Significance Testing for NEMI Cluster Validation

Jasper Dong

Background:

- The Native Emergent Manifold Interrogation (NEMI) algorithm utilizes manifold projection of data and agglomerative clustering in order to find underlying patterns and areas of interest in non-linear spaces

- Current validation methods for the NEMI clustering algorithm involves implementing entropy for uncertainty quantification

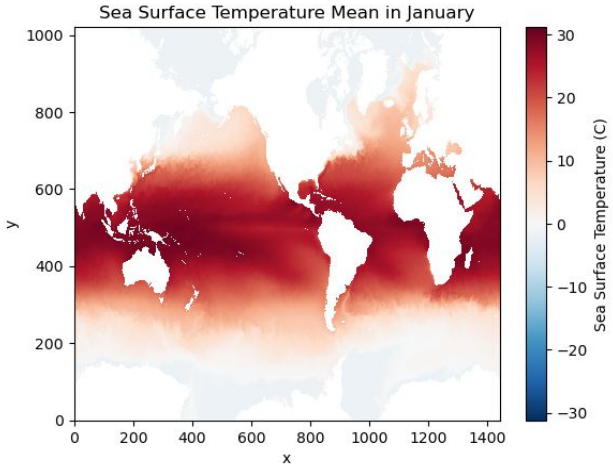
- Implementing a validation method based on statistical significance tests can be useful for users who would like to test for a new clustering result

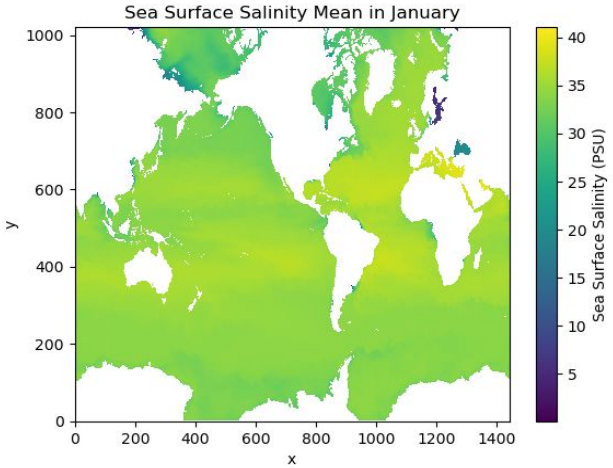
Data

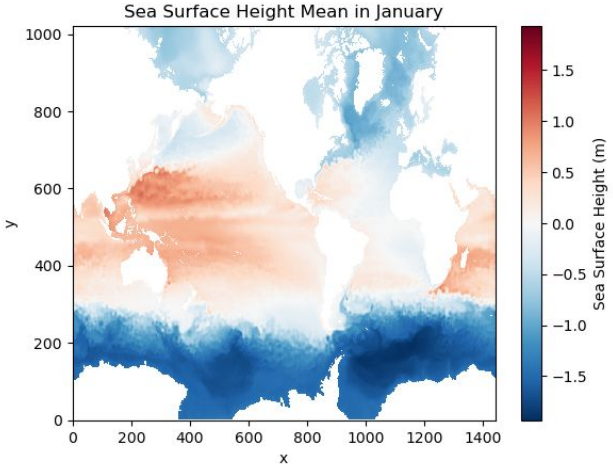
Source: European Centre for Medium-Range Weather Forecasts (ECMWF) OCEAN5 ocean analysis-reanalysis system (January 2004)

Variable	**Units**
SST (Sea surface temperature)	°C
SSS (Sea Surface Salinity)	PSU
SSH (Sea Surface Height)	m

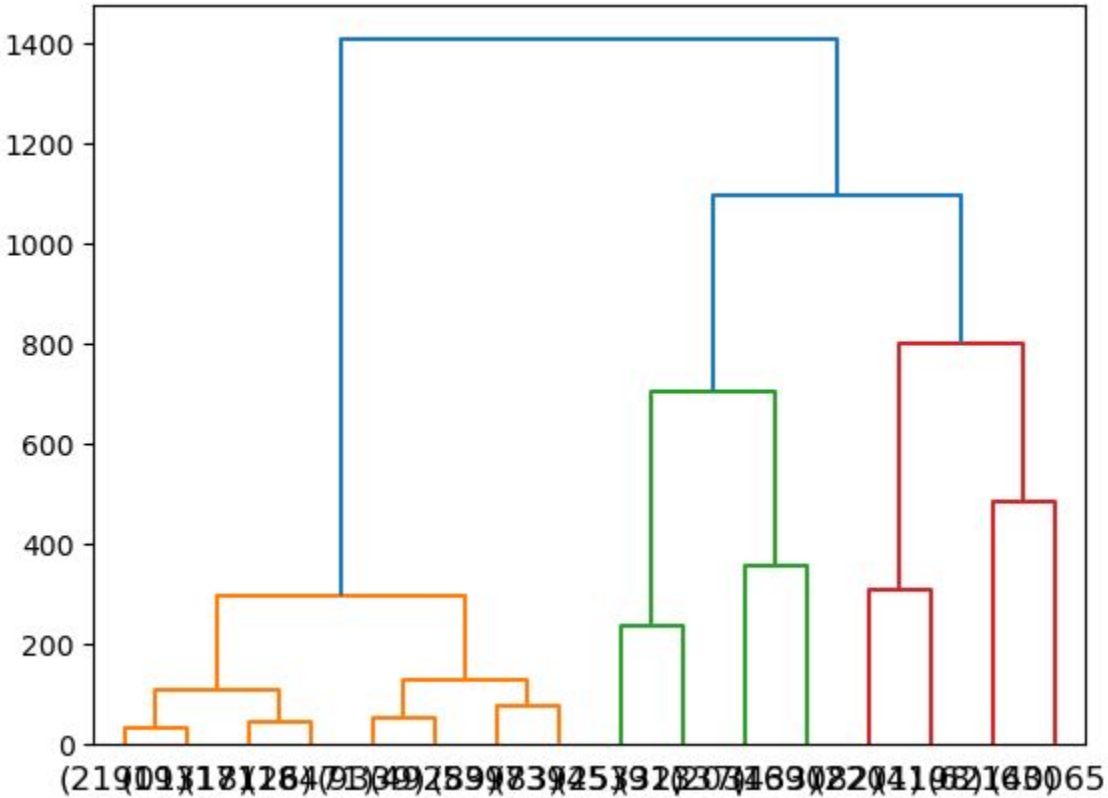
EDA


Sea Surface Temperature Mean in January
y
x
Sea Surface Temperature (C)


Sea Surface Salinity Mean in January
y
x
Sea Surface Salinity (PSU)

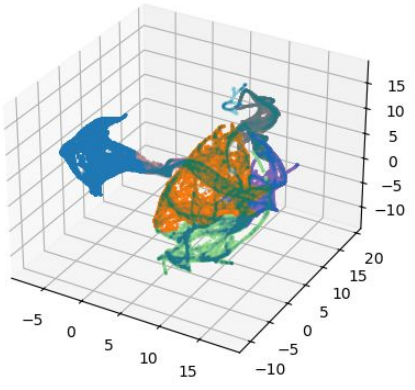

Sea Surface Height Mean in January
y
x
Sea Surface Height (m)

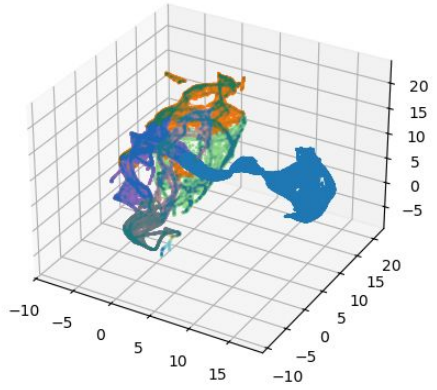
EDA

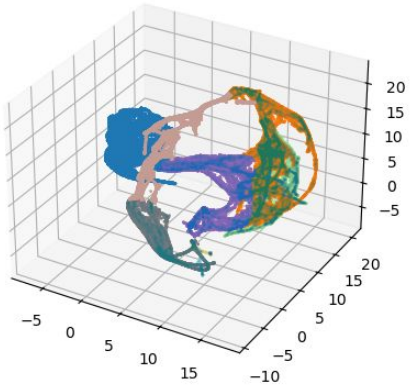


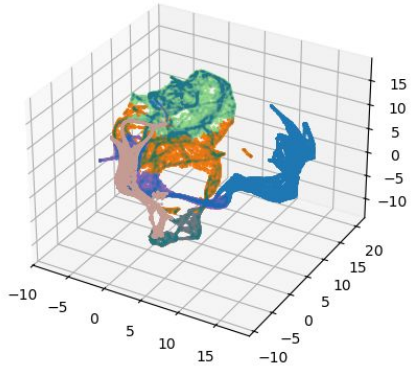
Dendrogram showing hierarchical clustering algorithm for a specific embedding

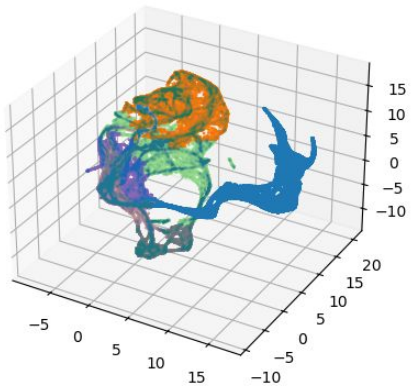
Embeddings and Clusterings



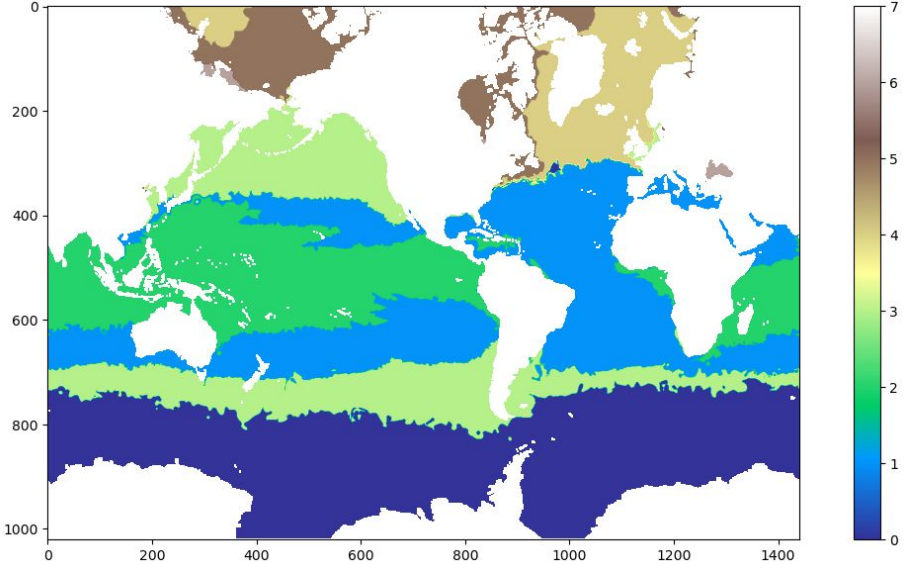


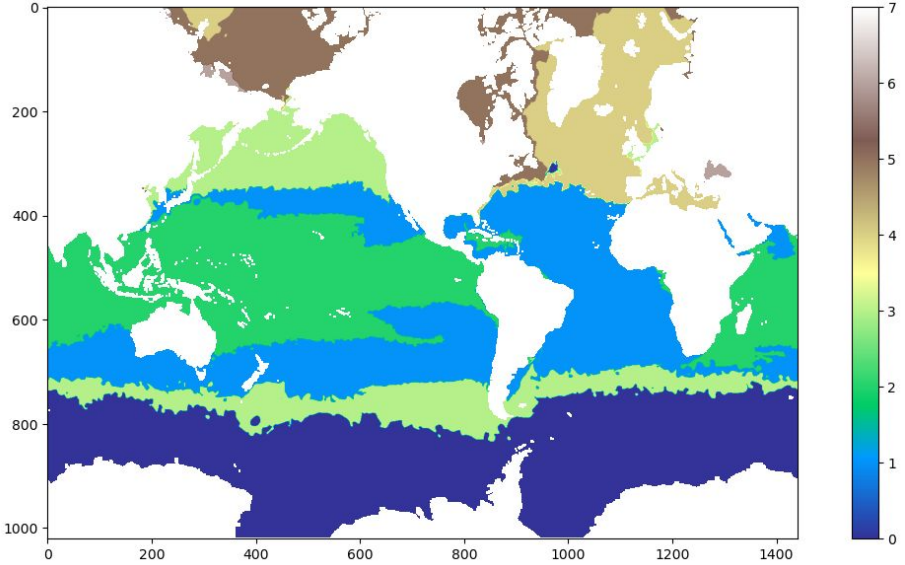






Embeddings and Clusterings





Methodology

- If we were to compare cluster results against each other, we can treat cluster labels as categorical variables with counts associated with them

- A user would be interested in how well a new cluster result will fit in with a set of existing clusterings of the same data

Methodology

The **Chi-squared Test for Goodness of Fit** tests whether the distribution of sample data fits a certain distribution.

$\mathbf{H_0}$: There is no difference between the distributions
$\mathbf{H_A}$**:** There is a difference between the distributions

Methodology

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

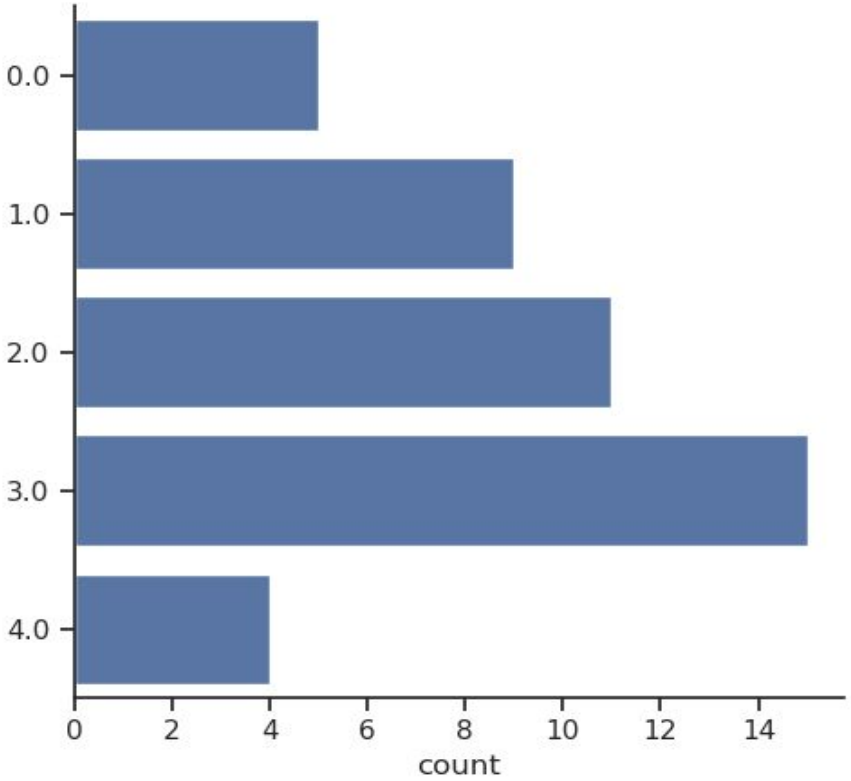
O_i = observed frequency of category *i*
E_i = expected frequency of category *i*

A p-value is obtained by comparing the Chi-squared test statistic to a chi-squared distribution with df = # of categories

Assumptions:

1. Large sample size
2. Expected counts are large enough
3. Independence

Sample Exercise

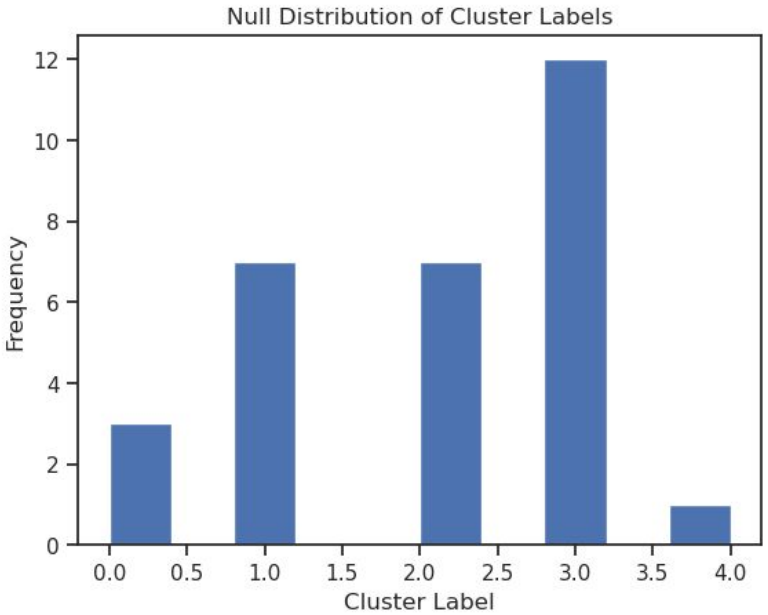


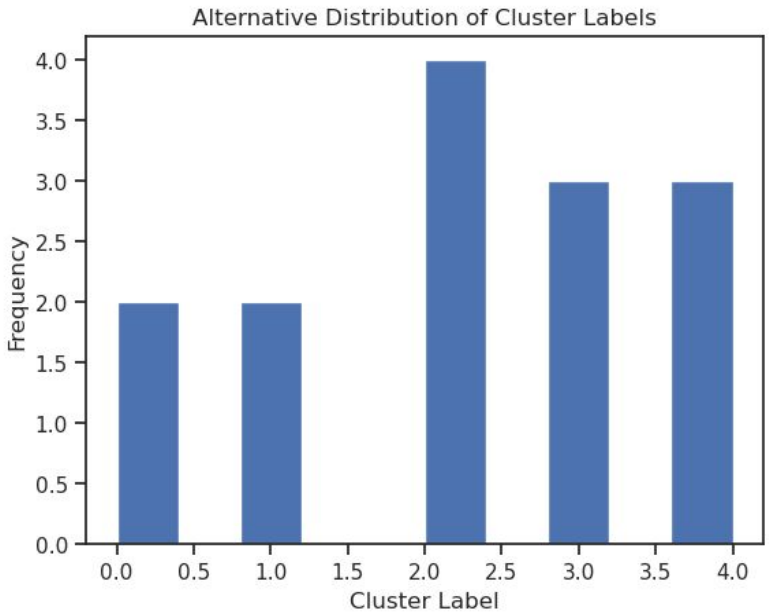
Cluster labels associated with a single point over 44 iterations

Cluster	Count
0	5
1	9
2	11
3	15
4	4

Sample Exercise

Null/Alternative Distribution 70/30 split





Sample Exercise

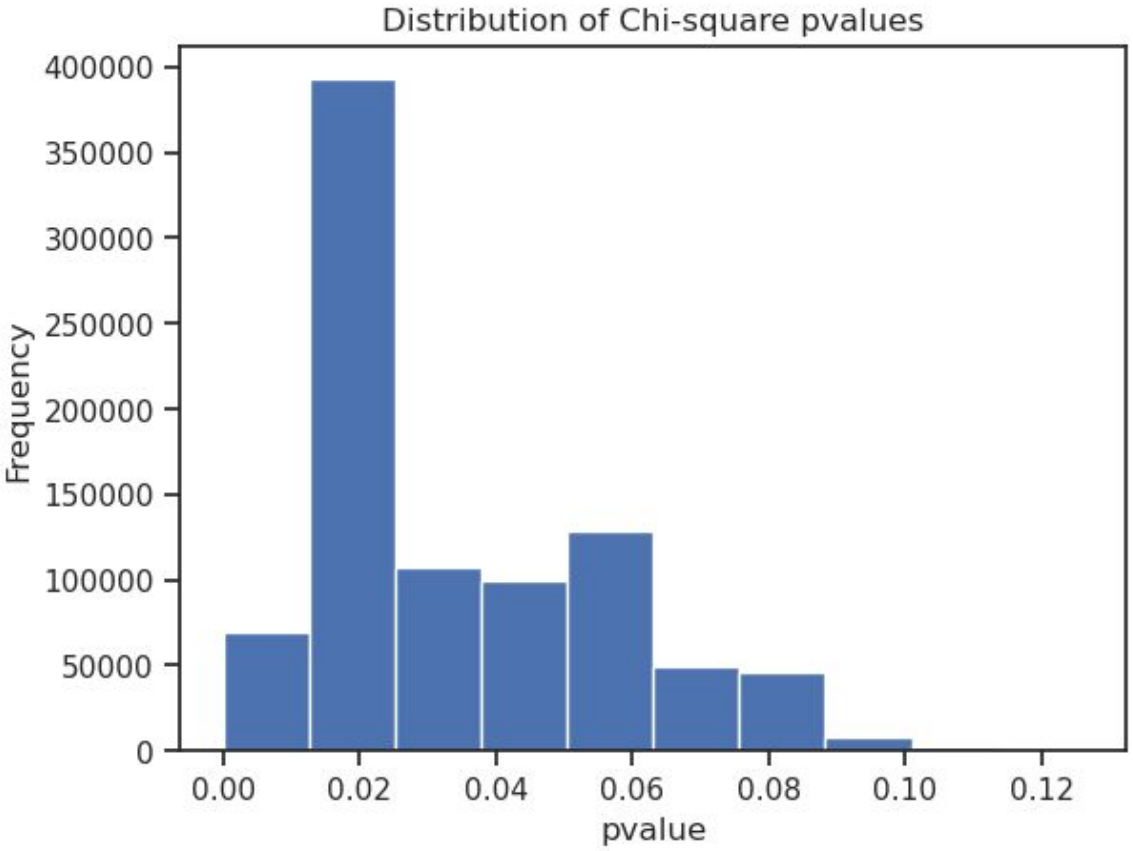
Cluster	Expected (Null)	Observed (Alternative)
0	3	2
1	7	2
2	7	4
3	12	3
4	1	3

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \approx 15.9405$$

p-value = 0.007

H_0 is rejected

Final Results



- P-values vary between 0.00 and 0.10, with a peak at around 0.02

- While most p-values would reject the null hypothesis, a considerable amount of p-values will also fail to reject

Discussion

- Chi-Square Goodness of Fit Test can be considered when carrying out tests of statistical significance for NEMI clusterings

- The nature of the results assumed that the user would compare a batch of new clustering results to a batch of existing ones

- Many assumptions of the test itself were most likely violated due to a small sample size/expected count values

Future Work

- Spatial mapping of p-values for better visualization

- Improvements to the significance testing method (test correction, tests with different assumptions/higher power, etc.)