# *Principal Component Analysis of Sea Surface Temperature via Singular Value Decomposition*

SYDE 312 – Final Project

Ziyad Mir, 20333385
Jennifer Blight, 20347163

Faculty of Engineering
Department of Systems Design Engineering
Waterloo, Ontario, Canada

***Abstract*— Principal Component Analysis is mathematical technique useful in reducing the dimensionality of a dataset, while preserving variation and determining its primary features. Using the Singular Value Decomposition of a de-trended, mean normalized covariance matrix of a dataset allows us to determine the orthogonal modes and their variance proportions. An example application is shown, which agrees with previous work in the field. Overall, Principal Component Analysis should be regarded as a powerful technique for determining important subsets of data to conduct in-depth analysis on.**

***Index Terms*— Principal Component Analysis, Linear Algebra, Weather Pattern Analysis**

## I. Background and Discussion

Weather and climate patterns hold importance with respect to both everyday life and scientific research. Data collection sensors and instruments allow information (e.g. sea surface temperature/pressure) to be harvested for predictive purposes. However, the immensity of the data collected poses a problem for those involved in analytical meteorology. Computational methods require much smaller sets of data in order to allow trends to be identified in a timely manner. Herein lies the power of mathematical techniques such as Principal Component Analysis (PCA), which help reduce the dimension of independent variables in a data set, while preserving as much variance as possible. PCA helps identify the primary components of a data set, allowing in-depth analysis to be conducted on a much smaller set of data. Furthermore, predictive models often rely on the first few principal components as relevant feature sets. Climate scientists depend heavily on this analysis to form conclusions regarding weather and climate patterns across the globe.

## II. Relevance

The primary mathematical technique employed in this analysis is known as Principal Component Analysis (PCA). PCA is a procedure that allows a set of independent variables to be reduced into their principal components. These components are variables which are useful in explaining the variation in the original data set; essentially reducing the size of the data set, while retaining as much relevant information regarding the set as possible.

Linear algebra plays an integral part in conducting a PCA, which is conducted using a singular value decomposition of a mean-centered covariance matrix of a data set. The techniques needed to compute a covariance matrix, remove its empirical mean, and compute its eigenstructure all have relevant ties to linear algebra.

## III. Case Study

The case study of interest concerns the analysis of ocean temperature and their impact on climatology. Sea surface temperature holds importance in the field of numerical weather prediction, as it has an influence on the atmosphere above it. For instance, low and high pressure zones can be determined by looking at variations in temperatures in ocean groups.

The National Climatic Data Center has a repository of data representing the surface temperature spanning the past 50 years [1]. In this case study, sea surface temperature for a 30 year time series (1980-2009) in the northern-hemisphere (from the equator to the 88th parallel, and between 90 degrees west to the Prime Meridian) is analyzed, with specific attention paid to the North Atlantic Oceans, for fluctuations in atmospheric pressure differences.

This data is preprocessed (detrended and normalized) and analyzed using PCA. Then, contour maps of the computed eigenstructure are drawn for analysis and comparison. The results are compared with previous research done regarding the North Atlantic Oscillation and Atlantic Multidecadal Oscillations, known patterns, which principally express themselves via sea surface temperature variations.

## IV. Analysis

The mathematical solution to the case study is outlined below. All techniques were implemented in Octave.

1) Data Organization
2) Data Preprocessing

i) Land Removal
ii) Detrending
iii) Mean Removal
iv) Normalization

3) Principal Component Analysis using Singular Value Decomposition

4) Computing the Variance Proportions of each Principal Component

*Data Organization*

For each year from 1980 to 2009, an A = MxN row matrix is constructed, with M being 30 (the number of years used in the analysis), and N representing the number of latitude-longitude gridpoints. Each column of this matrix is a 30-year timeseries of sea surface temperature values for some point in the Atlantic ocean. In this setup, each value represents the mean sea surface temperature over a specific three-month period within that year. Four separate experiments were conducted, corresponding to four seasons: Winter (January-February-March), Spring (April-May-June), Summer (July-August-September), and Fall (October-November-December).

*Data Preprocessing*

1) First, all grid points corresponding to land-mass are removed from the above matrix. This allows only the points pertaining to sea surface temperatures to be considered. These columns were comprised entirely of fill values and have no relevance to the analysis.

2) Next, the A = MxN matrix is detrended, by subtracting a k (where k = 1) degree polynomial from each column of the matrix. This accounts for the general increase in temperature over the past few decades, and instead produces a matrix which is more representative of inter-year variability. The result is a new matrix, A' = MxN, with the same size and shape as A.

3) Then, the mean of A' is removed, leaving A'' = A' - mean(A), where mean(A) is a row vector containing the mean of each column in A, and the mean of a column, C, in A is simply the arithmetic mean of the elements in C.

4) Then, A''' = normalized(A'') is computed, by dividing each entry in A'' by the standard deviation of A''.

*Principal Component Analysis using Singular Value Decomposition*

Now, A''' is ready to run through a Principal Component Analysis via Singular Value Decomposition, where we seek to find the dominant orthogonal modes of A''', and begin by calculating, C, the covariance matrix. Then, by using the singular value decomposition factorization $C = USU^T$ we can find, U, a matrix containing the eigenvectors of $CC^T$, and S, whose diagonal elements contain the singular values. The columns of U are orthogonal vectors that describe a series of uncorrelated features representing the original matrix A'''. The singular values indicate what percentage of the original variance is described by each of these components. By selecting the first K columns of U (the K most important principal components of the data), we have effectively reduced the dimension of the data from N to K, while preserving a significant amount of the variance in the data [2].

*Computing the Variance Proportions of each Principal Component*

The exact variance proportion preserved by selecting the first K principal components can be computed by finding the sum of the diagonals of the covariance matrix of A''', and dividing each entry in Y by that value.

## V. Results and Discussion

For each principal mode, the components of the eigenvector are mapped back to a latitude-longitude grid, in order to visually identify dominant patterns in the dataset. The results of the analysis are clear when a series of consecutive principal component maps are observed and compared.

*Principal Components by Season*

Figures 1-4 display the map of the first mode of the singular value decomposition for the time series data associated with each season. Upon inspection, the heatmaps display slight variations in centers during different times of the year. This corresponds with the variations in sea temperature (and resultantly, with other important facets of climatology) that vary according to the time of year. The heatmaps generated in the analysis resemble the North Atlantic Tripole, a known weather pattern observed by (Deser and Timing) and (Meizhu and Schneider). While this is not an original result, the agreement between this experiment and previous work should serve to demonstrate the usefulness and validity of this analysis.

*Principal Components by Variance Proportion*

Figures 5 through 9 display the eigenvalue contribution of the first five modes of the singular value decomposition for the fall data experiment, accounting for 60% of the variance in the original data. By inspection, the number of primary centers generally increases as the variance proportion of the modes decreases (as with consecutive principal components). The plots with clearer centers are associated with the higher order principal components, while lower order principal components result in heat maps that contain more noise.

## VI. Conclusion

The Principal Component Analysis (PCA) technique helps reduce the dimensionality of a dataset of interest, while preserving much of the variance of the data. The results of the analysis conducted within this report are supported by previous research into climatology patterns, and as such, help demonstrate the validity of the work presented. Overall, while PCA is not an end-to-end solution, it certainly has value in narrowing down a dataset for further analysis. Using such a technique allows analysts to efficiently determine the pockets

of data that hold the most value for in-depth analysis. By narrowing down the volume of data to analyze, the process of extracting meaning from raw datasets can be expedited while retaining as much quality in information as possible.

## REFERENCES


[1] L. K. Hansen, O. Winther. (2004, February) *Singular value decomposition and principal component analysis*, Class notes, IMM, DTU.

[2] Ray, R. (2012, April 19). *Extended reconstructed sea surface temperature (ersst.v3b).* Retrieved from http://www.ncdc.noaa.gov/ersst/

[3] Fan, Meizhu, Edwin K. Schneider. (2012). *Observed decadal north atlantic tripole sst variability. part i: weather noise forcing and coupled response*. J. Atmos. Sci., 69, 35–50. doi: http://dx.doi.org/10.1175/JAS-D-11-018.1

[4] Deser, Clara, Michael S. Timlin. (1997). *Atmosphere–ocean interaction on weekly timescales in the north atlantic and pacific*. J. Climate, 10, 393–408. doi: http://dx.doi.org/10.1175/1520-0442(1997)010<0393:AOIOWT>2.0.CO;2

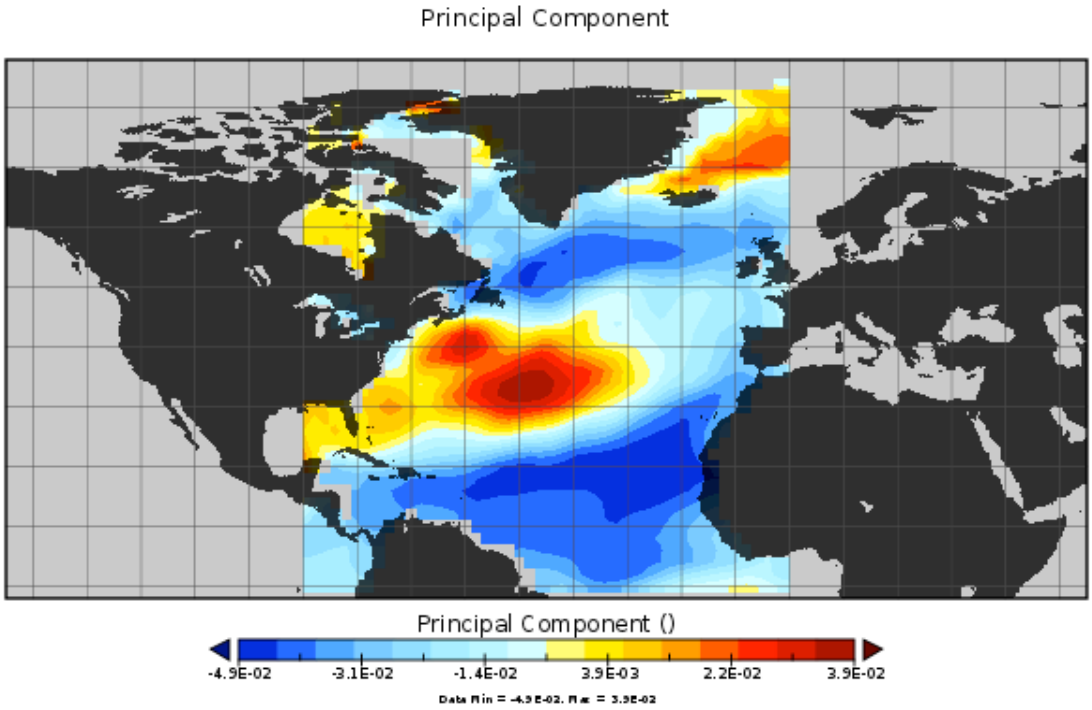


*Figure 1 - First Principal Component during Winter Season (January, February, March)*

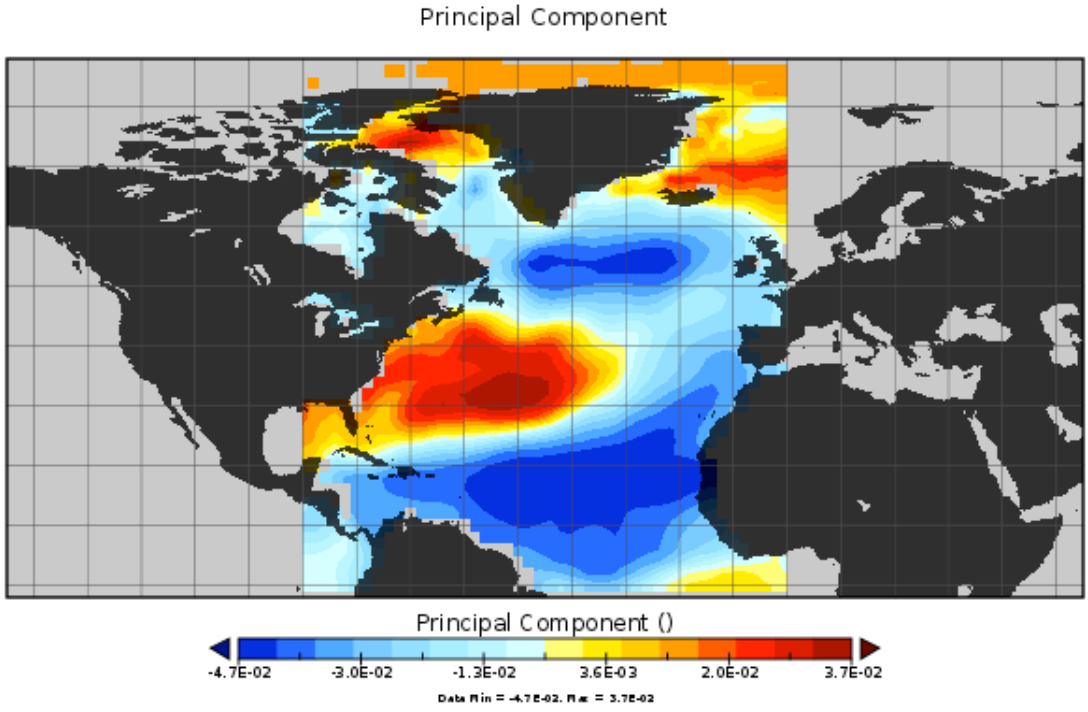


*Figure 2 - First Principal Component during Spring Season (April, May, June)*

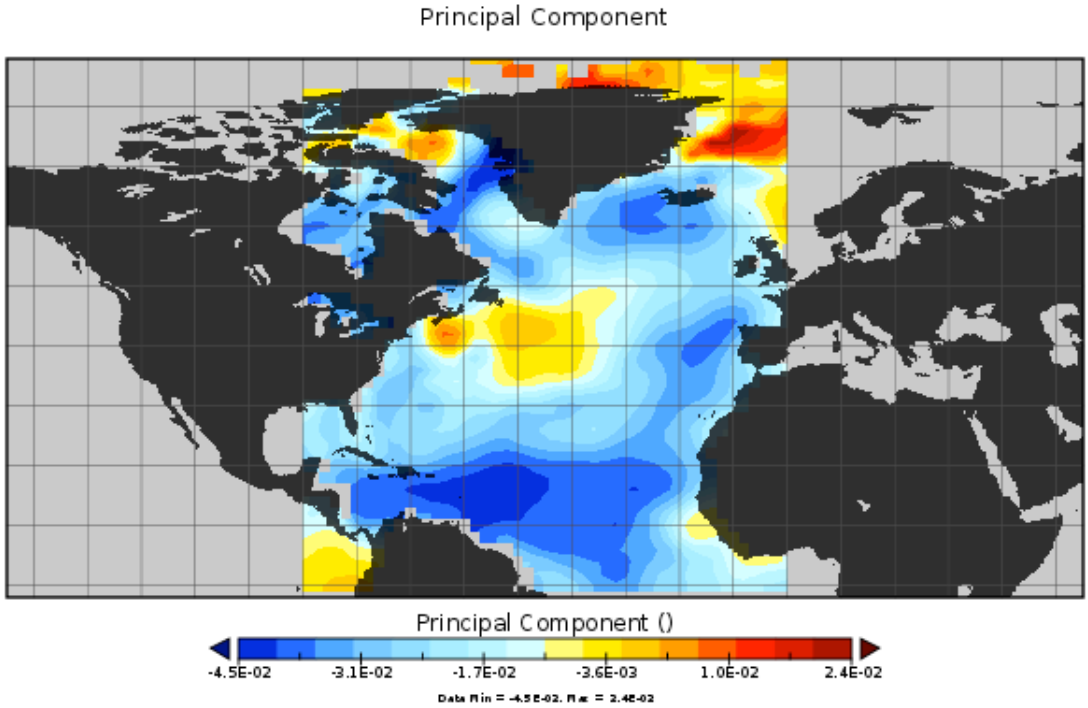


*Figure 3 - First Principal Component during Summer Season (July, August, September)*

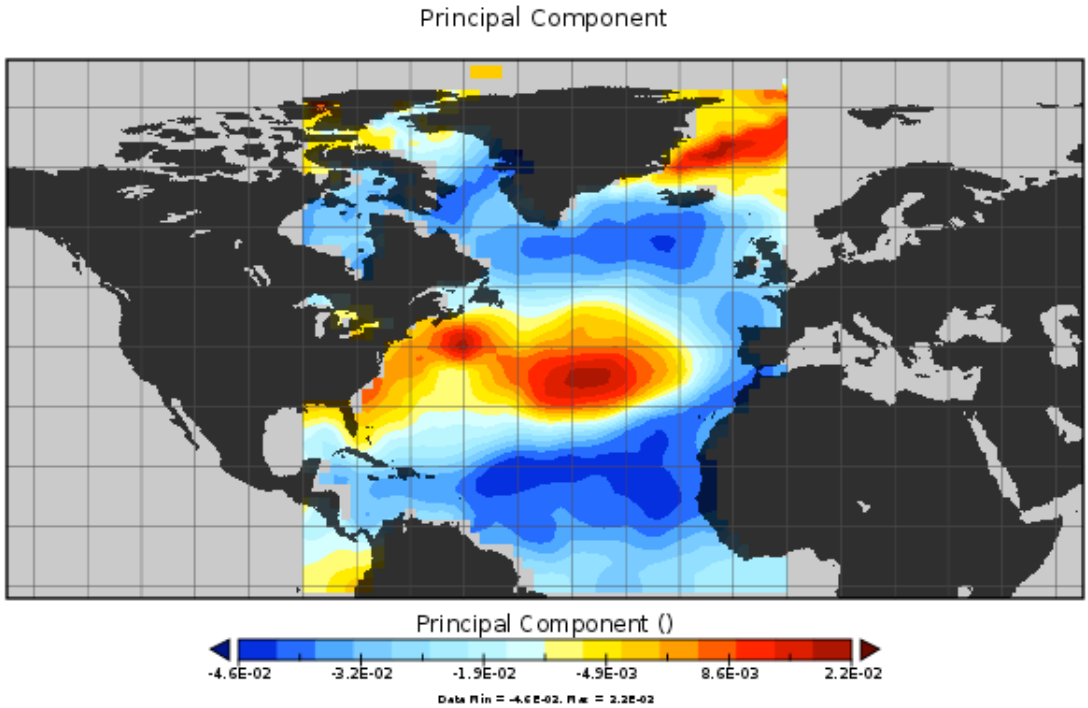


*Figure 4 - First Principal Component during Fall Season (October, November, December)*

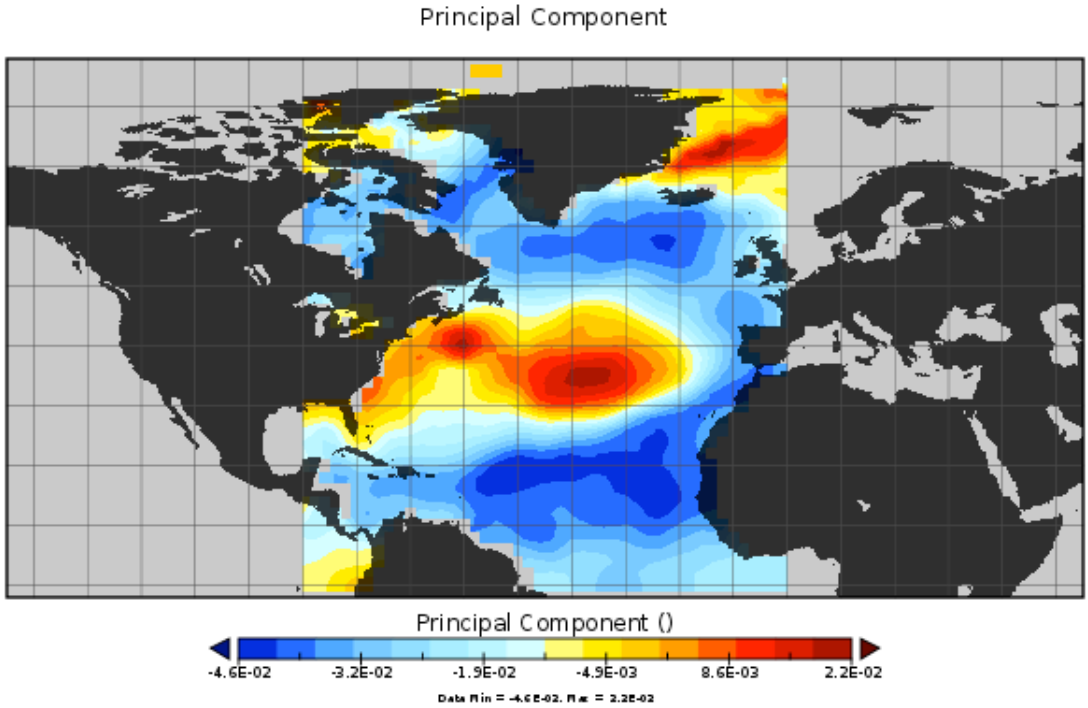


*Figure 5 - First Principal Component during Fall Season - Variance Proportion of 0.247*

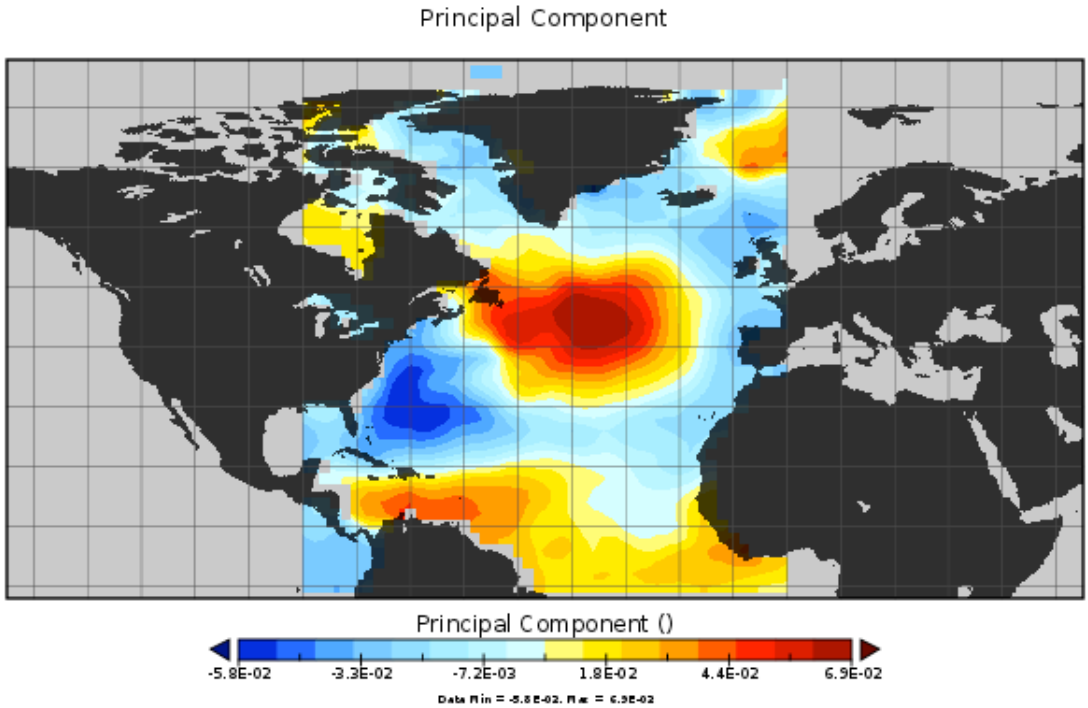


*Figure 6 - Second Principal Component during Fall Season - Variance Proportion of 0.1909*

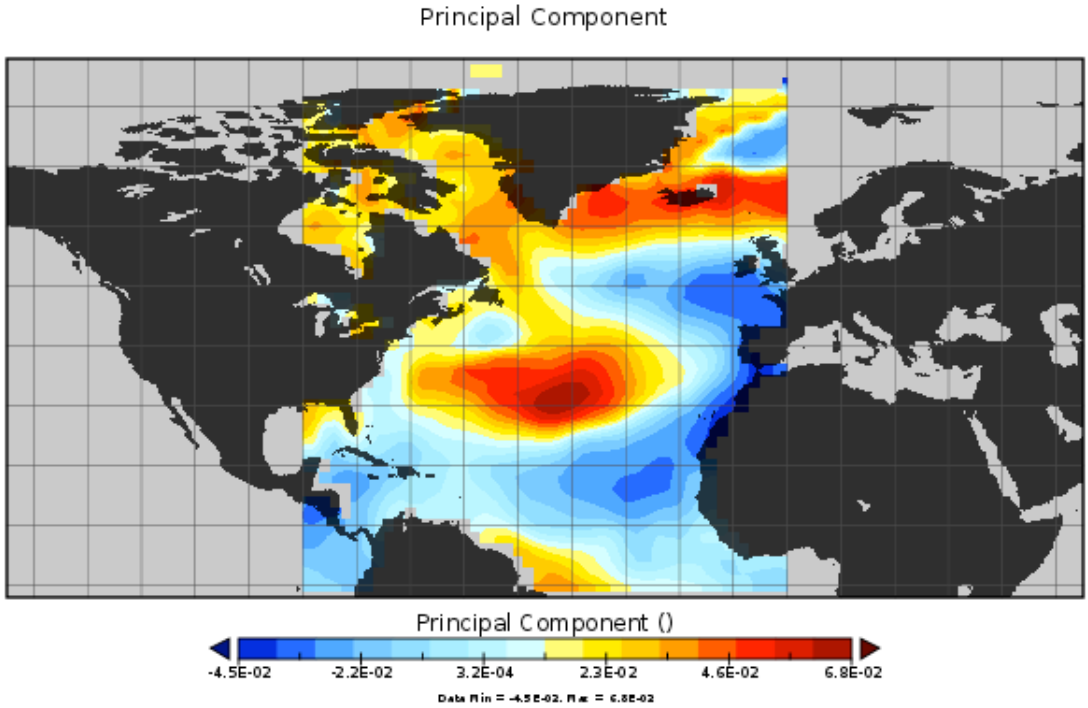


*Figure 7 - Third Principal Component during Fall Season - Variance Proportion of 0.090*

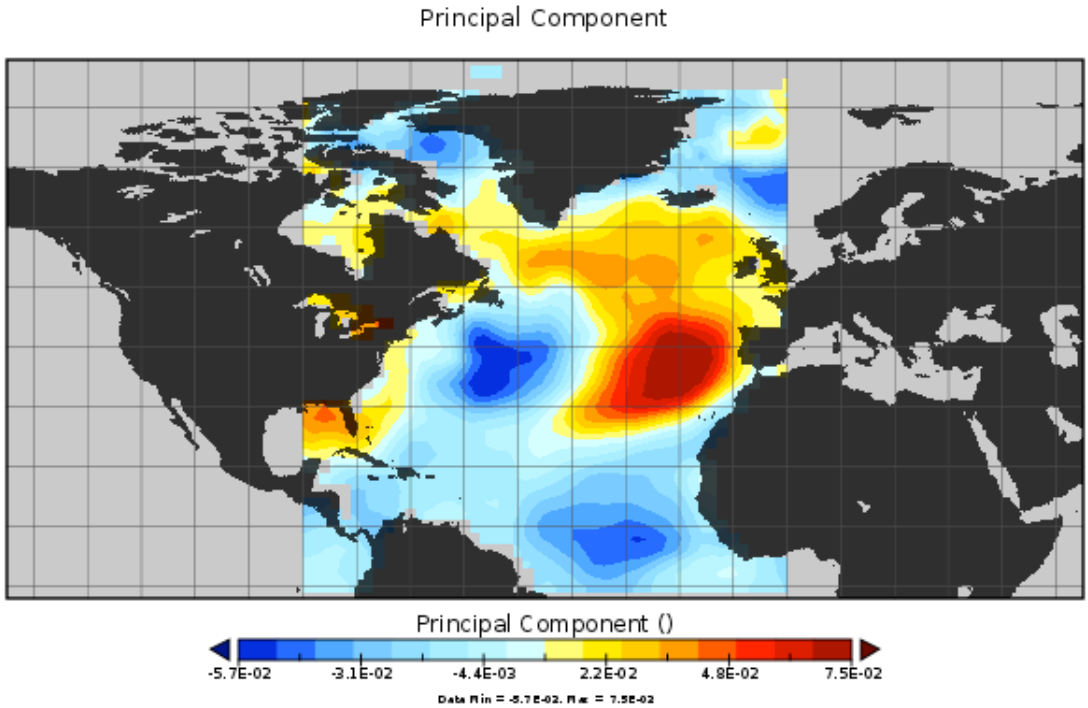


*Figure 8 - Fourth Principal Component during Fall Season - Variance Proportion of 0.088*

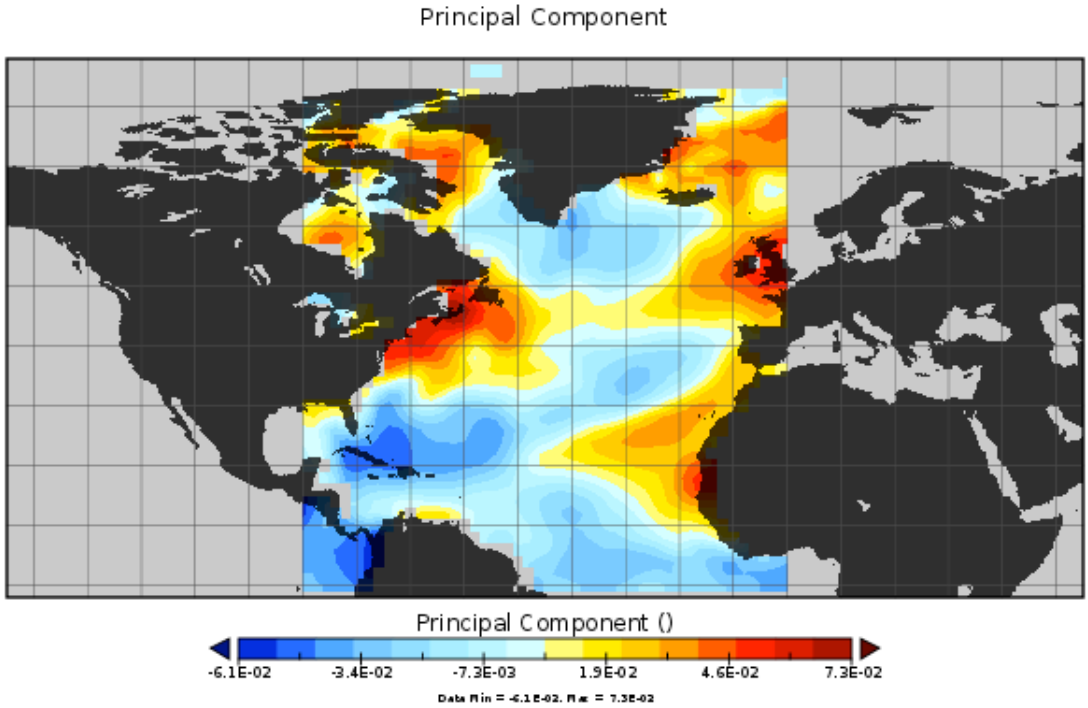


*Figure 9 - Fifth Principal Component during Fall Season - Variance Proportion of 0.063*