How accessible is machine learning for non-computer scientists? Application of deep learning in physical oceanography and deep-sea ecology

VIMS | William & Mary
Virginia Institute of Marine Science
Laboratory for Ocean Sensing

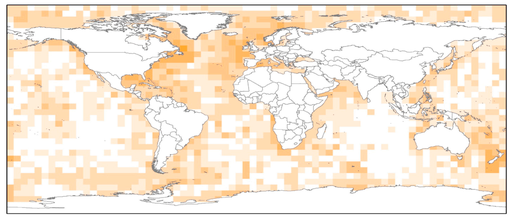
Donglai Gong (*gong@vims.edu*), Jeanna Hudson (*jeannak@vims.edu*)

ED53E-0761

Introduction

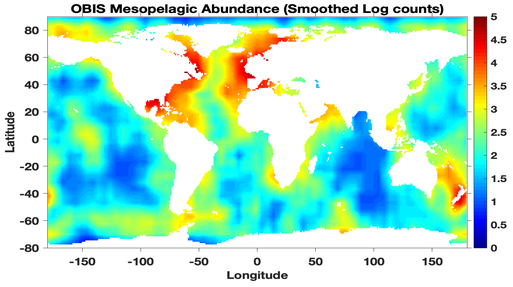
Estimates of global mesopelagic faunal abundance and biomass are critical components of biogeochemical research of the ocean. Many mesopelagic fauna are diel vertical migrators, consuming zooplankton and fish in the surface 200m and respiring, excreting, and egesting at depths of 200-1000m. This daily migration plays an important role in the carbon cycle and the transportation of nutrients from productive surface waters to the food limited deep ocean. The global distribution of mesopelagic faunal abundance and biomass are based on a very limited number of studies (i.e. Gjøsaeter and Kawaguchi (1980), Proud et al. (2017), Sutton et al. (2017)). We used 'off the shelf' supervised learning algorithms to model the observed mesopelagic abundance using physical and biogeochemical data from the World Ocean Atlas. We also applied the same methodology to model biomass distributions on the MAB continental shelf using the MOCHA T/S climatology and fisheries catch data from NOAA NEFSC. The Matlab-based ML tools was easy to use, however the processing of the data into the right format was the most time consuming step.

Distribution of Mesopelagic Fauna



Sutton et al. (2017)

Ocean Biogeographic Information System (OBIS) records of mesopelagic fauna from Sutton et al. (2017) were used as input data for the analysis.



Data from Sutton et al. (2017), log transformed and smoothed over 15 degrees, are used as training and validation data for building supervised machine learning models.

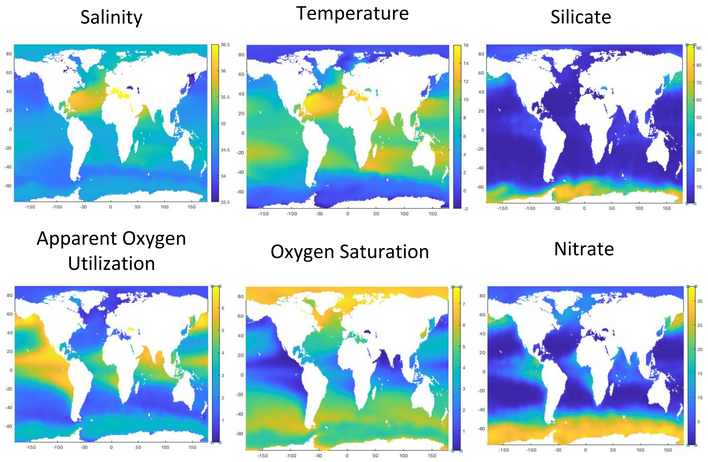
Questions & Hypotheses

- Can we predict mesopelagic faunal abundance based on known ocean physical and biogeochemical parameters?
- Areas of high mesopelagic abundance correlate with specific physical and biogeochemical conditions.
- The pattern of abundance distribution is sensitive to climate change. Trained machine learning models may be useful for climate forcing sensitivity studies in the mesopelagic.

Supervised Machine Learning

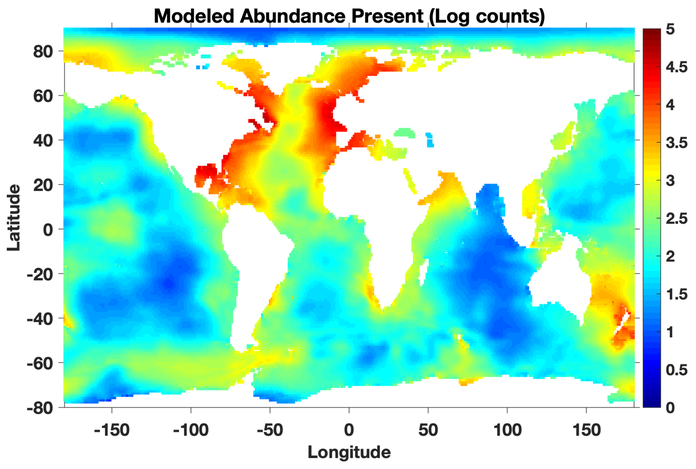
- Matlab's Regression Learner App was used to conduct supervised machine learning.
- Support Vector Machine (SVM) and Regression Tree models were used for learning.
- Physical and biogeochemical properties such as temperature, salinity, percent oxygen saturation, apparent oxygen utilization (AOU), nitrate, and silicate, as well as latitude and longitude, were used as model input features.
- Root mean square error was used to assess model skill.
- SVM with Gaussian Radial Basis Function (RBF) gives the best compromise between goodness of fit and sensible prediction.

Global Physical & Biogeochemical Data (Input Features)

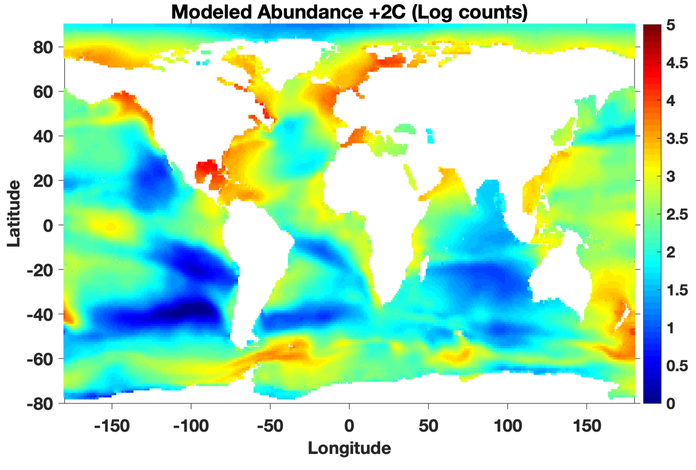


Data source World Ocean Atlas

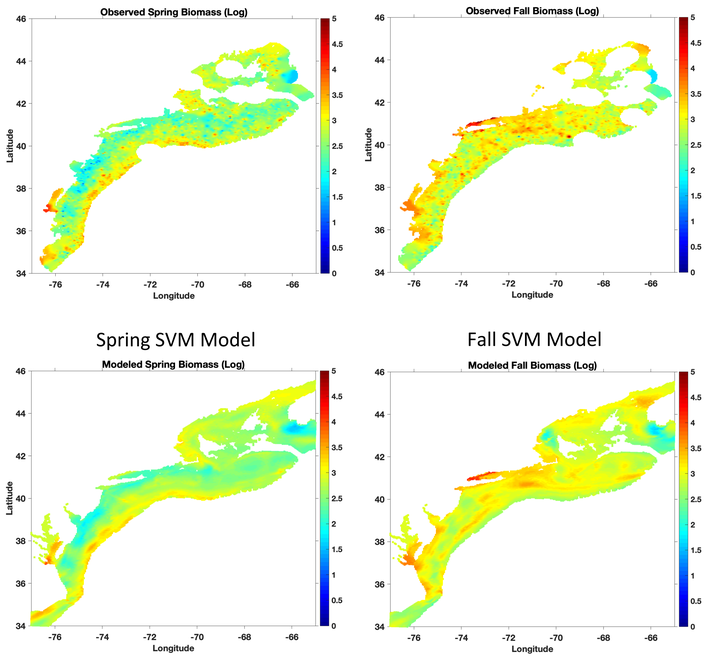
Global Model-Predicted Mesopelagic Abundance



Mesopelagic faunal abundance was predicted using a SVM Gaussian RBF model with a kernel scale of 1 (figure above). All physical and biogeochemical parameters were used, as well as latitude and longitude. Randomly selected, 50% of the data were used for training and 50% of the data were used for model validation. The root mean square error was 0.11. The SVM model was then used to predict the mesopelagic distribution for a hypothetical scenario of two degree increase in global ocean temperature (figure below).



Mid-Atlantic Actual & Model-Predicted Biomass



Preliminary Findings of Regional Analysis in the MAB

In the Mid-Atlantic Bight, the MOCHA climatology contains only temperature and salinity, therefore the input parameters for the ML algorithm was limited to physical parameters only. The resulting RMS error was around 0.2, higher than the global case when biogeochemical data are also used. Furthermore, the MAB fisheries catch data has much higher spatial patchiness than the 5 degree averaged global mesopelagic faunal abundance data. ML was able to capture most of larger scale variability.

The models tested include Regression Tree and SVM. The 'best' model was the SVM model with a Gaussian kernel. It is also one of the fastest models to train with an average training time of less than 1 min per run. There are three hyperparameters in the model, two of which are auto tuned, the only hyperparameter that was set manually was the 'kernel scale' which was set to 0.2 through trial and error. It's a good compromise between smoothing out data patchiness while capturing some of the mesoscale variabilities.

Preliminary Findings of Global Analysis

We found the SVM model with Gaussian RBF to strike a good balance between model skill and avoiding over fitting. While the Regression Tree method produced consistently low root mean square error scores with the training data, and the predicted global abundance distribution very closely resembled our initial reference data, it was prone to over fitting and was unstable to small variations introduced to the physical forcing data. This produced unreasonable patterns for different forcing scenarios (natural variability or human-induced change).

The full SVM model, trained using latitude, longitude, salinity, temperature, silicate, nitrate, oxygen saturation, and apparent oxygen utilization as input features, appears to be the best model for projecting future scenarios of faunal abundance and distribution. However, if the goal is to produce the best interpolation of the existing mesopelagic dataset then a Regression Tree model using only latitude and longitude produces good results.

References

Tracey T. Sutton, Malcolm R. Clark, Daniel C. Dunn, Patrick N. Halpin, Alex D. Rogers, John Guinotte, Steven J. Bograd, Martin V. Angel, Jose Angel A. Perez, Karen Wishner, Richard L. Haedrich, Dhugal J. Lindsay, Jeffrey C. Drazen, Alexander Vereshchaka, Uwe Piatkowski, Telmo Morato, Katarzyna Błachowiak-Samołyk, Bruce H. Robison, Kristina M. Gjerde, Annelies Pierrot-Bults, Patricio Bernal, Gabriel Reygondeau, Mikko Heino (2017) A global biogeographic classification of the mesopelagic zone. Deep-sea Research Part 1 126, 85-102.

Roland Proud, Martin Cox ,Andrew Brierley (2017) Biogeography of the global ocean mesopelagic zone. Current Biology 27, 113-119.

Gjøsaeter, J., Kawaguchi K., 1980. A review of the world resources of mesopelagic fish. FAO Fisheries Technical Papers 193, FAO, Rome.

World Ocean Atlas- https://www.nodc.noaa.gov/OC5/indprod.html

What we learned as non-computer scientists

- There is a steep learning curve for those without a strong programming and calculus background when using machine learning (ML). Much of the effort is focused on the front end getting the data formatted correctly in order to feed it into the ML framework.
- Data manipulation using a computing language is necessary (Matlab, R, Python, etc.).
- An understanding is needed of the various ML approaches (support vector machine, regression trees, Gaussian regression, etc.) as well as the frameworks to conduct the analysis (Matlab Regression Learner app, TensorFlow, Numpy, etc.).
- Some frameworks require a more in depth understanding of calculus than others. The Matlab Regression Learner app requires less of an understanding of the nuances of ML but interpretation of the results can be tricky (i.e. over fitting).
- Application of the most cutting edge ML model such as deep neural net would require significant ML training and computing resources.
- Familiarity with global climate models and datasets is useful as a source of comparative environmental physical and biogeochemical data.

Discussion

Due to the localized nature of oceanic sampling in productive regions, the training data contains bias which influences the models ability to identify true patterns in faunal distribution. Based on the OBIS dataset provided to the model for training, however, our model was able to accurately predict regional patterns of abundance. The modeled abundance map is not a true representation of mesopelagic abundance because it does not take into account non-uniform fishing efforts around the globe. There could be other factors, beyond the features we selected, that may significantly impact the distribution (e.g. ocean turbulence, fronts, etc.).

Given the upward trend in our ocean's temperature, it is not unrealistic to expect to see a two degree increase. Our model portrays one possible scenario as a response to such an increase in temperature. Abundance could increase in the Pacific, Arctic, and Southern Oceans and decrease in the north Atlantic.

Supervised Machine Learning is a useful tool for characterizing distribution patterns in the poorly sampled mesopelagic ocean; however, in order to properly validate the model, additional data is needed.

Future Work

- Our first endeavor into ML focused broadly on the global distribution of marine fauna using Matlab's Regression Learner app. We followed that with a more focused examination of both region and species using the same app. Next, we plan to investigate other ML frameworks like PyTorch or Flux to expand our predictive capabilities.
- We may further investigate the regional timeseries data to track changes in faunal abundance and biomass from the 1960s to the present as it relates to additional physical and biogeochemical properties.
- For the regional study in the MAB, we plan to include climatology of biogeochemical data to try to improve the model and reduce errors.
- Compare ML models of distribution with other statistical models such as GAM to evaluate their relative strengths and weaknesses.
- Document and share our experience with other interested people.

Acknowledgement

We would like to Michael Vecchione for project advice and guidance.

We also would like to thank NOAA NEFSC and John Manderson for providing us the fish catch data for the Mid-Atlantic Bight.

We like to thank Andrew Ng at Coursera for providing the online machine learning lectures that were invaluable for helping us to gain the necessary understanding of machine learning.