Gap Filling of Surface Ocean pCO2, with Uncertainty

Data Science Institute
COLUMBIA UNIVERSITY

Xiaoqi Wang, Bin Xu, Nan You, Tianyi Zhang, Yisu Zhong
Faculty Mentors: Galen McKinley, Lucas Gloege

Data Science Capstone Project with Prof. McKinley, Department of Earth and Environmental Sciences

DATA VISUALIZATION

Our project presents a new machine-learning-based method to estimate surface ocean pCO_2 and evaluate the uncertainty of estimates. Given the fact that only small percentage(around 1.56%) of our observed data owns pCO_2 value, we decide to make use of the simulated data(left) to help us train and test our model. The full-filed "Truth" could also help us evaluate our method performance. From the following two pictures, we could see the simulated data is a good representation of observed data.

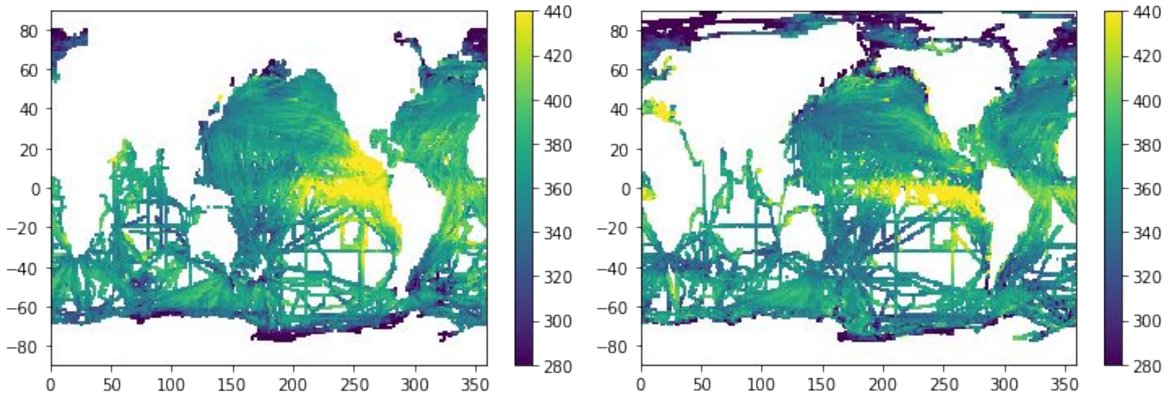

Figure 1. Color plot of pCO2 from simulated and observed Dataset

KRIGING SYSTEM

The first conditional distribution of pCO_2 given spatial data is estimated by nearby data points. This leverages the kriging mean and variance of simple kriging system using nearby observations to parametrize the conditional Gaussian distribution.

We use GeoStatTools package in Python to conduct the interpolation, which result becomes our estimated mean, and the errors become our variances, which will be used to parametrize the conditional Gaussian distribution.

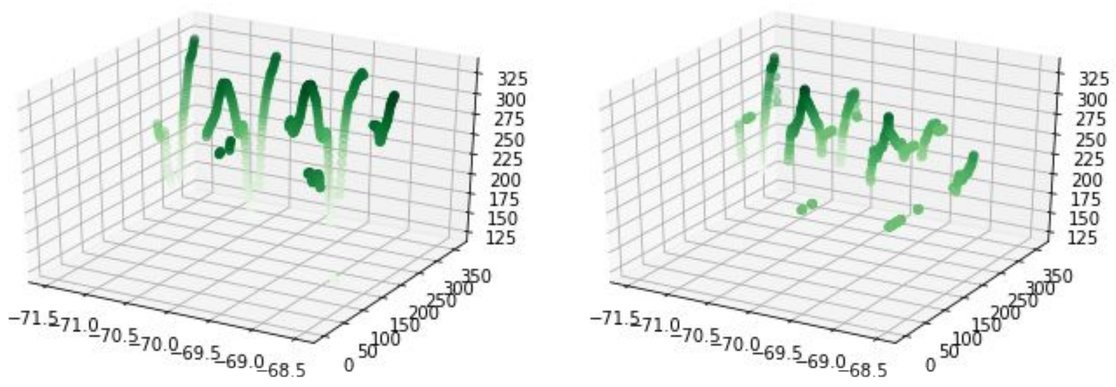

Figure 2. True distributions and predicted distributions by kriging system.

GMM-EM ALGORITHM

The second conditional distribution of pCO_2 given the colocated observed feature is estimated by taking the advantage of Expectation Maximization (EM) algorithm and Gaussian Mixture Model (GMM). To be specific, it is defined as the marginal of the conditional GMM given other features observed at same data location. Therefore, to obtain this conditional distribution, we first implemented GMM to get the soft clustering boundary for simulated data with GMM, then imputed the missing value of pCO_2 conditioned on other observed features by using EM algorithm and clustering statistic generated from GMM.

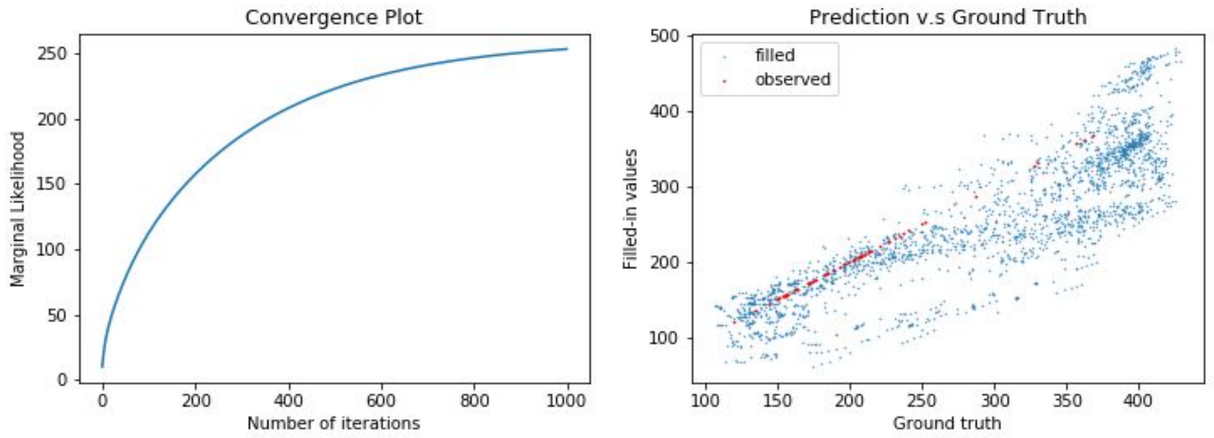

Figure 3. A Demo: GMM-EM for filling missing values in a cluster

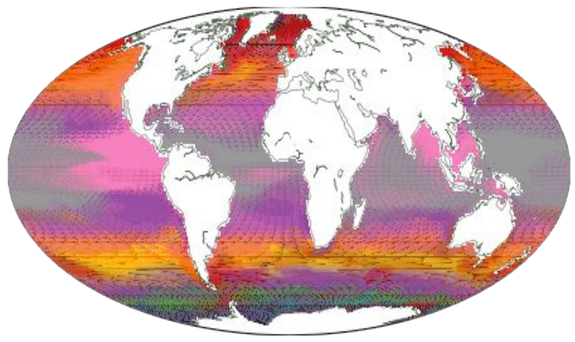

Figure 4. GMM Clustering Result

As shown in Figure 4, the data points which are closer to each other spatially are more likely to be assigned to the same cluster, even though the spatial information of data points is not used during the training of GMM. Therefore, the clustering result follows our initial assumption about data and GMM and is trustworthy for our cases.

Conclusions and Recommendations

The ultimate goal for our project is to predict oceanic pCO_2 level for unknown regions and get statistical inference for our prediction. The final step is to apply the algorithms to the whole dataset and to evaluate their performance in different clusters.

Acknowledgments

We thank Professor Galen McKinley and Luke Gloege for their expertise support.

References

Silva, D., et al. "Multivariate Data Imputation Using Gaussian Mixture Models."

Mckinley, G. A., et al. "Natural Variability and Anthropogenic Trends in the Ocean Carbon Sink."