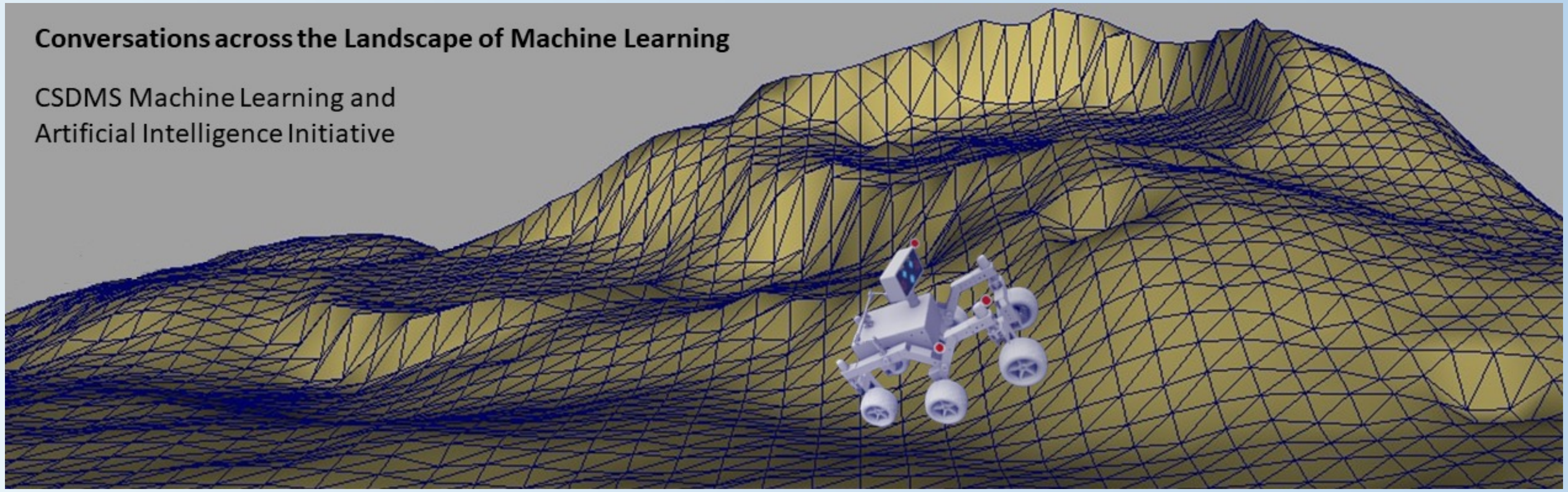
Chris Jenkins, INSTAAR CU Boulder

PI for 'dbSEABED'

Machine Learning AI&ML Group in CSDMS 2020

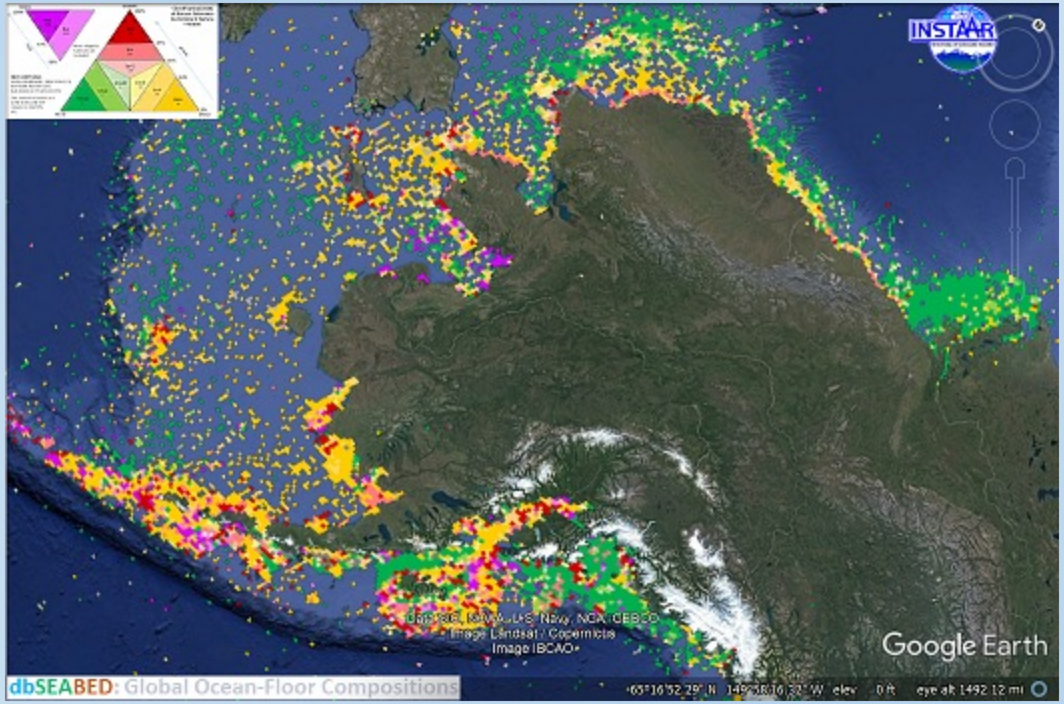
(Chris Jenkins, Daniel Buscombe, Evan Goldstein, Kelly Kochanski, Jeffrey Obelcz

- Decided: best strategy to help CSDMS members to pick up & use ML is for CSDMS to provide training data
- At the same time it's been observed that the deep-sea tends to be neglected as an activity in CSDMS



Background

- dbSEABED Project objective – maps and database for the **materials of the seabed**, everywhere maximizing data and optimizing the math / algorithm methods; resolve facies on 10km scales worldwide
- Research **Heterogeneous Data methods** – leveraging the massive amounts of data that have been collected by oceanographers over the decades; difficult, but with many advantages



Get the world coverage: *tinyurl.com/dbseabed/kml/*

- **Applied to research and ocean management** – fisheries, biogeochemistry, habitats, contamination monitoring, safe navigation, marine conservation, sonar prediction, mine-countermeasures, stratigraphy, deposition/erosion, seafloor stability, paleoceanography; over 100 cases
- A persistent problem: How to **create the best gridded data from sparse point data** ? Avoid pitfalls ! Realistic spatial predictions. THAT is why we look to MACHINE LEARNING

Overview of ML in our field

- “learn from already labeled data how to predict the class of unlabeled data “; “machine memory, not machine learning”

- Informative **example papers (see refs)** …
    - Lee Wood & Phrampus 2019. A machine learning (kNN) approach to predicting global seafloor total organic carbon.
    - Dutkiewicz Müller O’Callaghan & Jónasson 2015. Census of seafloor sediments in the world’s ocean
    - Restreppo Wood Phrampus 2020. Oceanic sediment accumulation rates predicted via machine learning algorithm: towards sediment characterization on a global scale

- The range of **AI methods** includes Random Forest, Neural Network, Support Vector Machine, K-Nearest Neighbors; SciKit-Learn is our starting package

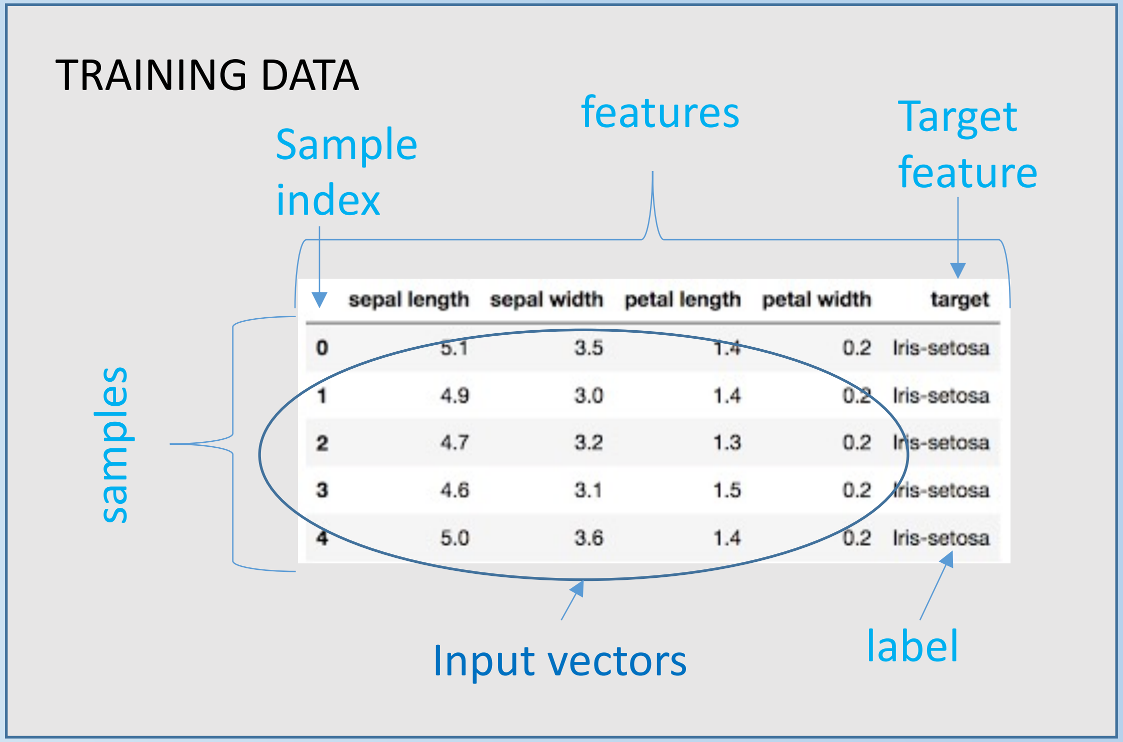
- **Training data** – absolutely critical, and difficult to compile on large scales

- Important separations: **Supervised** and Unsupervised, **Regression** and Classification

Overview of ML in our field (cont)

Primer on Training Data

- Column-wise data, ‘input vectors’, with headers
- ‘Features’ – each parameter, attribute
- ‘Labelled Data’, ‘Target Values’
- ‘Labels’ – the desired output values on the training data
- ‘Target’ the attribute to predict
- ‘Dimensionality’ – here 4

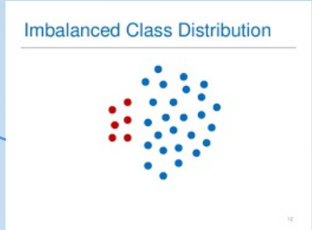


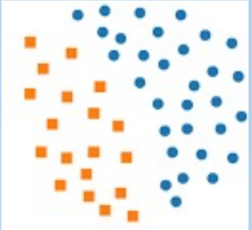
	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Overview of ML in our field (cont)

Common Training Data issues

- Missing attribute values: fix by imputation, completing (e.g., with feature medians), or deletion
- Standardizing – usually centered on MEAN and scaled by the STDEV
- Not too many features (Random Forest)
- Imbalanced data (fix by down- or up-sampling, or reduction)
- Incomplete parameter values coverage
- Feature selection / reduction
- Overfitting

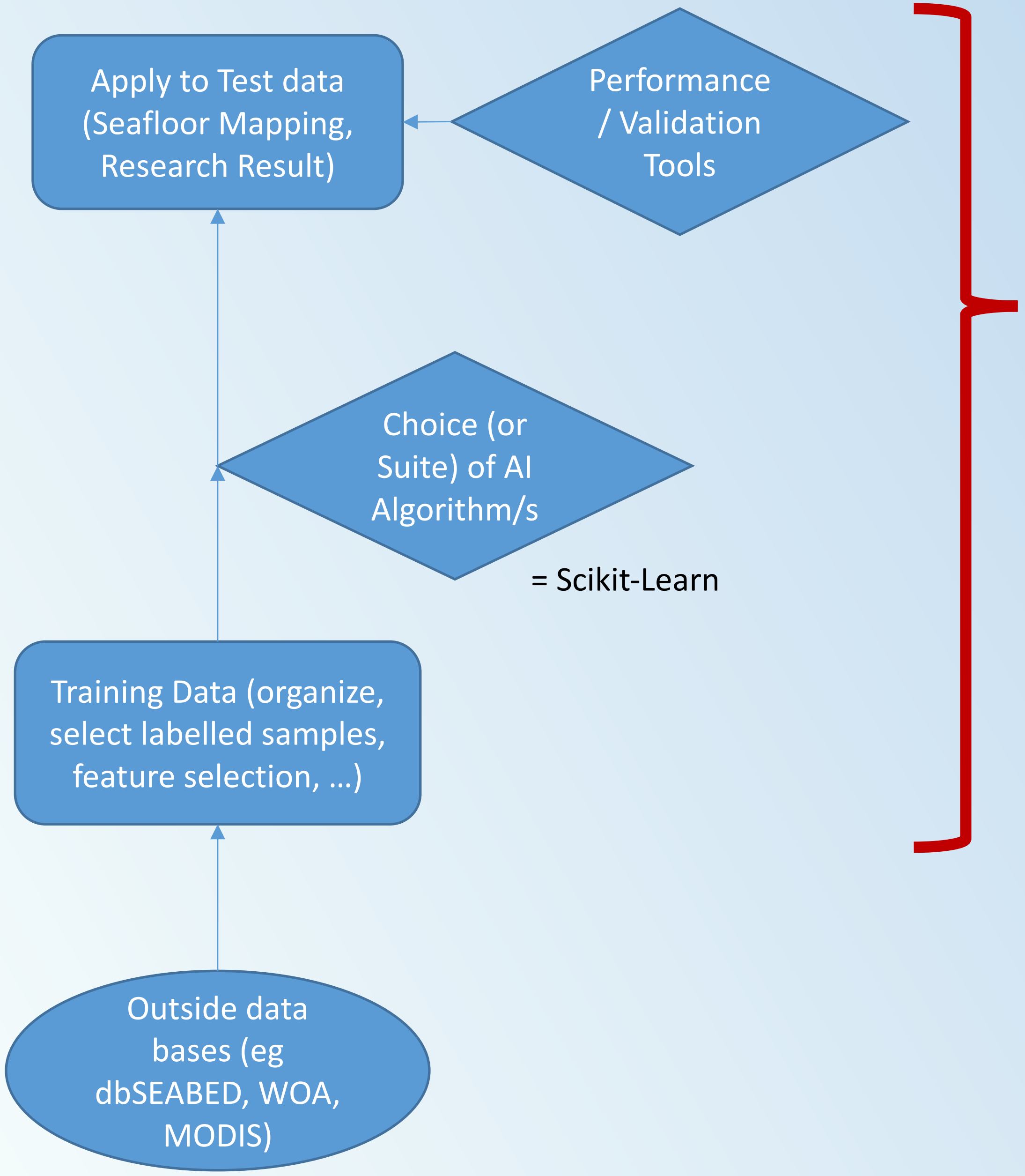




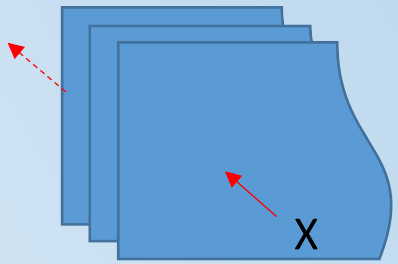
- The modeler is not a specialist on the data (e.g., geologist using physical oceanography data)

Specifically in dbSEABED …

The processing flow for ML-Mapping of Seafloor Properties



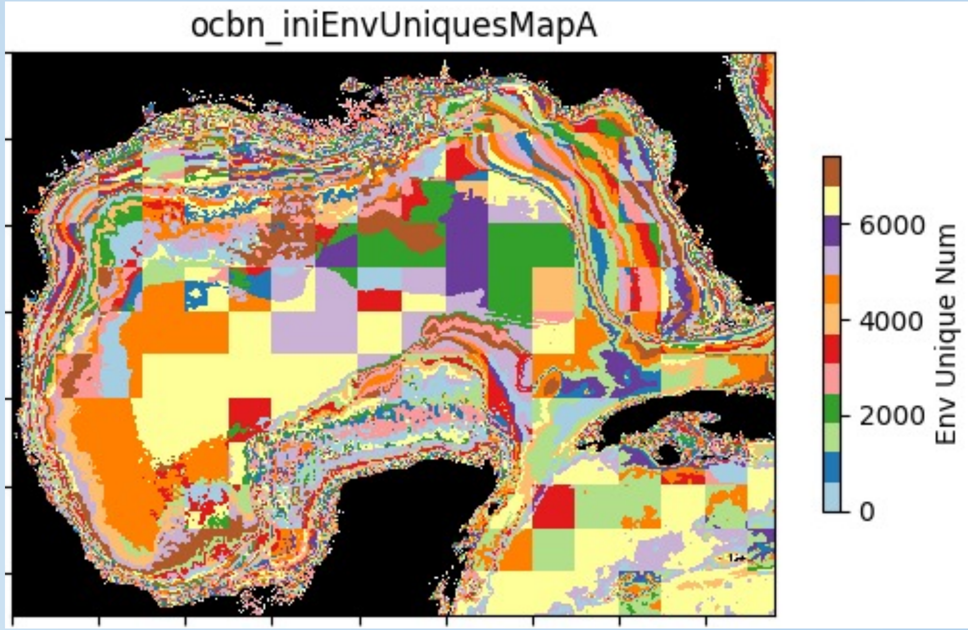
Software package
“Contributed Model”



Digital stack of the Terrain, Geo and Enviro data grids, sampled by the Training and Test data; skewering at Labelled data locations

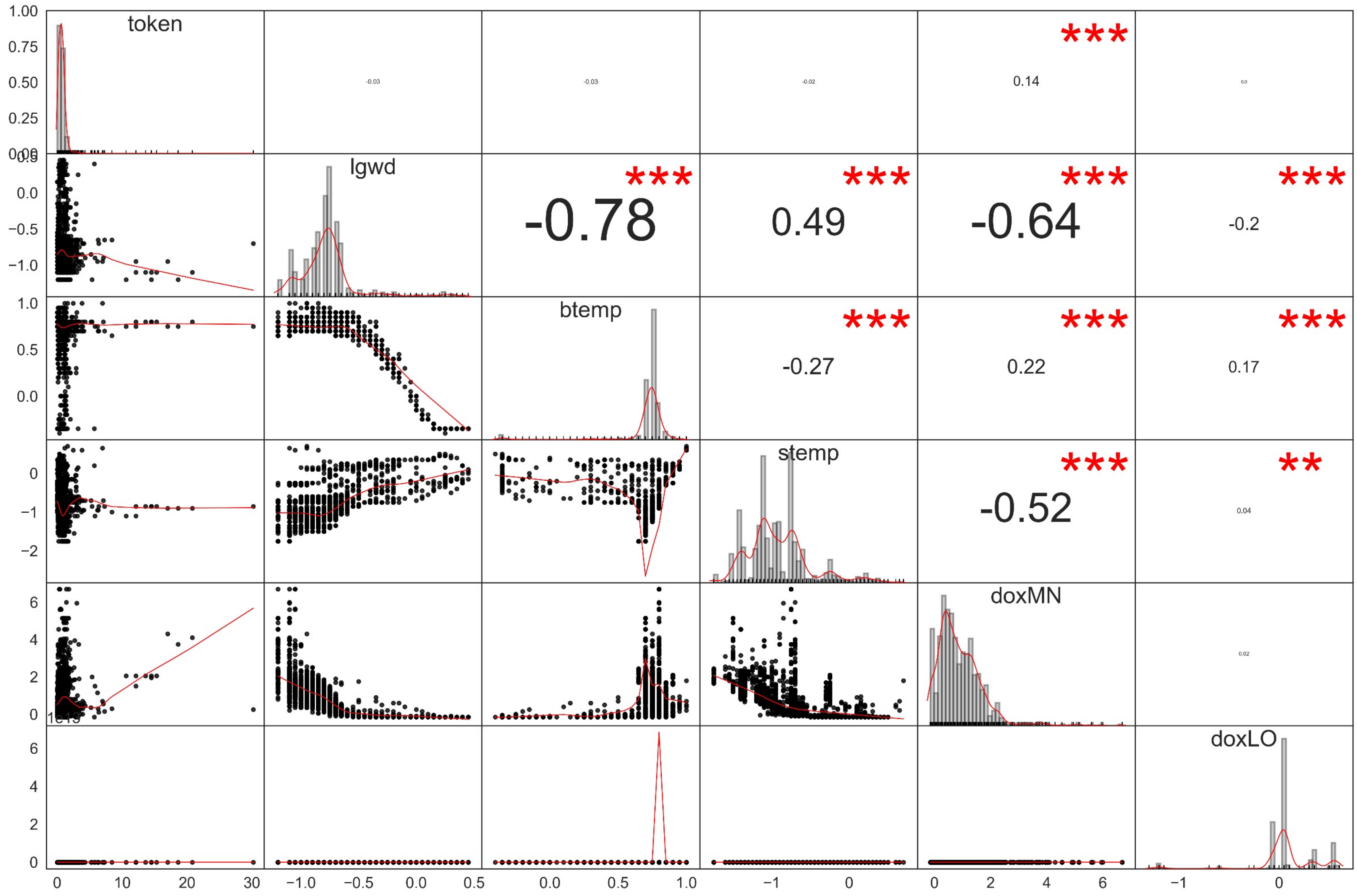
Specifically in dbSEABED (cont)

- Assemble griddings of the Training parameters
- Normalize and step the grid-values
- Create uniques-classified map area of the grid-stack; assign an index to each unique area, and associate the training parameter values
- Assemble labeled Training samples with Target-parameter values and locations; pare down to a set of uniques with no blanks
- Attach Training parameter values from the stack to the Labelled locations using their locations
- Run the ML algorithm/s, collect performance metrics
- Transfer the outputs to
- Assess the outputs; adjust and re-run



Correlation matrix

Here, ‘Token’ = Organic Carbon content



Pearson correlation coefficient **R**, **p**

p <= 0.05: p_stars = '*'

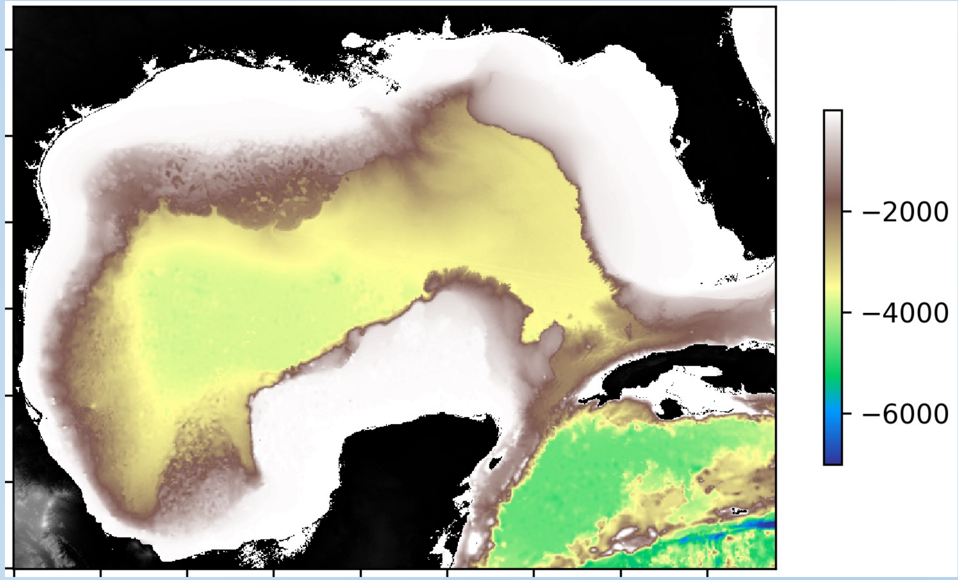
p <= 0.01: p_stars = '**'

p <= 0.001: p_stars = '***'

https://stackoverflow.com/questions/48139899/correlation-matrix-plot-with-coefficients-on-one-side-scatterplots-on-another

This data release from dbSEABED

Gulf of Mexico 0.02deg (~2km), but also **Global** (0.25 deg to match World Ocean Atlas)



Terrain – bathymetry, slope, aspect, geomorphic provinces, hessian largest eigenvalue (ridged-ness)

Geologic* – gravel content (% wt), sand content (% wt), mud content (% wt), clay content (% wt), central grainsize (phi), sorting (phi), rock presence (%), color (rgb), red/green (/1), grain component memberships (%), feature memberships (%)

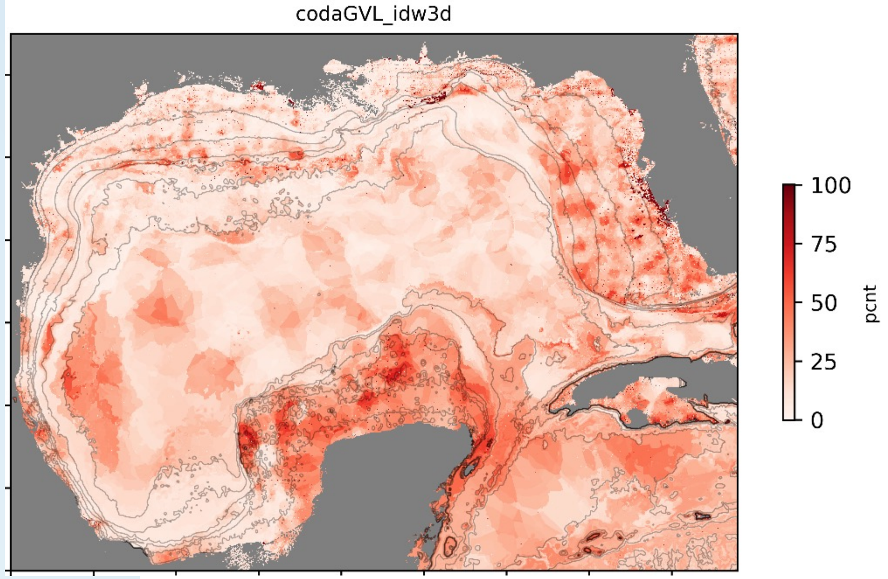
Compositional* – carbonate content (% wt, dry), organic carbon content (%wt, dry),

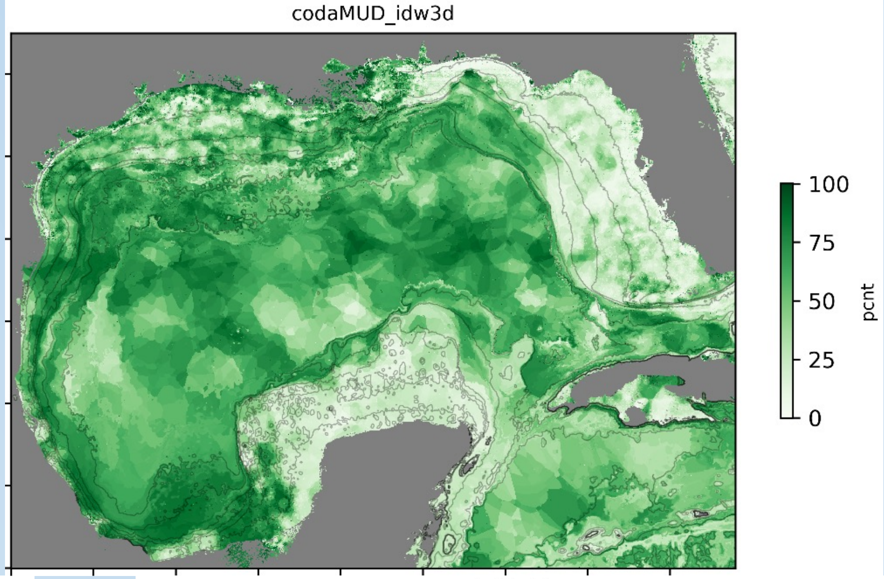
Geophysical* – porosity (%), critical shear stress (kPa), shear strength (kPa), sound velocity (m/s)

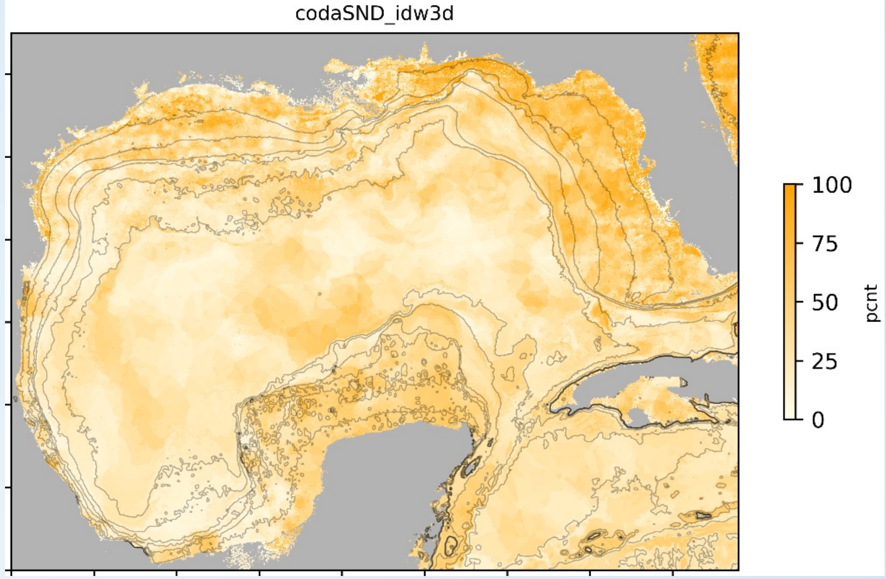
Environment – bottom & surface water temperatures, turbidity, surface chlorophyll-a, bottom dissolved oxygen (µmol/kg), bottom oxygen lows (µmol/kg), …

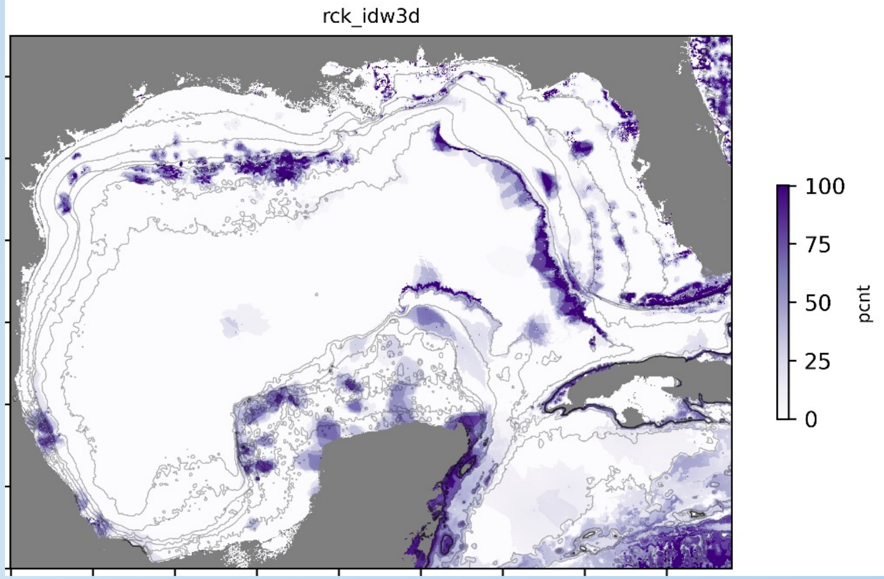
This data release from dbSEABED (cont)

Seabed textures



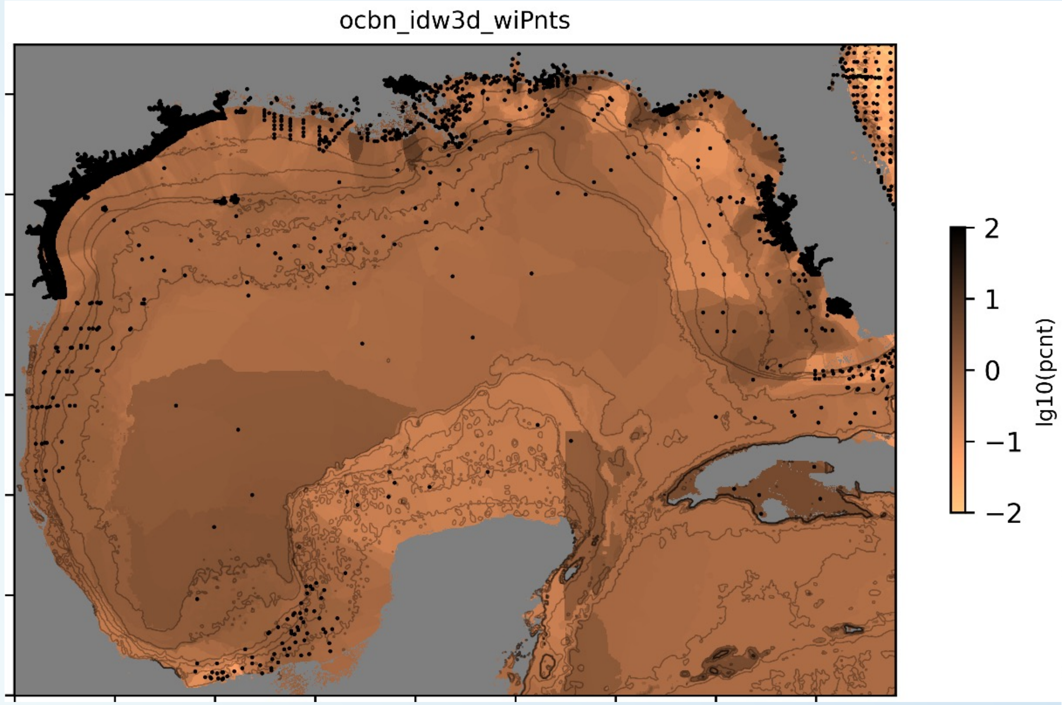


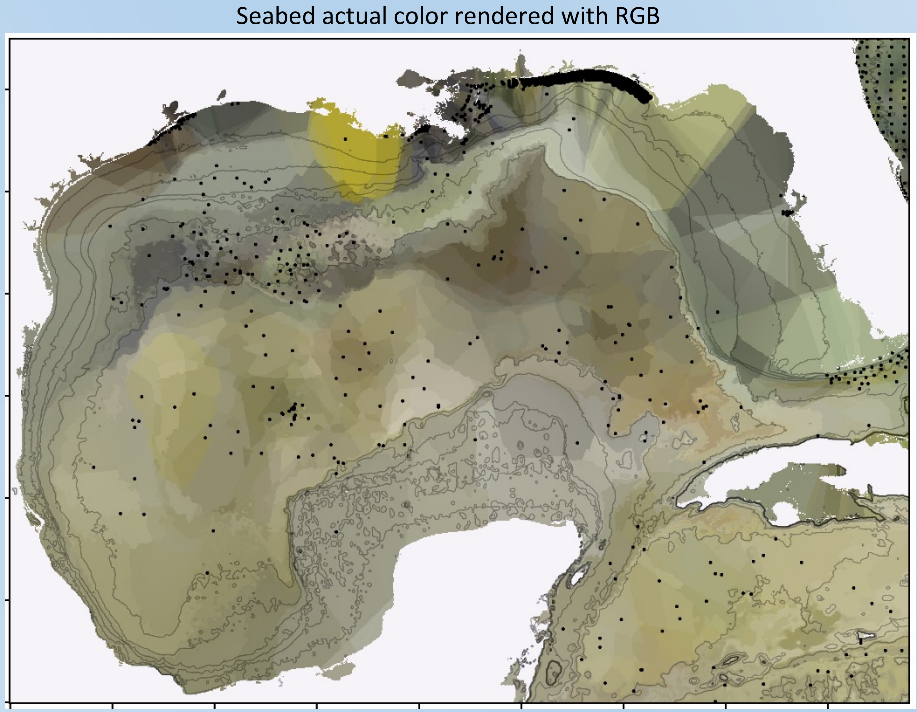


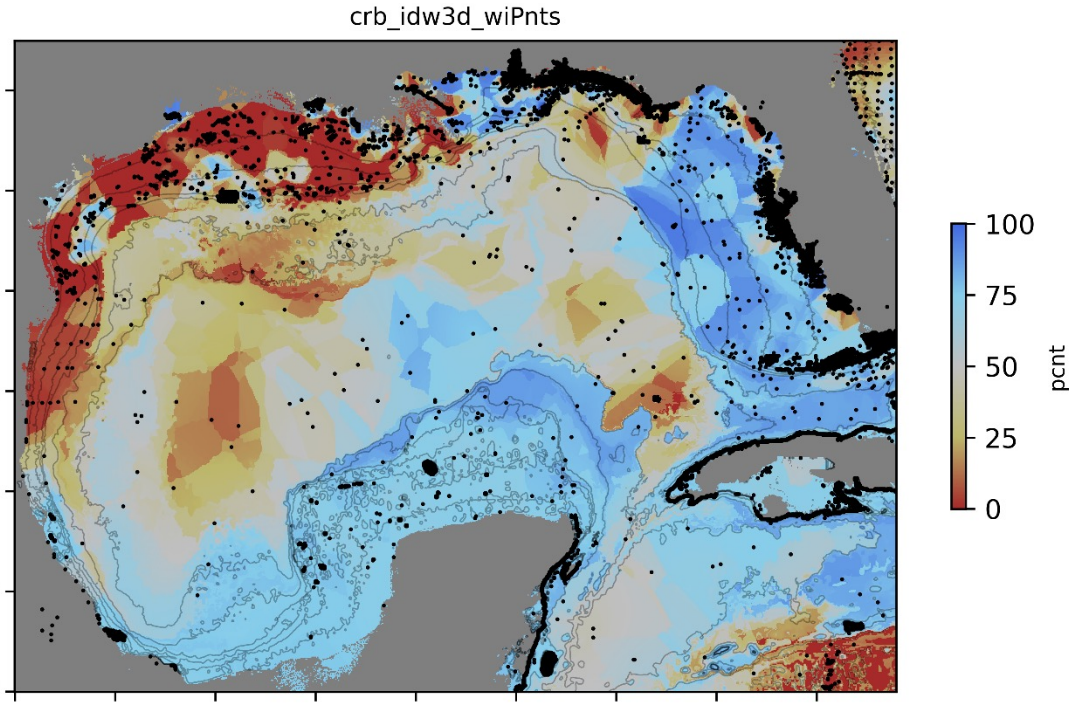


This data release from dbSEABED (cont)

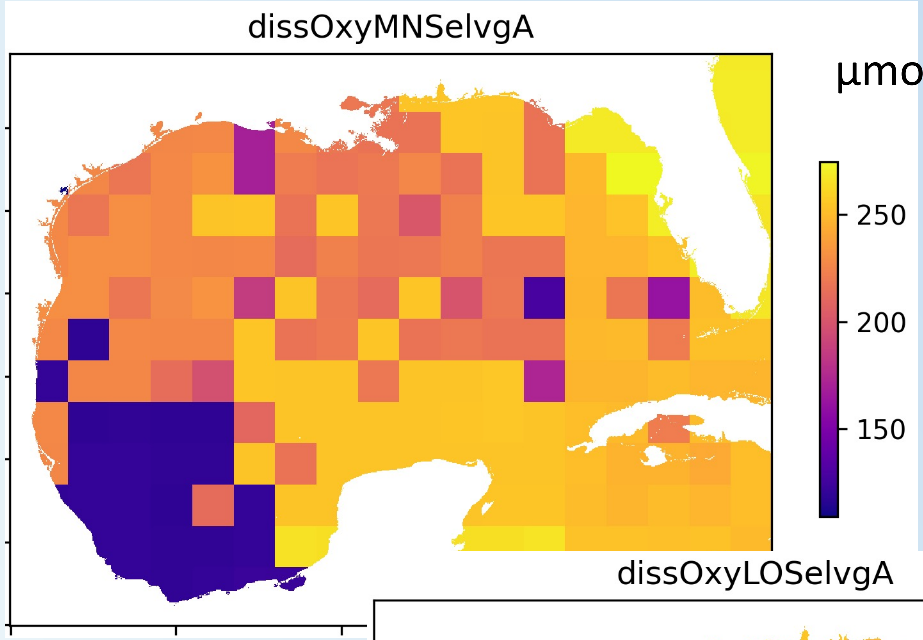
Compositions & Color

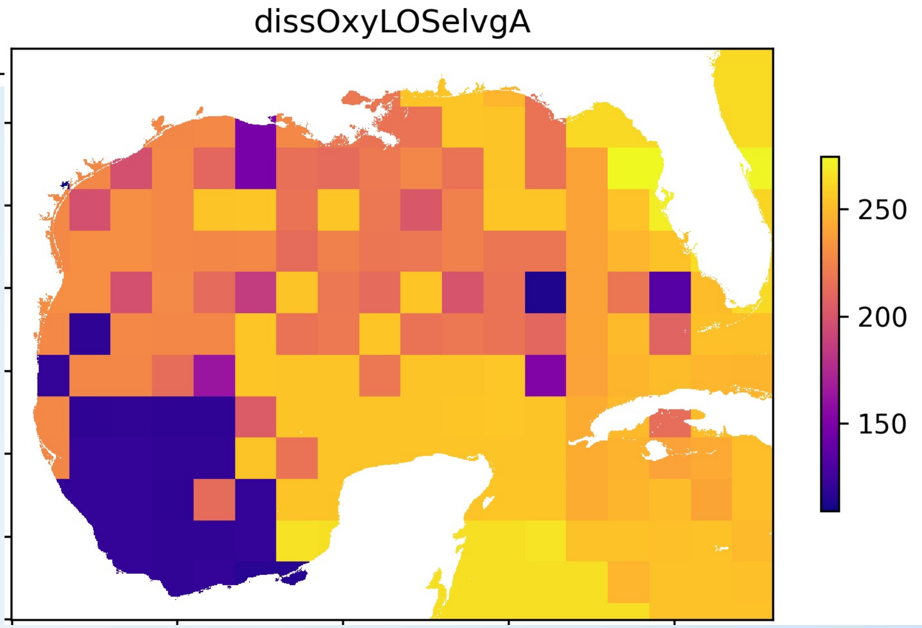






Modeling and Machine Learning





Dissolved Oxygen

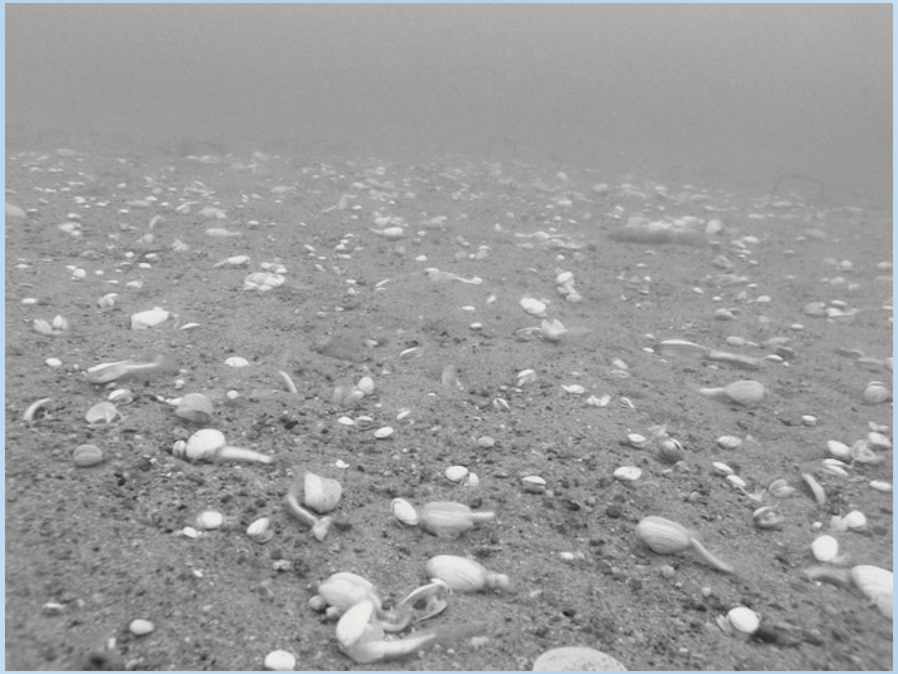
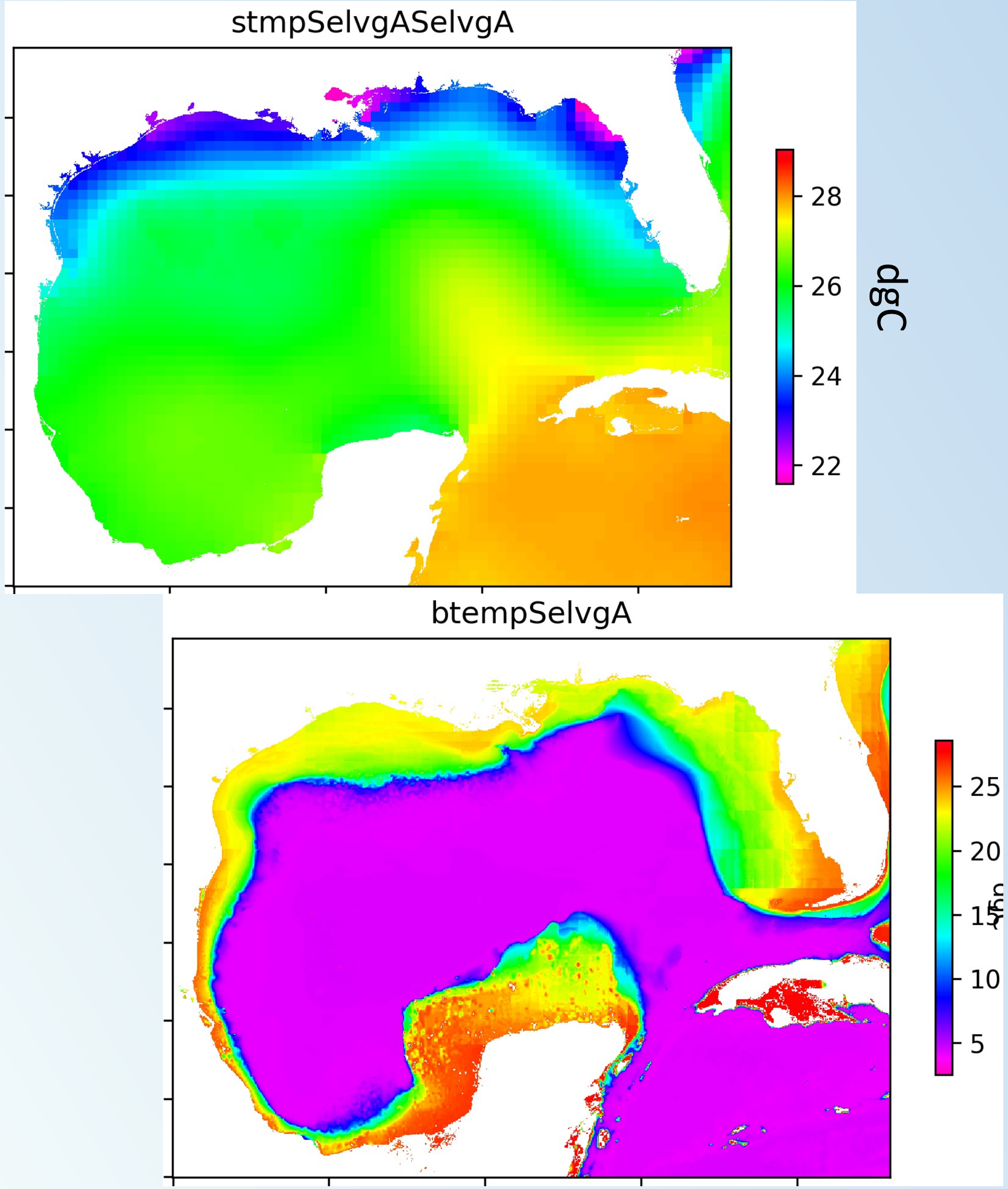


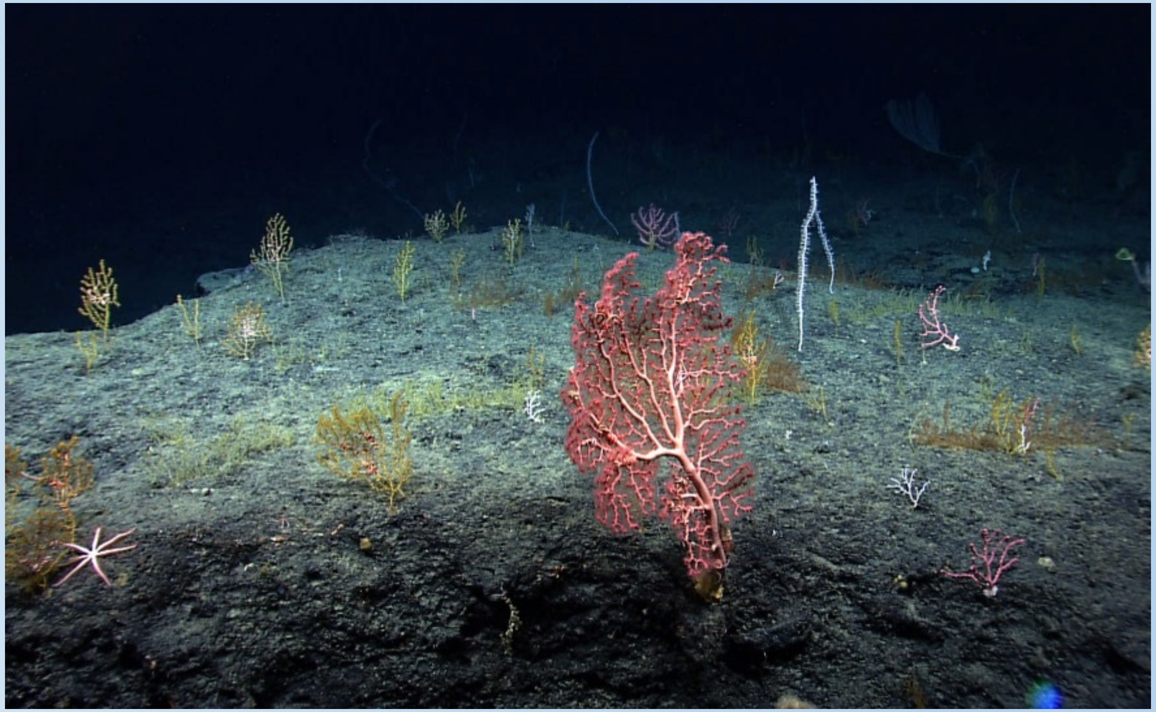
Figure 43. Dead bivalves at the seafloor after an oxygen starvation event, still with soft tissues visible. From: Norkko & others (2013). *Baltic Sea; width of view is about one metre.*

Source: World Ocean Atlas 2018
1dg averages, statistical mean & stddvn

https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl



Temperature (Surface & Bottom)

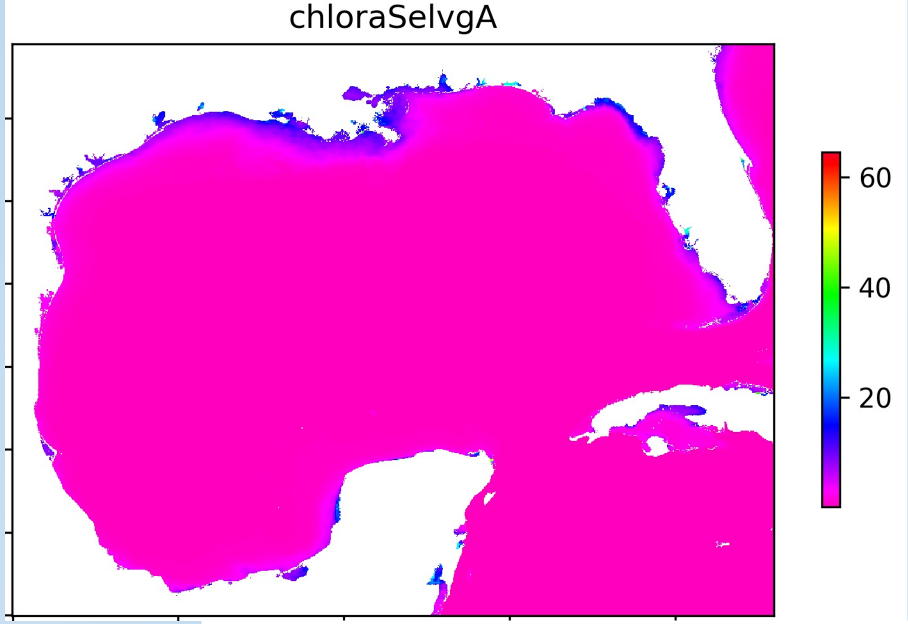


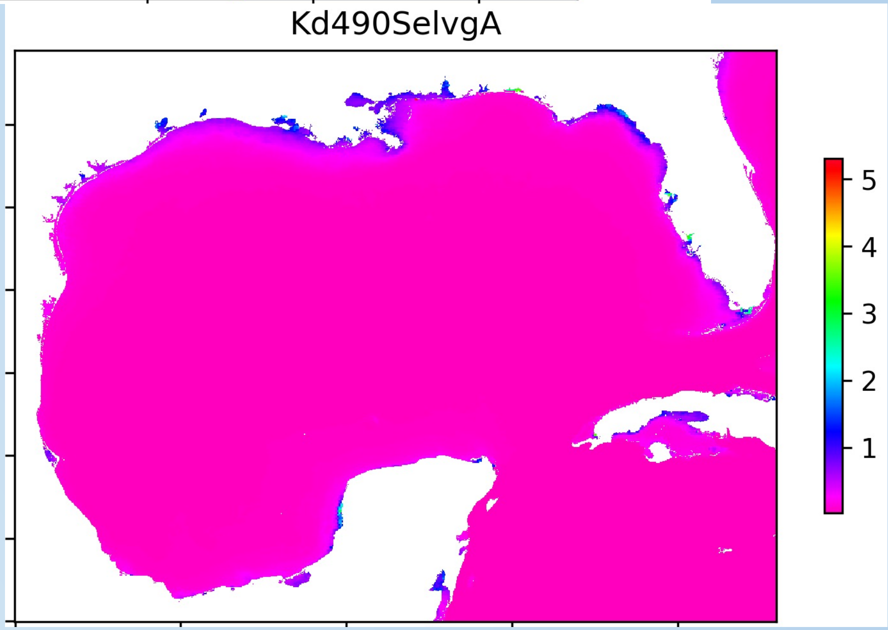
https://wusfnews.wusf.edu/environment/2020-10-23/deep-sea-gulf-corals-are-now-federally-protected

Modeling and Machine Learning (cont)

Turbidity, productivity

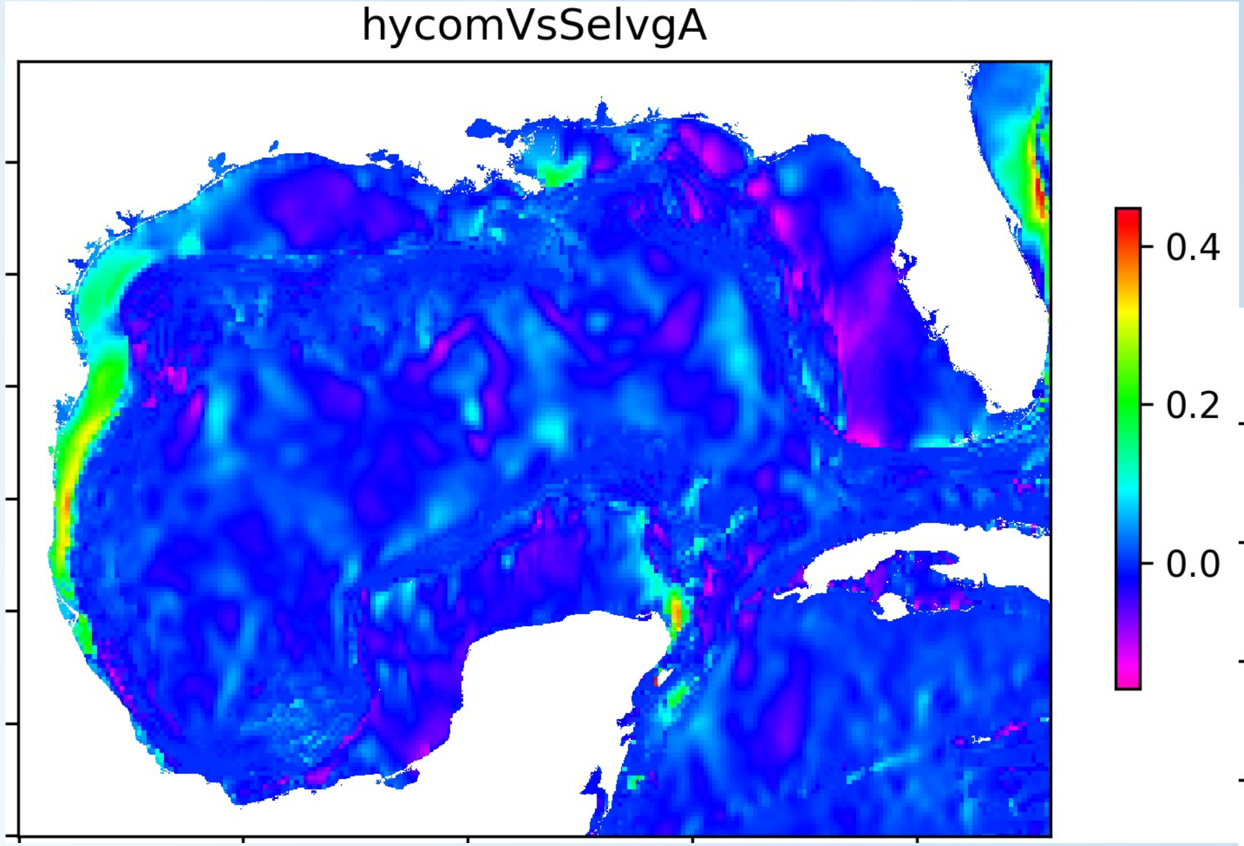
Source: MODIS Aqua
1dg averages, time averaged,
visible spectral proxies

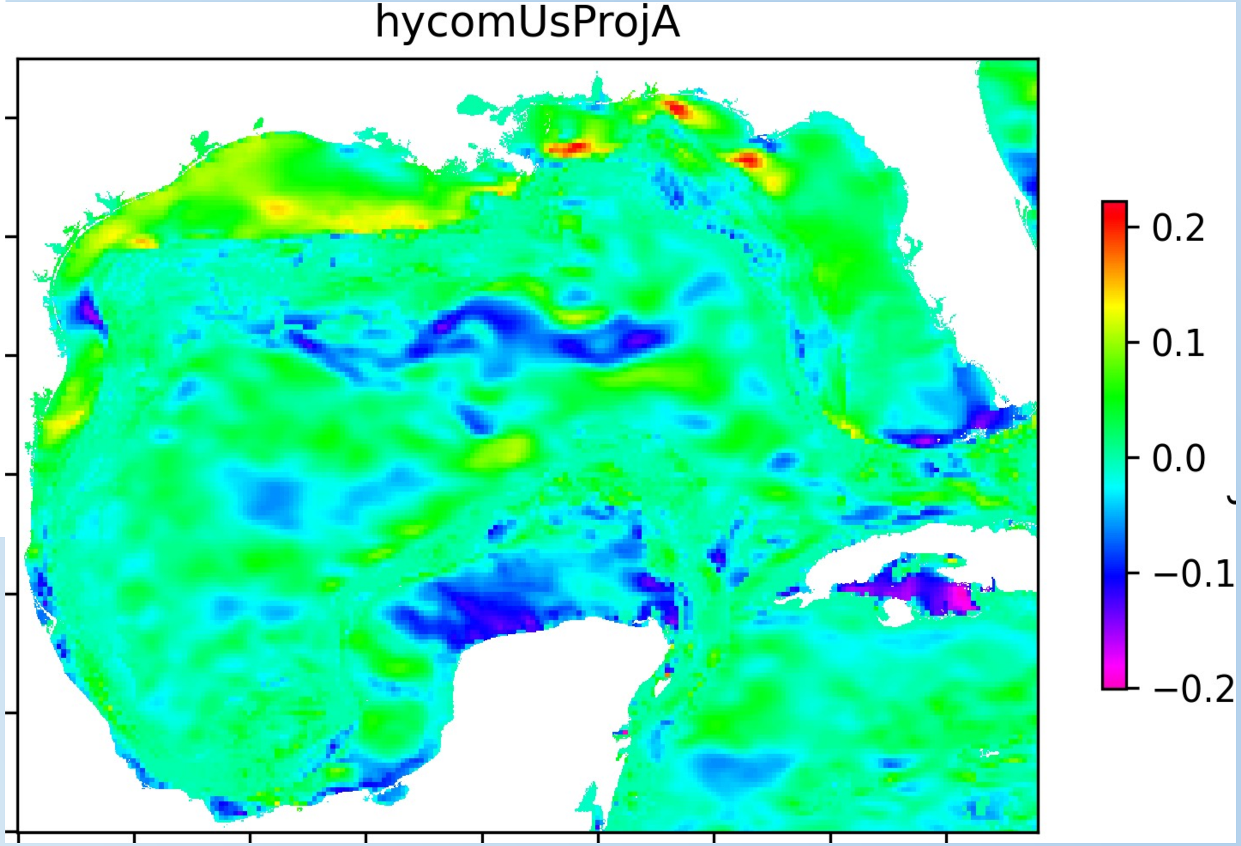




https://modis.gsfc.nasa.gov/data/dataprod/kd_490.php

Modeling and Machine Learning (cont)





https://www.hycom.org/dataserver
(short sample of)

Current flows

Exercises

For the clinic:

What seabed applications do YOU and your institution want / do you imagine for ML ?

- Habitat Suitability models
- Inter-parameter relations
- Substrate mappings
- … add more (with links ?)

- Get the data from “HERE” and enter it into your GIS (QGIS or ESRI)

Exercise (cont)

HSM – <u>Macoma balthica</u>

<u>Habitat suitability model</u>
- standard management tool
- here, physical parameters
- usually a logistic equation

What are the chief correlates for the living <u>M. balthica</u> ?
- dbSEABED _Enviro layers
- GBIF (OBIS) occurrence data

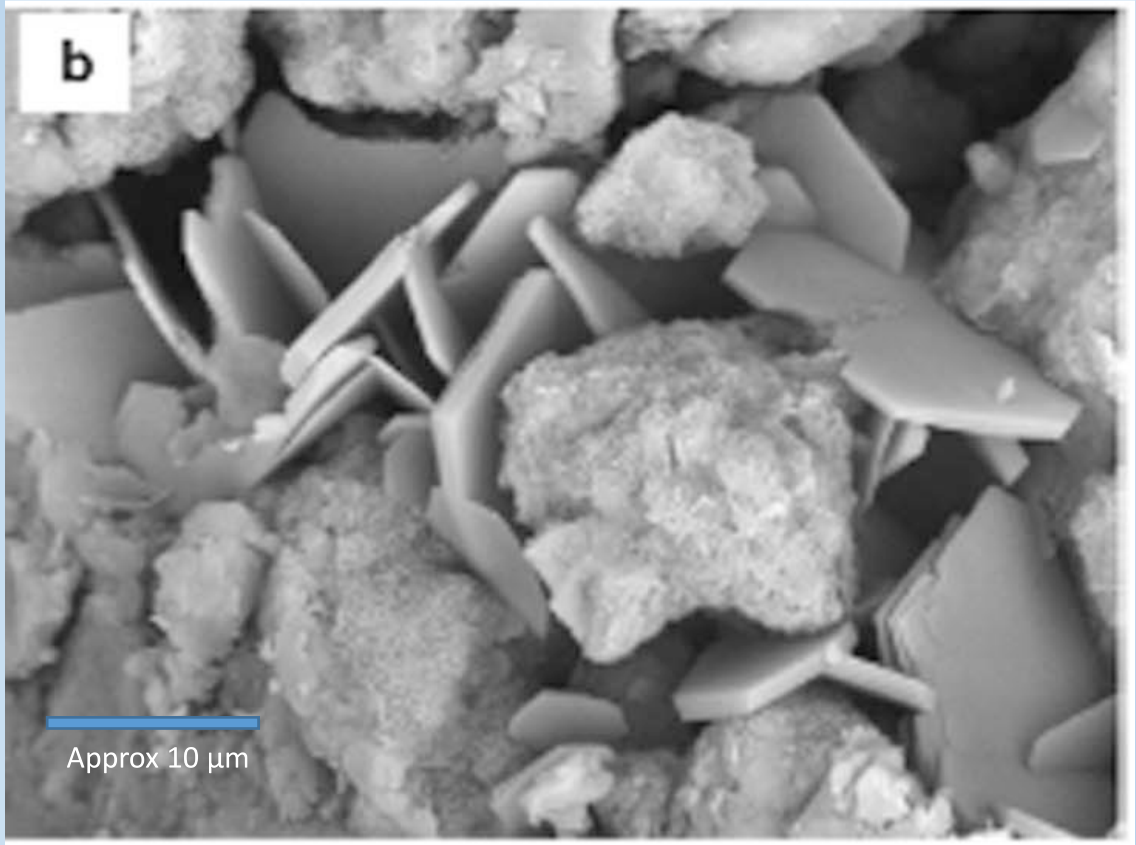
From FLIKR - https://www.flickr.com/photos/gridarendal/31636259221

Exercise (cont)

Clay/silt ratio
Important for seabed uptake and sequestration of radionucleides, e.g., from the Fukushima releases

Method:

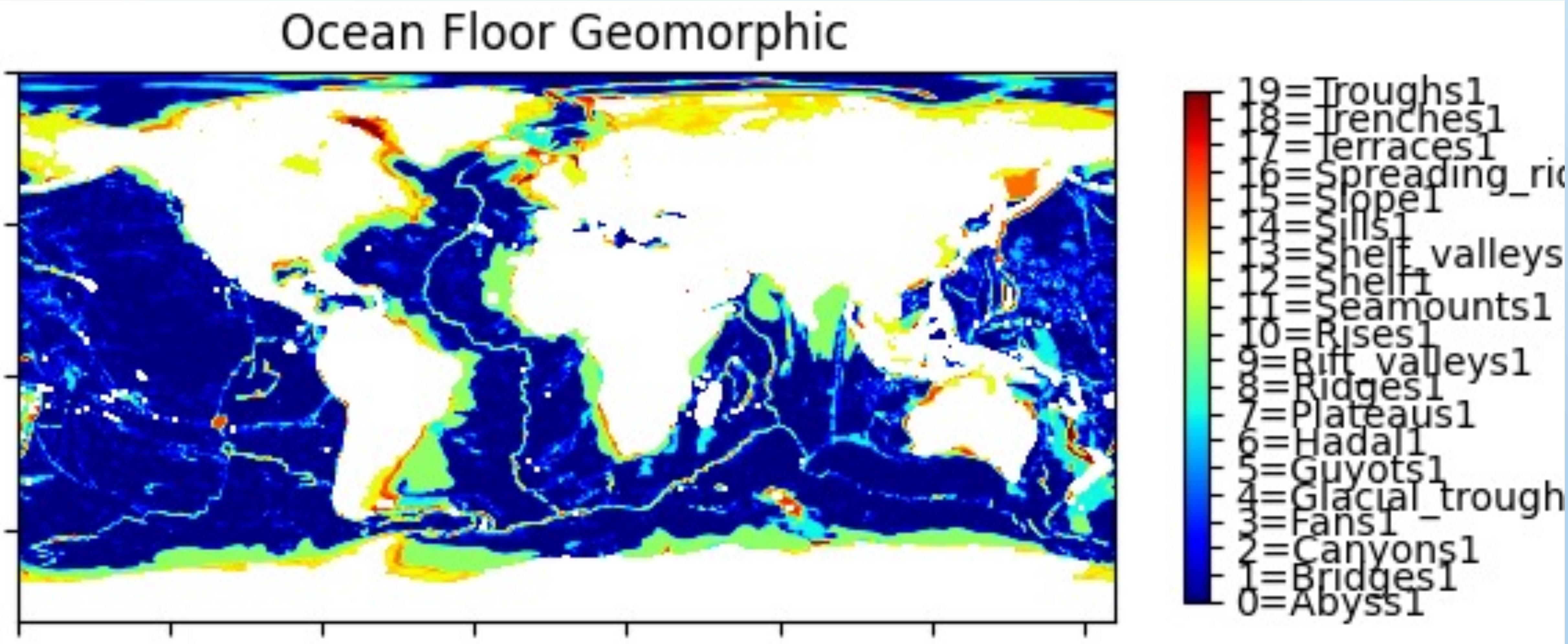
- Collect all conceivable parameter inputs, with a rationale for their use
- Collect all the silt/clay or clay/mud data from dbSEABED, separately for analysis and description data
- Associate the silt/clay & parameter locations and water depths.
- Test the correlations



From de Gruyter Open Geosciences - https://doi.org/10.1515/geo-2020-0145

Wrap-up

- What do researchers want in global gridded data-layers ?
- What are the questions they want to answer ?
- How will we obtain/build those layers ?
- What exactly is the parameter / statistic that we want in each case ?
- How important is having a process-model rationale for each parameter ?



Harris, P. T., MacMillan-Lawler, M., Rupp, J., and Baker, E. K. (2014). Geomorphology of the oceans. *Mar. Geol.* 352, 4–24. doi: 10.1016/j.margeo.2014.01.011

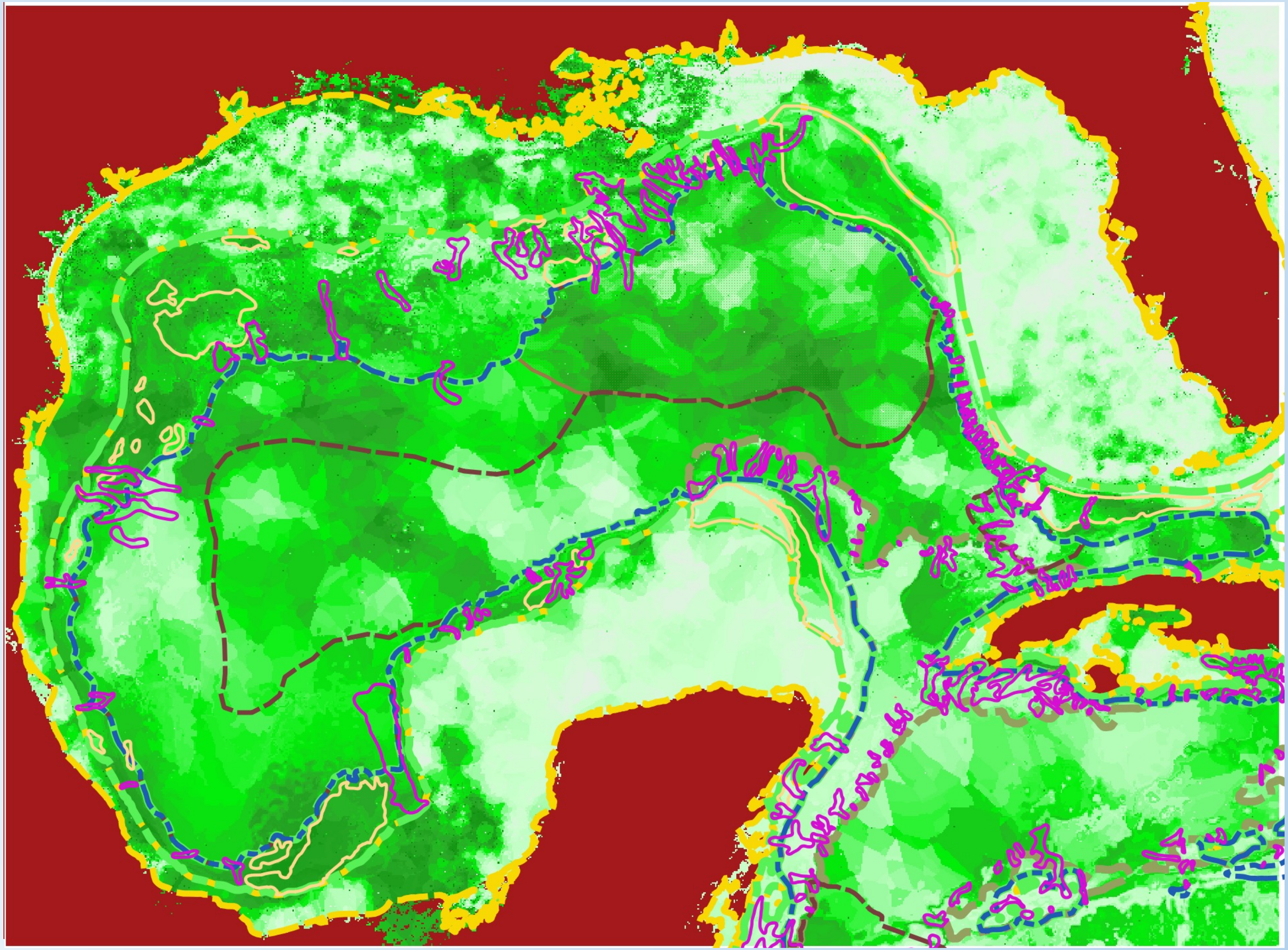
Wrap-up (cont)

dbSEABED Mud contents (%)

	0.000000
	20.000000
	40.000000
	60.000000
	80.000000
	100.000000

HYCOM Bottom Currents (Vs, m/s)

	-0.5000
	-0.2500
	0.0000
	0.2500
	0.5000



References

Dutkiewicz, A., Müller, R. D., O'Callaghan, S., & Jónasson, H. (2015). Census of seafloor sediments in the world's ocean. *Geology*, G36883-1. doi: 10.1130/G36883.1.

Jenkins, C.J. 2002. Automated digital mapping of sediment colour descriptions. *Geo-Marine Letters*, 22(4), 181-7.

Jenkins, C.J. 1997. Building Offshore Soils Databases. *Sea Technology*, 38(12), pp. 25-28. [Article draft: "http://instaar.colorado.edu/%7Ejenkinsc/dbseabed/seatchfz.pdf"]

Levitus, S., US DOC/NOAA/NESDIS - National Oceanographic Data Center 2013. *NOAA NODC Standard Product: World Ocean Atlas 2009* (NOAA NCEI Accession 0094866).

Lee, T. R., Wood, W. T., & Phrampus, B. J. 2019. A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Global Biogeochemical Cycles*, 33, 37–46. [URL: https://doi.org/10.1029/2018GB005992]

Norkko, A., Villnäs, A., Norkko, J., Valanko, S. & Pilditch, C. 2013. Size matters: implications of the loss of large individuals for ecosystem function. *Scientific Reports 3(*2646) [DOI "doi:10.1038/srep02646"]

Restreppo, G.A., Wood, W.T. & Phrampus, B.J. 2020. Oceanic sediment accumulation rates predicted via machine learning algorithm: towards sediment characterization on a global scale. Geo-Mar Lett 40, 755–763 (2020). https://doi.org/10.1007/s00367-020-00669-1

Extra for questions …

Target Feature	Predictive Features	
Carbonate	lgwd, prox, btemp, stemp, kd490, chlora	
Rock exposure	lgwd, slp, hsn	
Mud content	lgwd, slp, kd490, chlora	
Sand content	lgwd, slp, prx	
Gravel content	lgwd, slp, prox	
Organic carbon	lgwd, btemp, stemp, kd490, chlora, doxMN, doxLO	
Porosity	lgwd, btemp, stemp, kd490	

The End