Unsupervised machine learning for ocean profile classification and outlier detection using the Pacific Ocean temperature-conductivity-depth profile data

Introduction

General Background

Temperature-conductivity-depth (CTD) profiles captures the thermodynamic state of oceans (figure 1). These profiles exhibit different shapes, which change seasonally and geographically. They encode information about the state of the ocean, including scientifically interesting events (e.g. heatwaves).

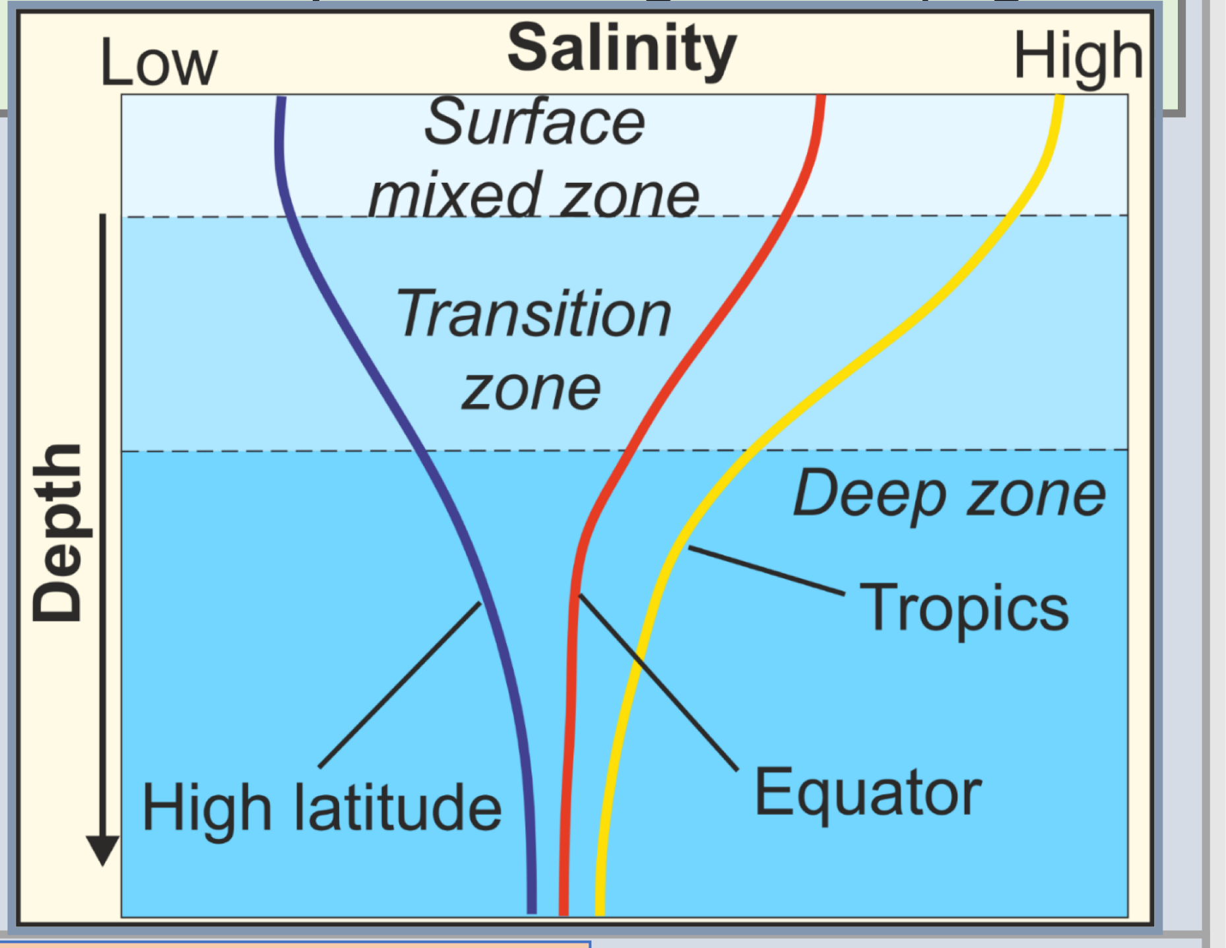


Figure 1. Representative CTD profiles of three geographical zones are shown.

Problem Definition

The Science Branch of the Pacific Region of Fisheries and Oceans Canada (DFO) has accumulated ~10^5 CTD profiles of variable quality from 1980 to 2020 (figures 2 and 3). Funded by the Results Fund for 2020-2021, the Office of the Chief Data Officer, IM/TS and the Science Branch of the Pacific Region partnered to experiment with the application of artificial intelligence in a cloud-environment using CTD data.

Problem: Classify CTD profiles based on their similarity to understand the quality of the data and to detect anomalies. This would enable scientists to monitor and understand ocean structural evolution.

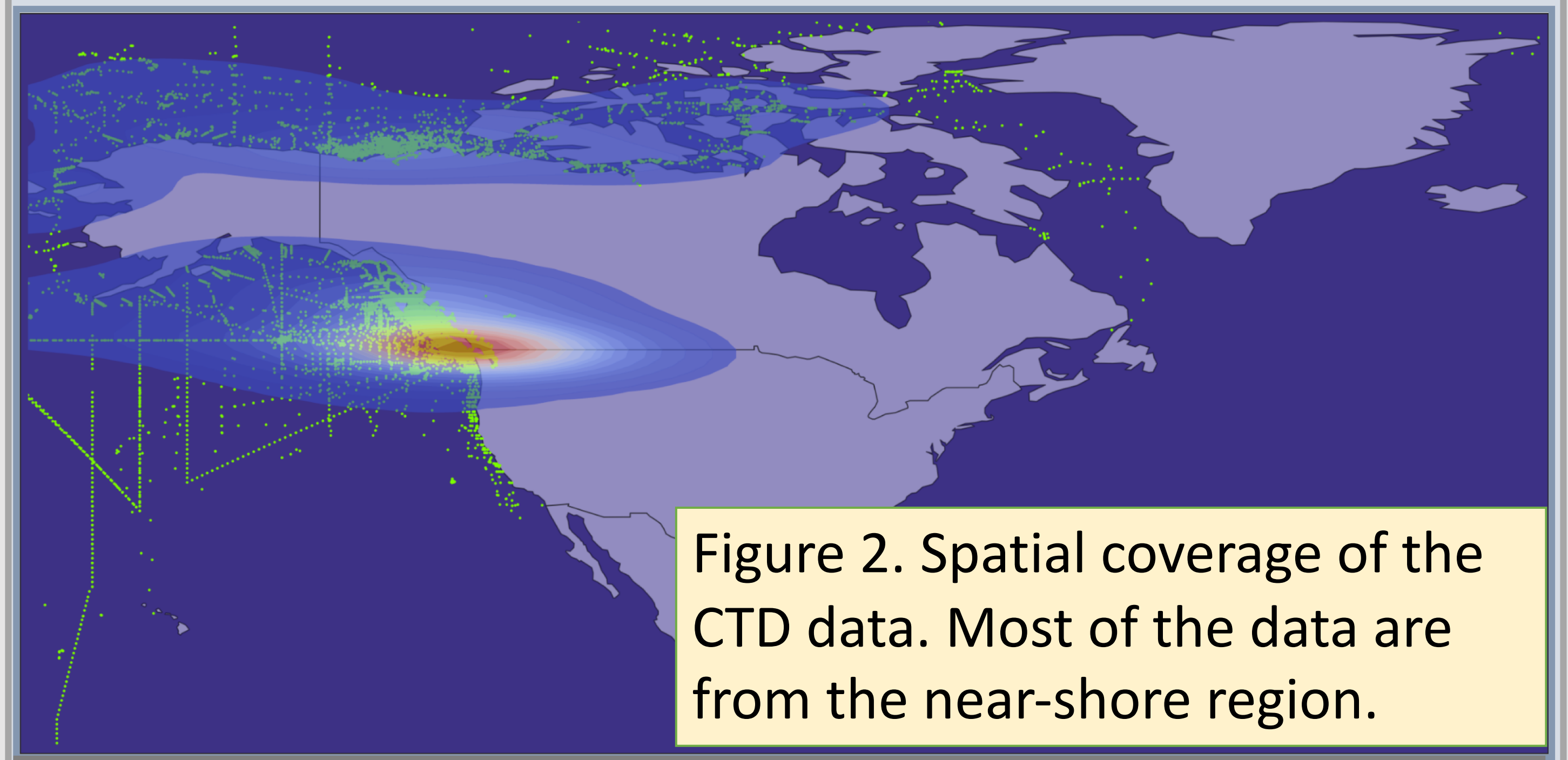


Figure 2. Spatial coverage of the CTD data. Most of the data are from the near-shore region.

Solution: Use unsupervised machine learning to automatically classify CTD data and identify anomalies.

Methodology

Pacific Region CTD data (1980 – 2020)

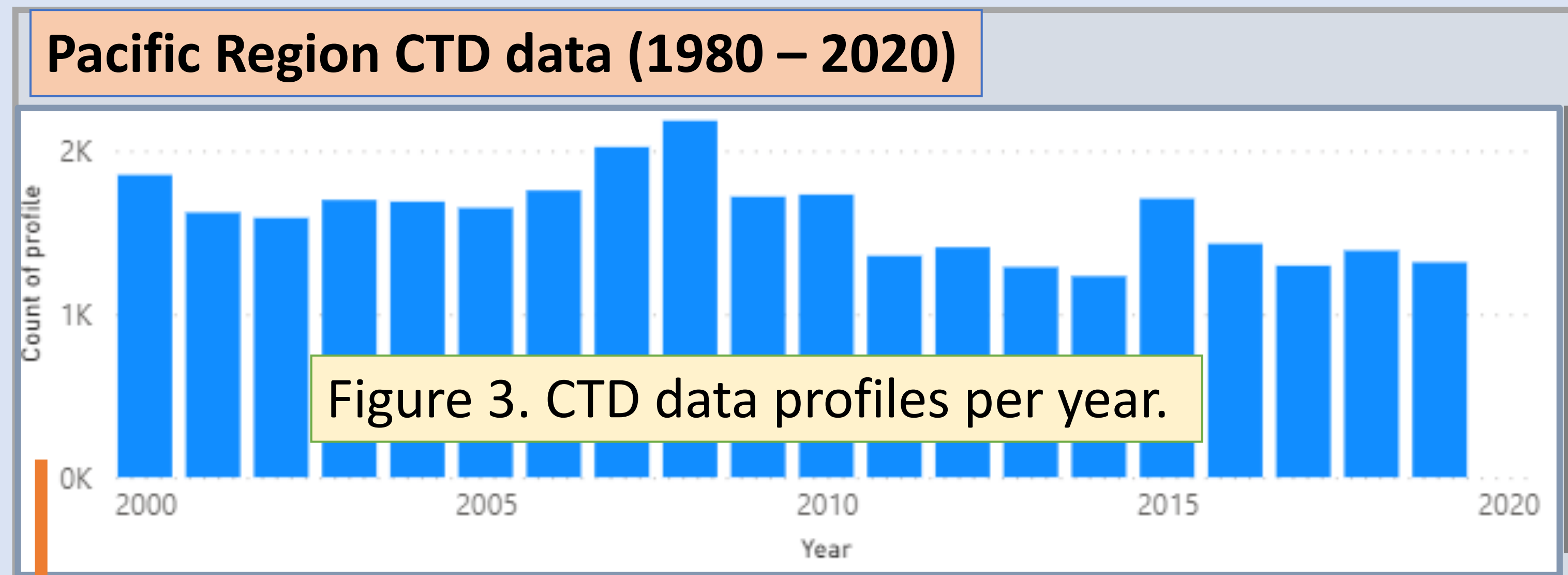


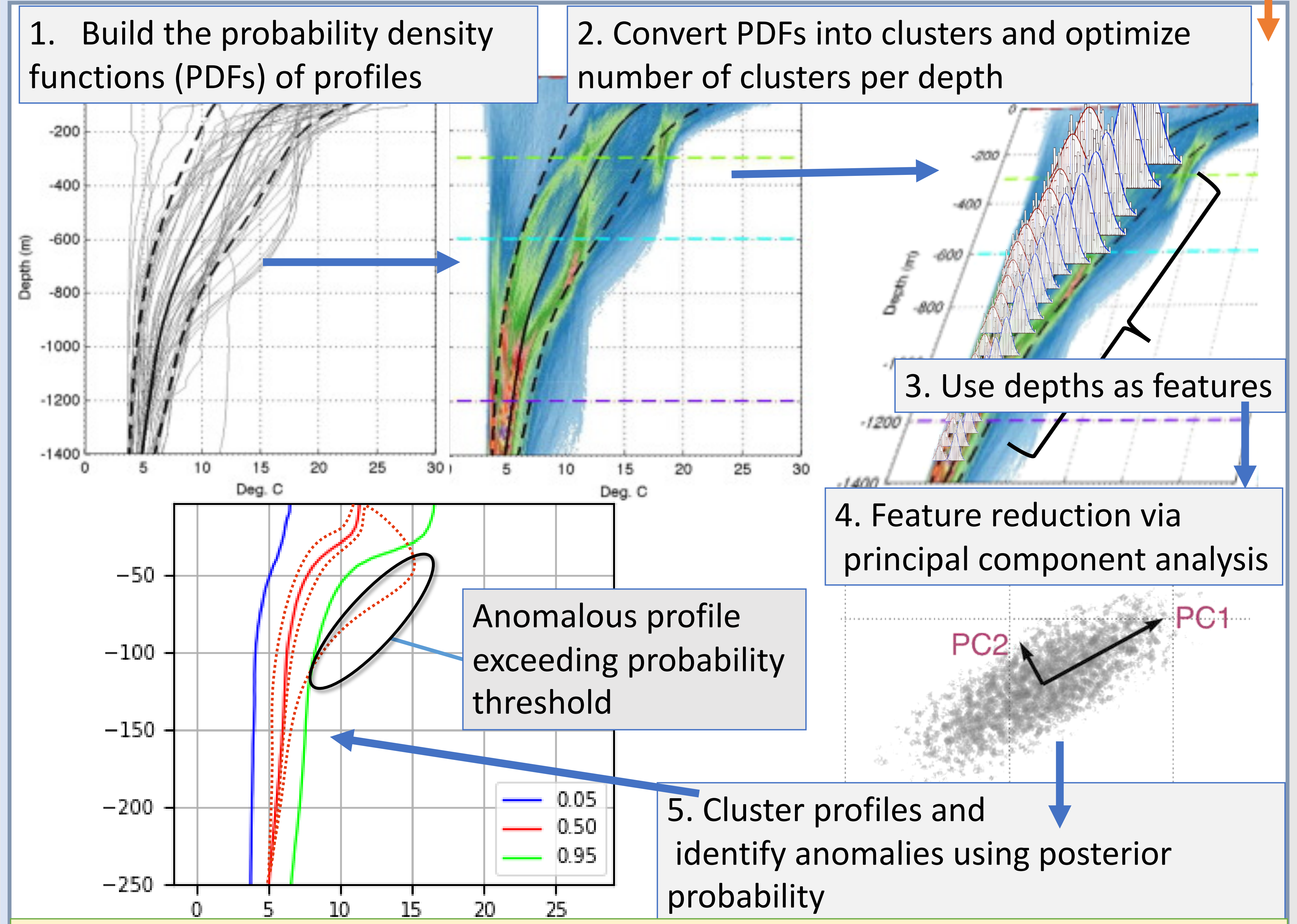
Figure 3. CTD data profiles per year.

Machine Learning

Following Maze et al., 2017, we use pyXpcm on a cloud computing environment.

Data preprocessing and quality control - challenges with the use of legacy CTD data (and remediation methods)

- Inconsistent number of data elements (reformat of data into a single standard)
- Inconsistent metadata (matching of metadata by inference and manual verification, rebuilt database into a single standard)
- Missing data (kept only data that contains both temperature and salinity)
- Spatial and temporal bias (not important for proof of concept)
- Duplicated profile names (removed duplicates)
- Depth coverage rate inconsistency (depth is gridded and interpolated)



1. Build the probability density functions (PDFs) of profiles

2. Convert PDFs into clusters and optimize number of clusters per depth

3. Use depths as features

4. Feature reduction via principal component analysis

5. Cluster profiles and identify anomalies using posterior probability

Anomalous profile exceeding probability threshold

Figure 4. Machine learning-based clustering and anomaly detection process, after Maze et al., 2017 (*Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean*).

Results and Outlook

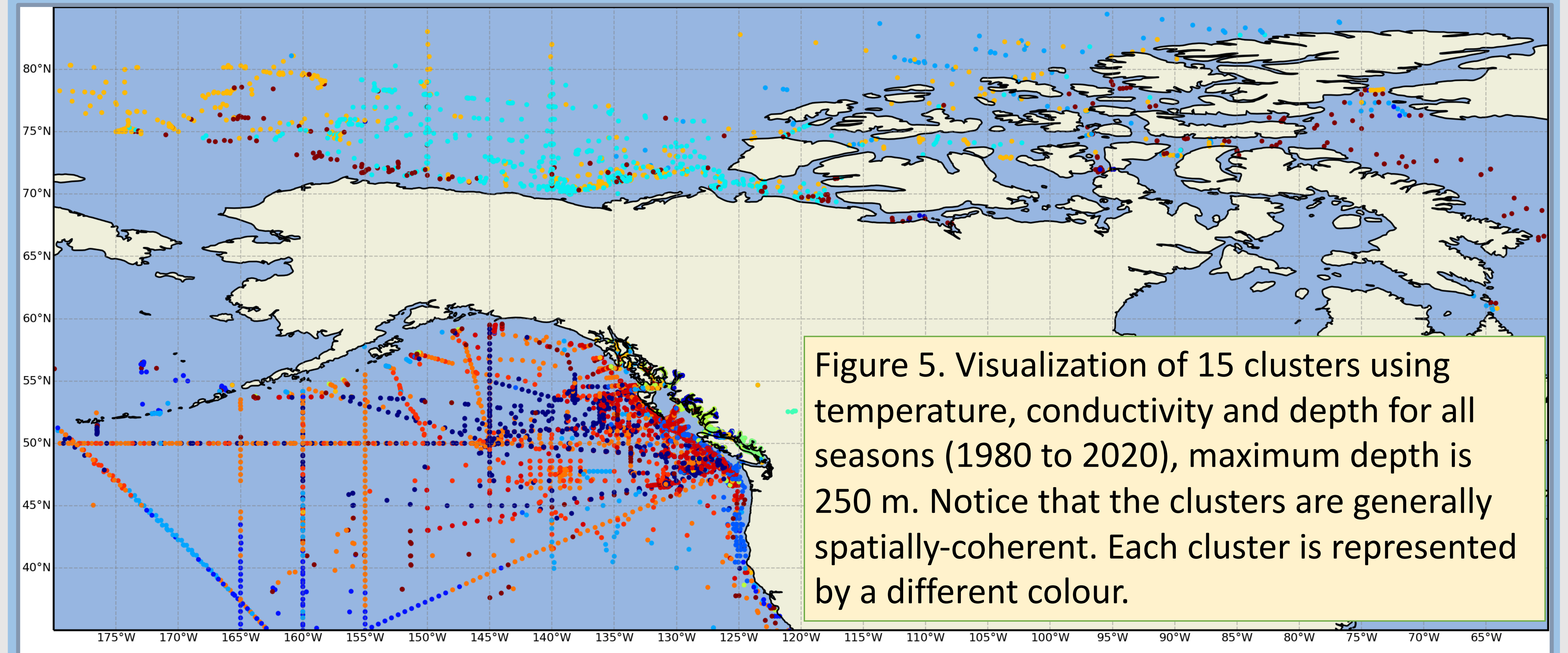
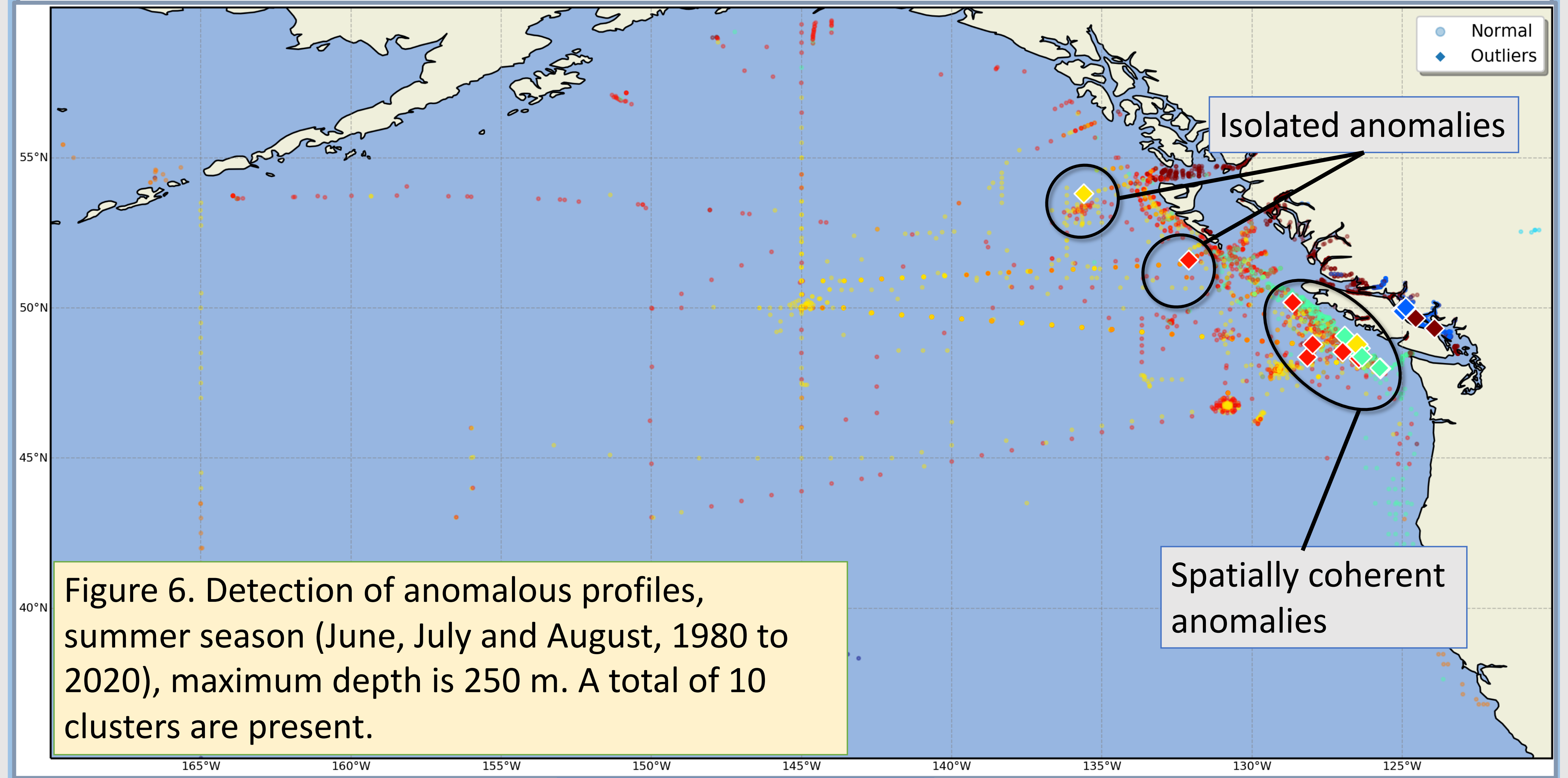


Figure 5. Visualization of 15 clusters using temperature, conductivity and depth for all seasons (1980 to 2020), maximum depth is 250 m. Notice that the clusters are generally spatially-coherent. Each cluster is represented by a different colour.



Isolated anomalies

Spatially coherent anomalies

Figure 6. Detection of anomalous profiles, summer season (June, July and August, 1980 to 2020), maximum depth is 250 m. A total of 10 clusters are present.

Summary and Outlook

- Unsupervised machine learning is highly appropriate for ocean profile classification.
- The method is robust to the quality of legacy CTD data and highly automatable.
- Spatially-coherent patterns are consistently observed and the structure varies by season.
- Members that are the most dissimilar within each cluster can be deemed to be anomalous.
- Spatially-coherent anomalies may be associated with physical oceanographic events, whereas isolated anomalous profiles may be data quality issues.
- Legacy CTD data is of enormous value for both scientific and data science-based method development and insight discovery.
- Cloud computing environments may be ideal for the use and sharing of large datasets.

Fisheries and Oceans Canada Pêches et Océans Canada

The Office of the Chief Data Officer

Steven E. Zhang[1], Riham Elhabyan[1], Di Wan[2]
[1]Office of the Chief Data Officer, Fisheries and Oceans Canada. E-mail: Steven.Zhang@dfo-mpo.gc.ca
[2]Science Branch, Pacific Region, Fisheries and Oceans Canada.

Canada