Ocean Carbon States based on k-means cluster analysis

Rebecca Latto, Anastasia Romanou
Applied Physics and Applied Math Columbia University and NASA GISS

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Goddard Institute for Space Studies

Introduction

- Advanced data mining techniques are becoming widely used in Climate and Earth Science with the purpose of extracting new meaningful information from large and complex datasets.
- In studies of the global carbon cycle, lack of understanding of the interacting physical and biogeochemical drivers confounds our ability to accurately describe, understand, and predict CO2 concentrations and their changes in the major planetary carbon reservoirs
- We employ cluster analysis as a means of identifying and comparing spatial and temporal patterns[1] of pCO2 (Landschuetzer product) and temperature at 10m (ARGO Coriolis product) for 2000-2015
- We assess how researchers could potentially use this exploratory analysis tool to better understand complex systems by correlating the interannual and spatial variability of relevant climate indices (ENSO, AO, NAO, etc.) and other physical fields like salinity and chlorophyll with cluster variability

Multivariate Analysis

Cluster Analysis: The k-means algorithm interprets a dynamical Earth system as a geophysical, climate network, with spatial nodes that are connected by a time series.

Choosing Optimal Variables for Clustering

- The air-sea exchange of CO_2 defines 2 main pathways that determine the ability of the ocean to uptake CO_2:
  1. The chemical disequilibrium expressed by pCO_2 of surface water, dissolved inorganic carbon, and nutrients in biogeochemical processes
  2. Physical processes (e.g. air-sea interaction and ocean circulation)
- The physical variables partial pressure of CO_2 (pCO_2) [Landschuetzer SOCAT product] and sea surface temperature (SST) [ARGO T profile at 10m] are selected because their joint parameter space can be used to understand CO_2 flux distributions and variability for 2000-2015

Determining optimal number of clusters

- Checklist:
  - ✓ There cannot be degenerate clusters
  - ✓ The number of clusters needs to make physical sense for the given system
  - ✓ No new, *significant* information can be gained by adding a new cluster

Methodology

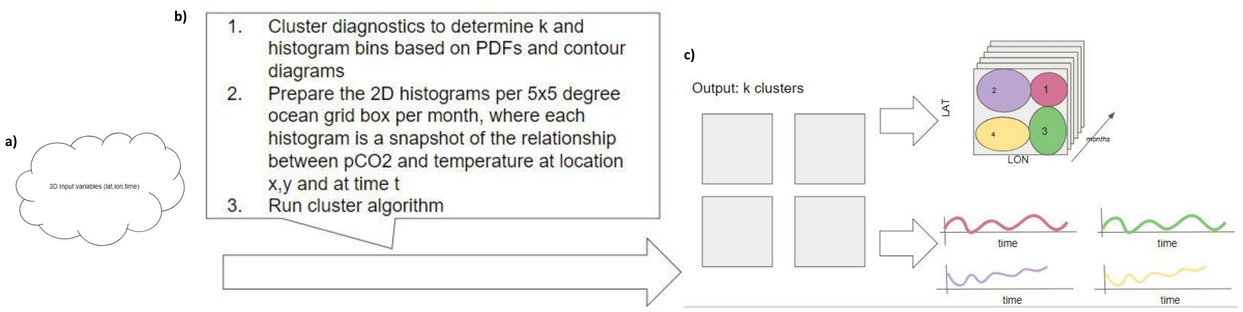


Figure 1: a) Prepare input variables on same grid, b) Run pre-clustering analysis, c) Investigate cluster output by analyzing spatial and temporal attributions.

Ocean Carbon States

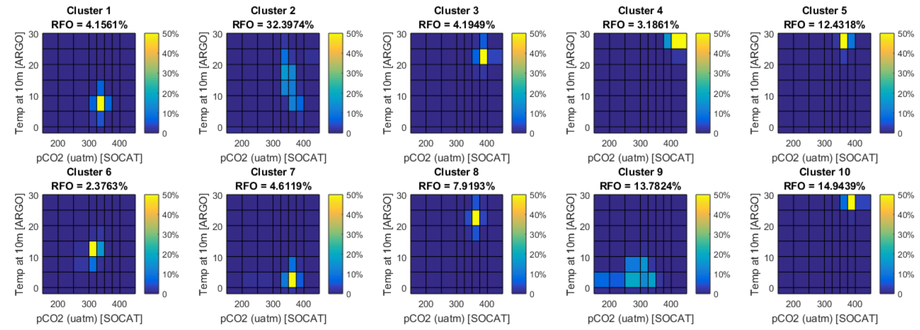


Figure 2: For k = 10, each cluster histogram is the average distribution of the histograms assigned to each cluster. The color bar represents the abundance of the variable-pair relationships between pCO2 and SST within each cluster and the relative frequency of occurrence (RFO) denotes how many histograms out of the total number of histograms are represented by each cluster.

Preliminary Post-Clustering Analysis for Cluster 1

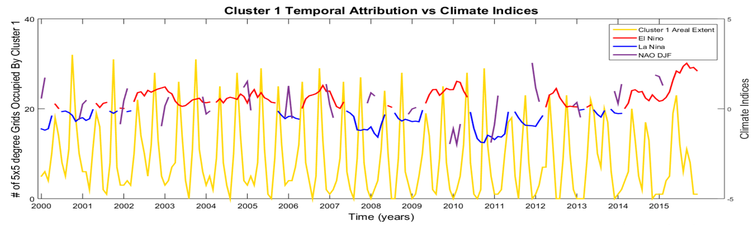


Figure 3: Temporal Attribution time series plotted with El Nino, La Nina, and North Atlantic Oscillation (DJF) variation to demonstrate possible relationship.

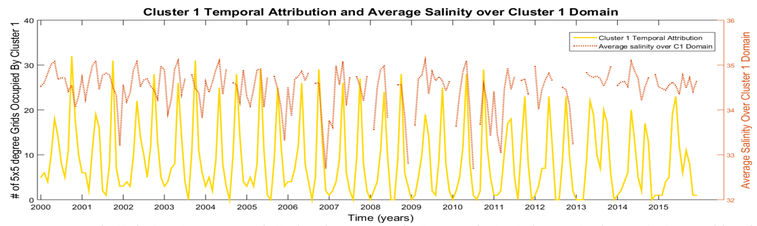
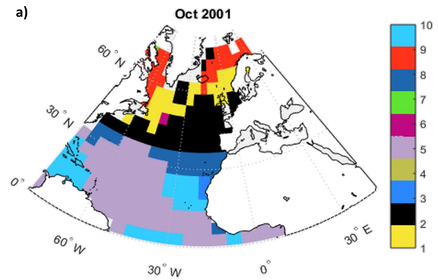


Figure 4: Temporal Attribution time series plotted with average salinity over cluster 1 domain to demonstrate possible relationship.

Cluster parameterization: defined as between 25th-75th percentile

• SST : 5 - 10°C	SST over N. Atlantic: 10 - 25°C
• pCO2: 325 – 350 uatm	pCO2: 290 – 425 uatm
• Flux: -1 - -4 mol/m² (outgassing)	Flux: -5 – 5 mol/m²
• Salinity: 34.2 – 35.7	Salinity: 35 – 36.5
• Chlorophyll: 0.223 – 1.65 mg/m³	Chlorophyll: 0.105 – 0.47 mg/m³



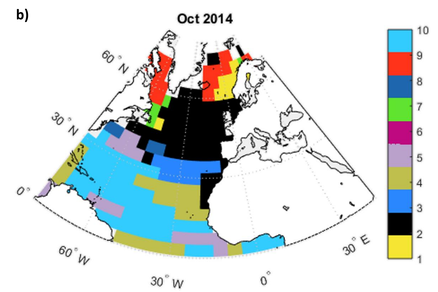


Figure 5: a) Spatial assignments of clusters 1-10 in October 2001. b) Spatial assignments of clusters 1-10 in October 2014

Discussion and Future Work

- Post-clustering analysis needs to be further explored for each cluster
- A comprehensive analysis is necessary to fully understand if and what physical significance the clusters have
- Optimization analyses can be performed to better the k-means cluster analysis outputs
- Standardization of the methodology will enable other scientists to conduct their research using this analysis

Acknowledgements

Computing resources used: NASA High-End Computing (HEC) Program of the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. Funding: NASA-ROSES Modeling, Analysis and Prediction 2013 NNX14AB99A-MAP and NNX15AJ05A. Clustering analysis: MATLAB ver 2016 computing environment.