Classifying Coastal Upwelling Using Environmental Variables

Derya Gumustel
2021-05-13 13:00:00

Pieces of This Project

Problem Statement:

Building a model to identify upwelling using environmental variables.

Oceanography side:

- Studying natural processes in the oceans: upwelling and downwelling
- Using data collected by imperfect sensors

Data science side:

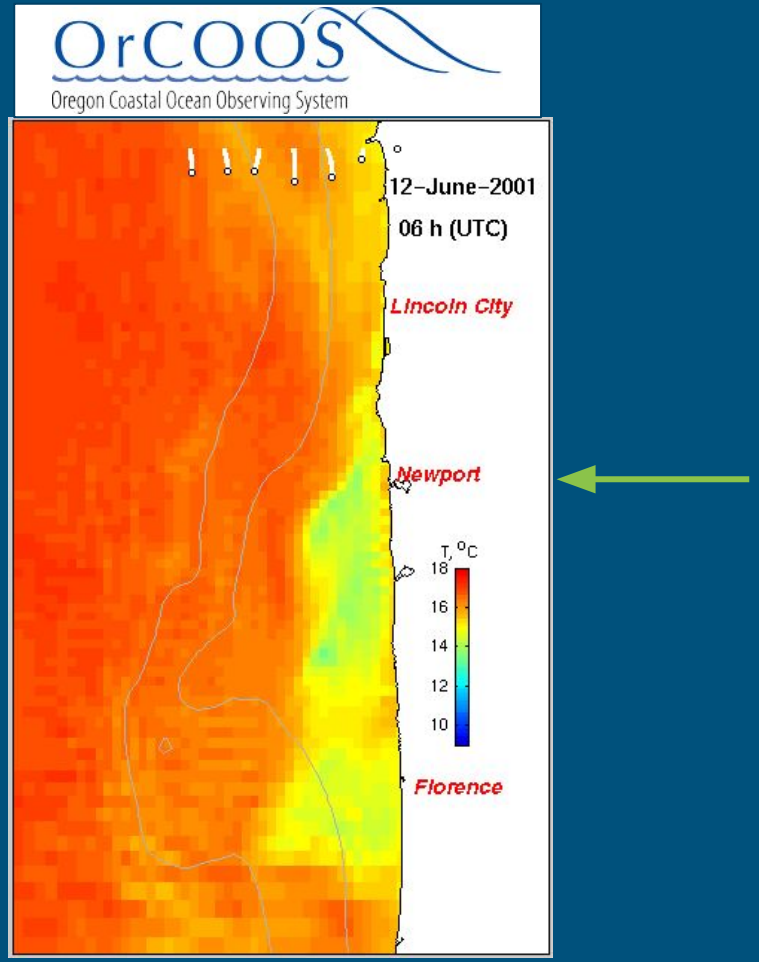
- Collecting and cleaning real-world data
- Gaps in data availability
- Processing the data for modeling

Background & Data Collection

What is Coastal Upwelling?

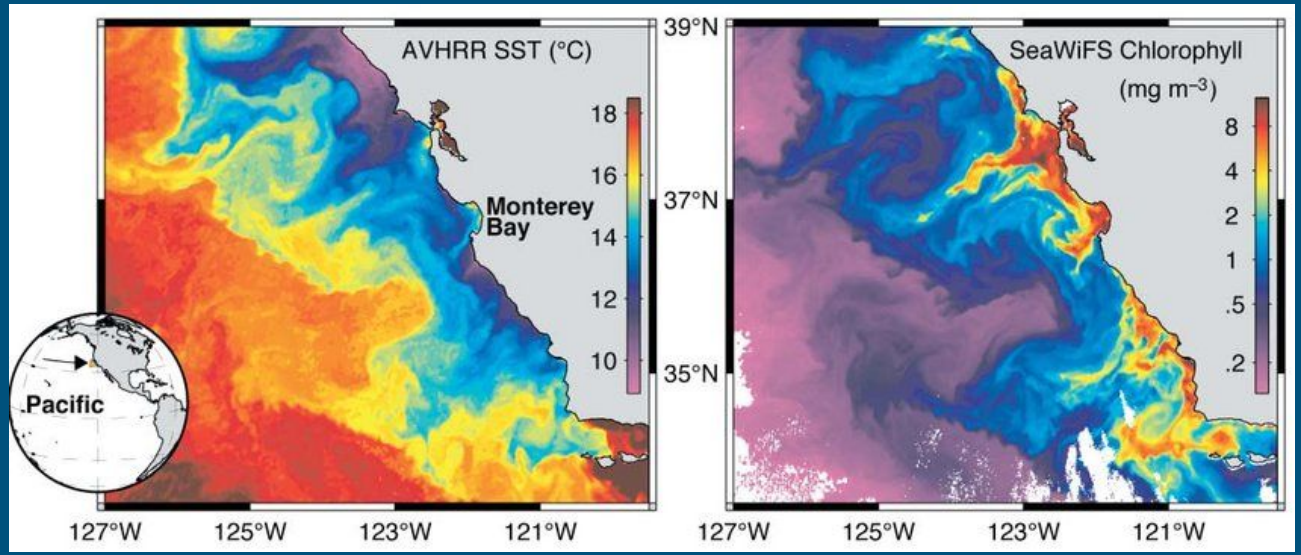
Winds blowing parallel to coastlines push surface water away from shore. Deep water is pulled up to the surface to replace it - this is called upwelling!

Downwelling describes the opposite - surface water is pushed towards shore and sinks when it reaches land.



Source: OrCOOS, http://agate.coas.oregonstate.edu/ocs/coastal_current.html

Why Do We Care?



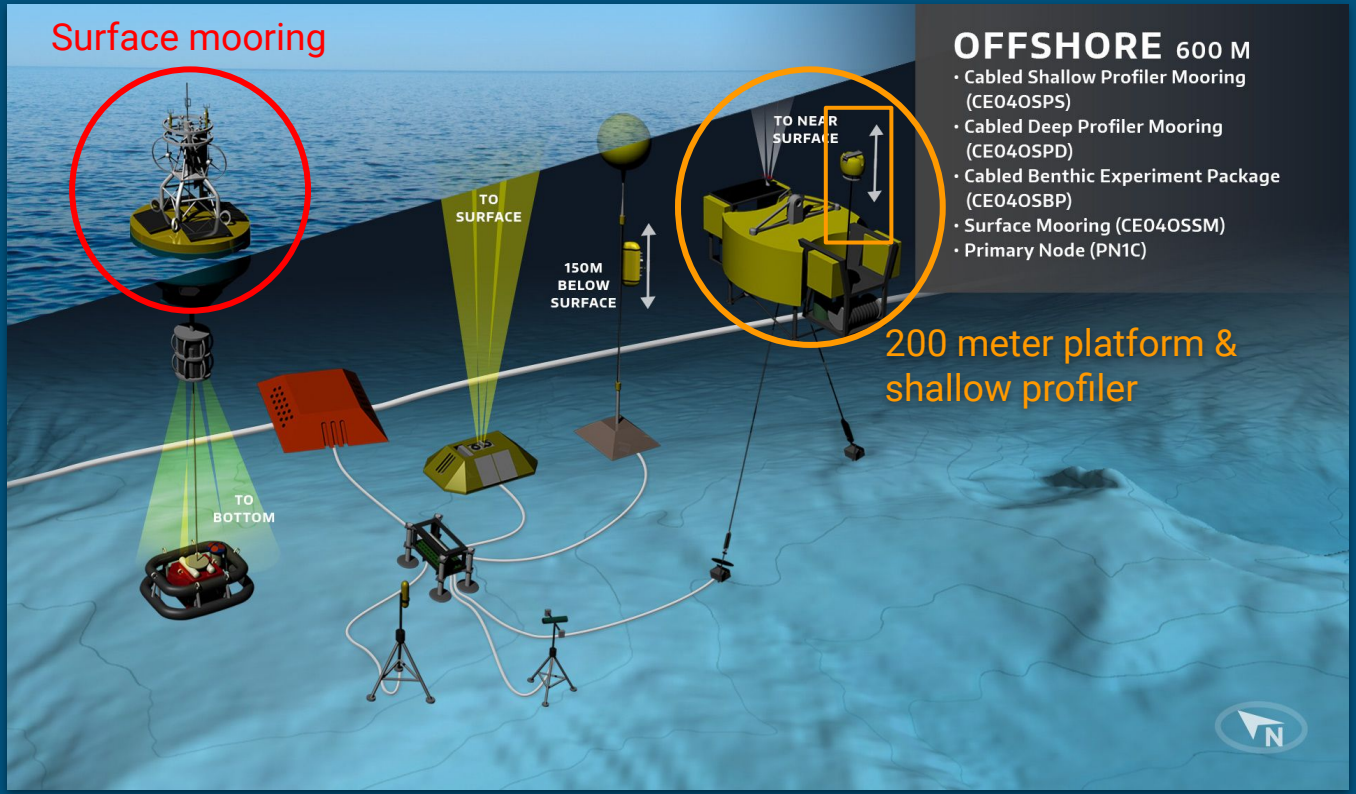
Source: Physical-biological coupling in Monterey Bay, California: Topographic influences on phytoplankton ecology

Upwelling carries nutrients from the deep sea to the surface and feeds the phytoplankton at the bottom of the food web. In turn, these phytoplankton feed the rest of the food web - the zooplankton, the fish, and us!

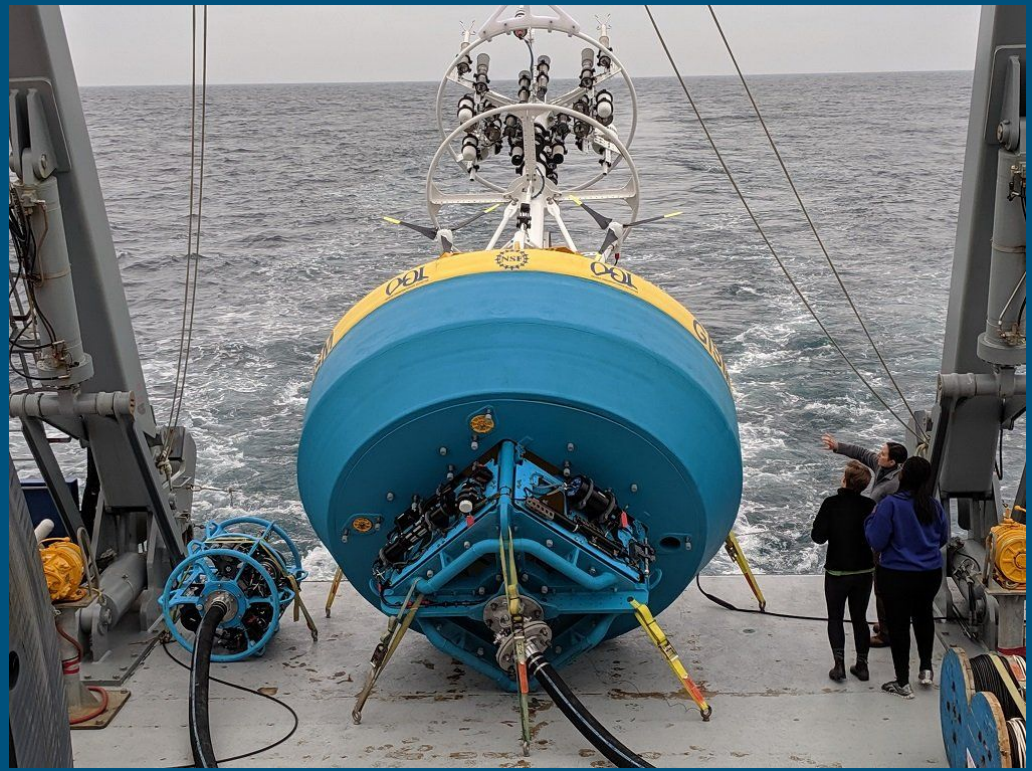
Ocean Observatories Initiative (OOI)

Source: https://oceanobservatories.org/research-arrays/

Oregon Offshore Site

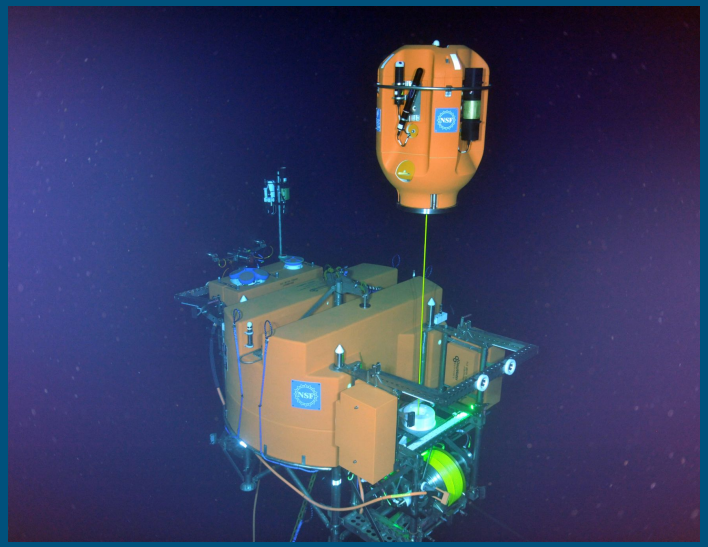


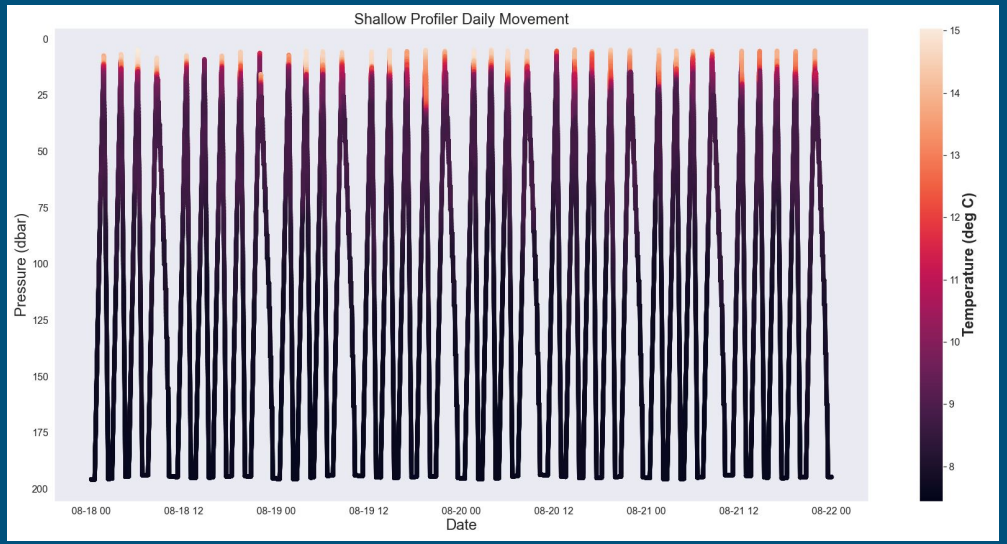
Surface Mooring: Air-Sea Interface



Left: https://oceanobservatories.org/2019/09/ooi-achieves-milestone-with-irminger-sea-deployment/
Right: https://twitter.com/HilaryPalevsky/status/1158122464263790592/photo/1

Shallow Profiler Daily Movement





Above: Ocean Observatories Initiative
Top right: https://interactiveoceans.washington.edu/platform-winch_-deck_-082914_145722/

Data Selection and Access

- Accessed data using OOI API
- Pulled data from the Oregon Offshore location, 600 m depth; ~40 miles west of Newport, Oregon, 1 of 6 sites in the Coastal Endurance array
- Three sensor types: bulk meteorology package on the surface mooring, CTD-O on the 200 meter platform, CTD-O on the shallow profiler

Exploratory Data Analysis

Coastal Upwelling Transport Index (CUTI)

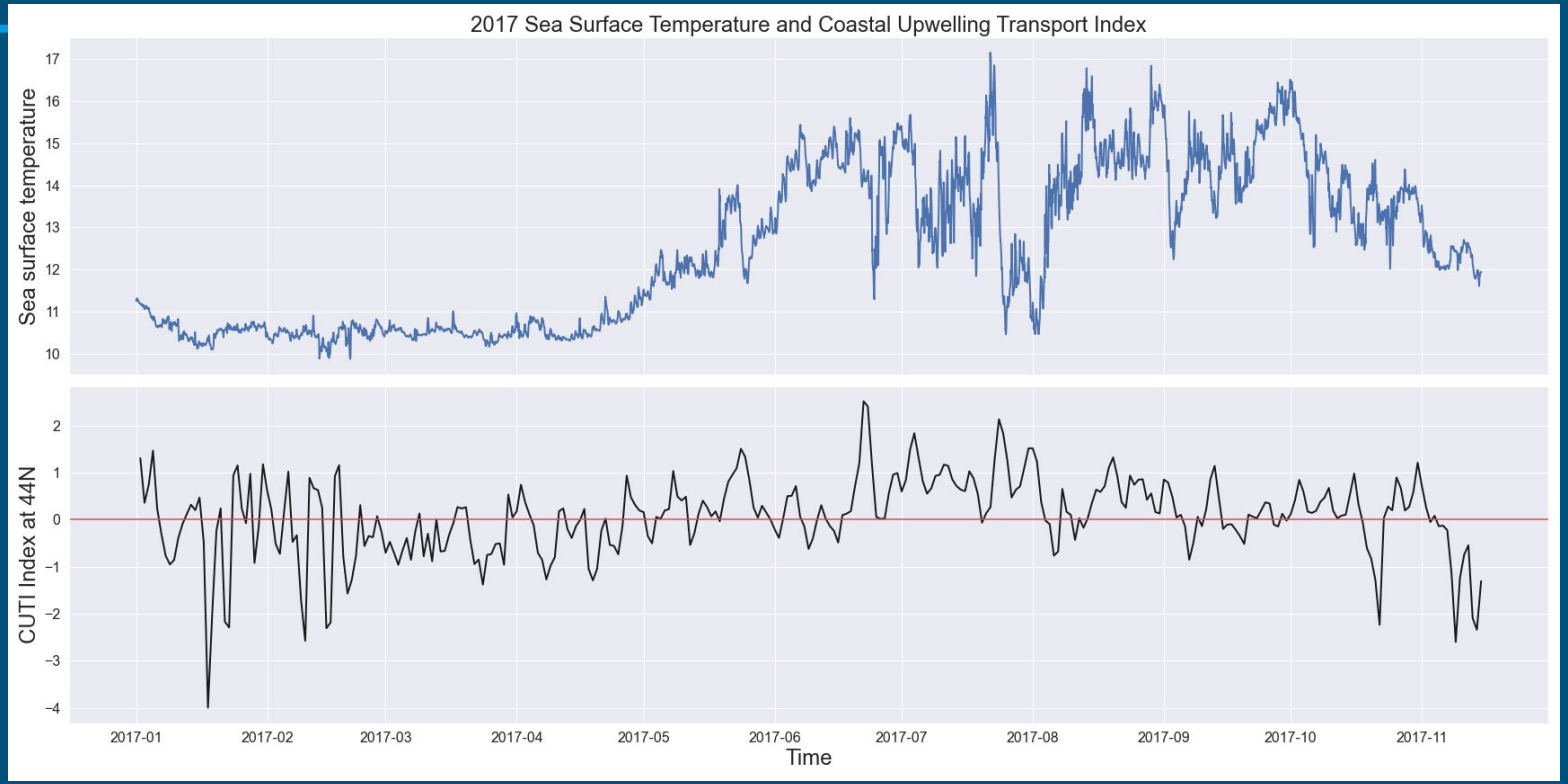
CUTI estimates total vertical transport through the base of the mixed layer (W) as the sum of two components: transport due to Ekman transport, and transport associated with geostrophic transport

$CUTI = U^{Ek} + U^{Geo}$

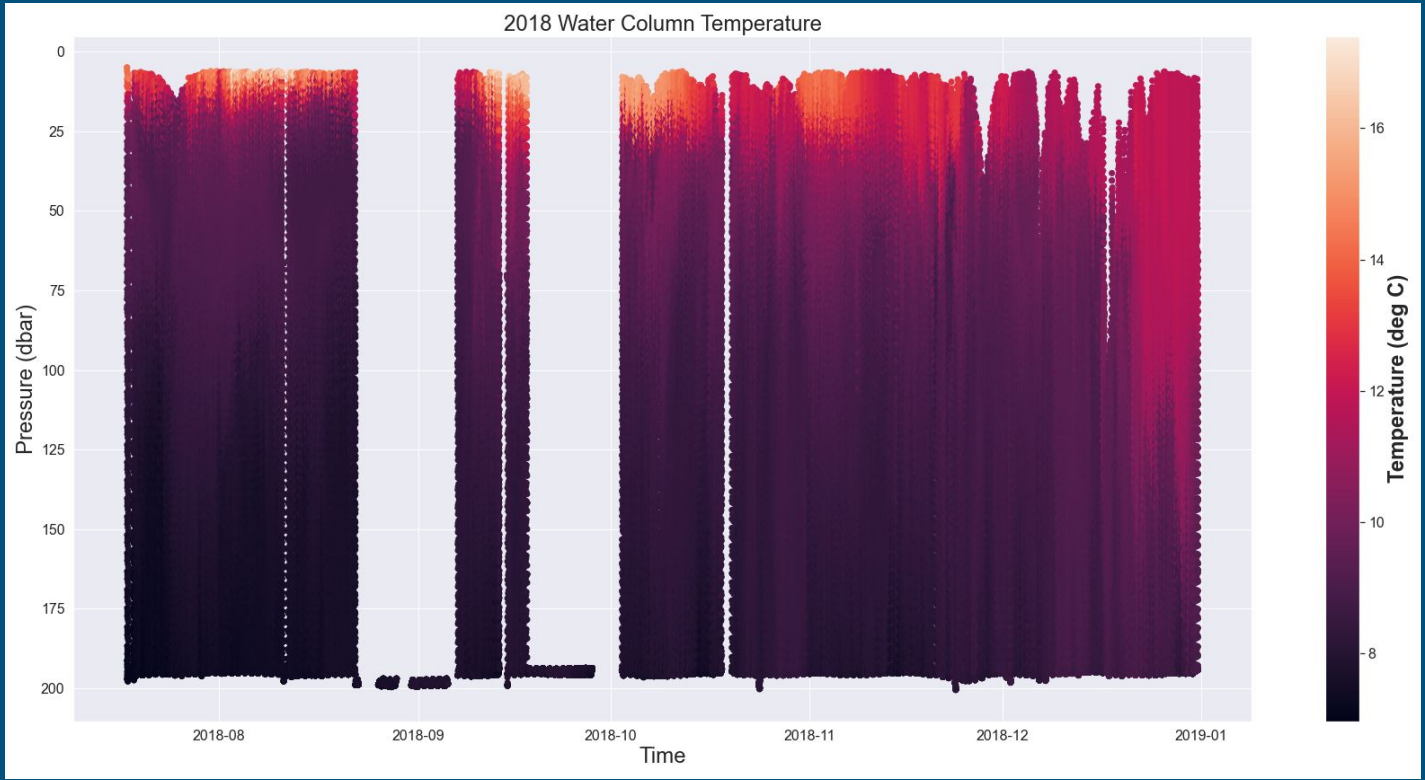
Positive values indicate upwelling, negative values indicate downwelling.

Source: https://oceanview.pfeg.noaa.gov/products/upwelling/cutibeuti

Sea Surface Temperature and CUTI



Temperature of the Water Column



EDA Findings

- SST is cold and stable in the winter, and warm and erratic in the summer
- The correlation between SST and CUTI is visible in the data
- The shallow profiler data from 2017 doesn’t provide a full view of the water column, but the 2018 data does.

Preprocessing Data Before Modeling

To turn this into a binary classification problem, I binarized the CUTI index to 1s for upwelling and 0s for not upwelling.

To identify and remove outliers, I used z-scores with a threshold of 3 standard deviations.

Modeling

Model Selection

Prioritized interpretability over accuracy, so I used a logistic regression classifier and a decision tree classifier.

Breaking one of the LINE assumptions of logistic regression: strong multicollinearity in model features!

Addressed this by adding interaction features and regularization to the logistic regression model.

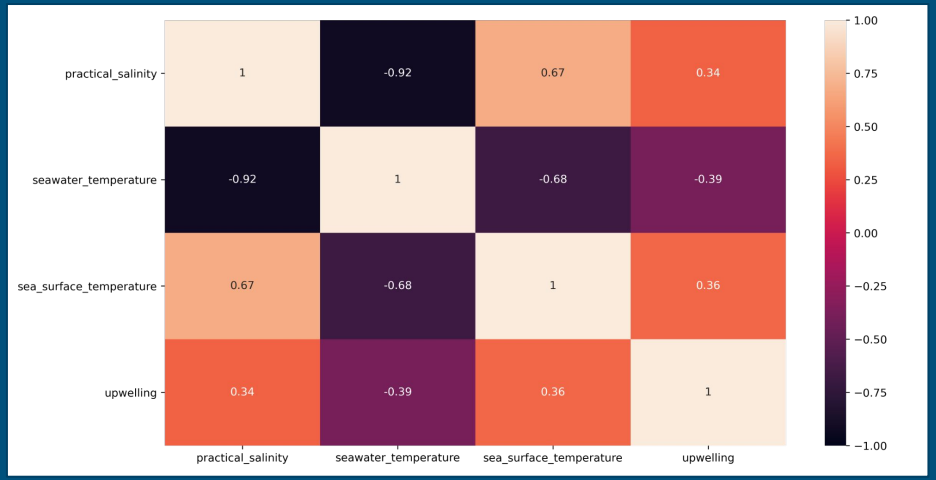
Feature and Target Variables

Selected features:

- Sea surface temperature
- Seawater salinity at 200 meters
- Seawater temperature at 200 meters

Target: Binarized upwelling index

Feature Correlation and Multicollinearity

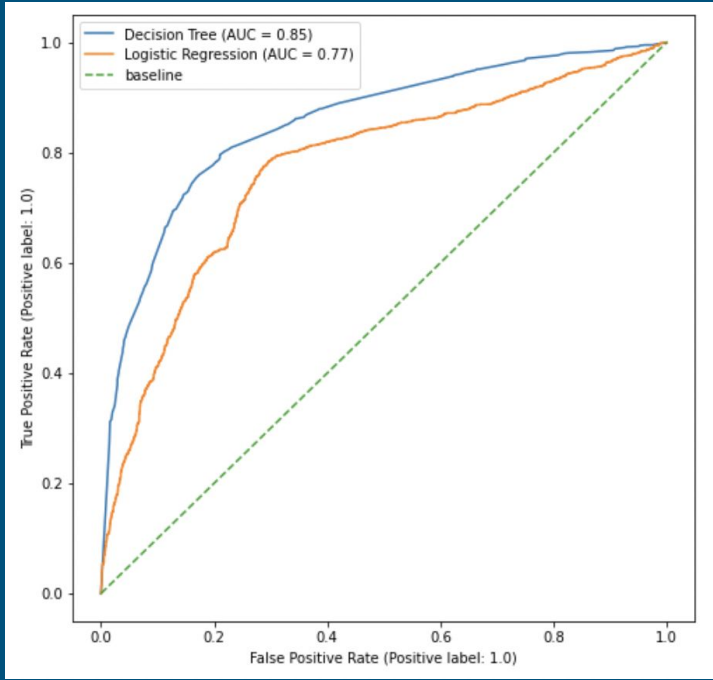


Features are correlated to each other, but they're independent of each other in nature.

Strongest correlation to upwelling is seawater temperature at 200 meters.

Seawater temperature and salinity are nearly 1:1 inversely correlated!

Model Comparison



Baseline accuracy: 61.95%

Logistic regression accuracy

Train: 74.64% Test: 75.34%

Decision tree accuracy

Train: 81.89% Test: 79.41%

In the AUC plot, the green dashed line represents no class separation capacity. The closer the AUC is to 1, the better the model is able to tell 0s from 1s.

Logistic Regression Classification

Baseline accuracy: 61.95%

Train accuracy: 74.64%

Test accuracy: 75.34%

Best parameters:

- C = 5.79
- Penalty = l2

Coefficients:

- seawater_temperature^2 = 1500469
- seawater_temperature * sea_surface_temperature = 38.73

Decision Tree Classification

Baseline accuracy: 61.95%

Train accuracy: 81.84%

Test accuracy: 79.43%

Best parameters:

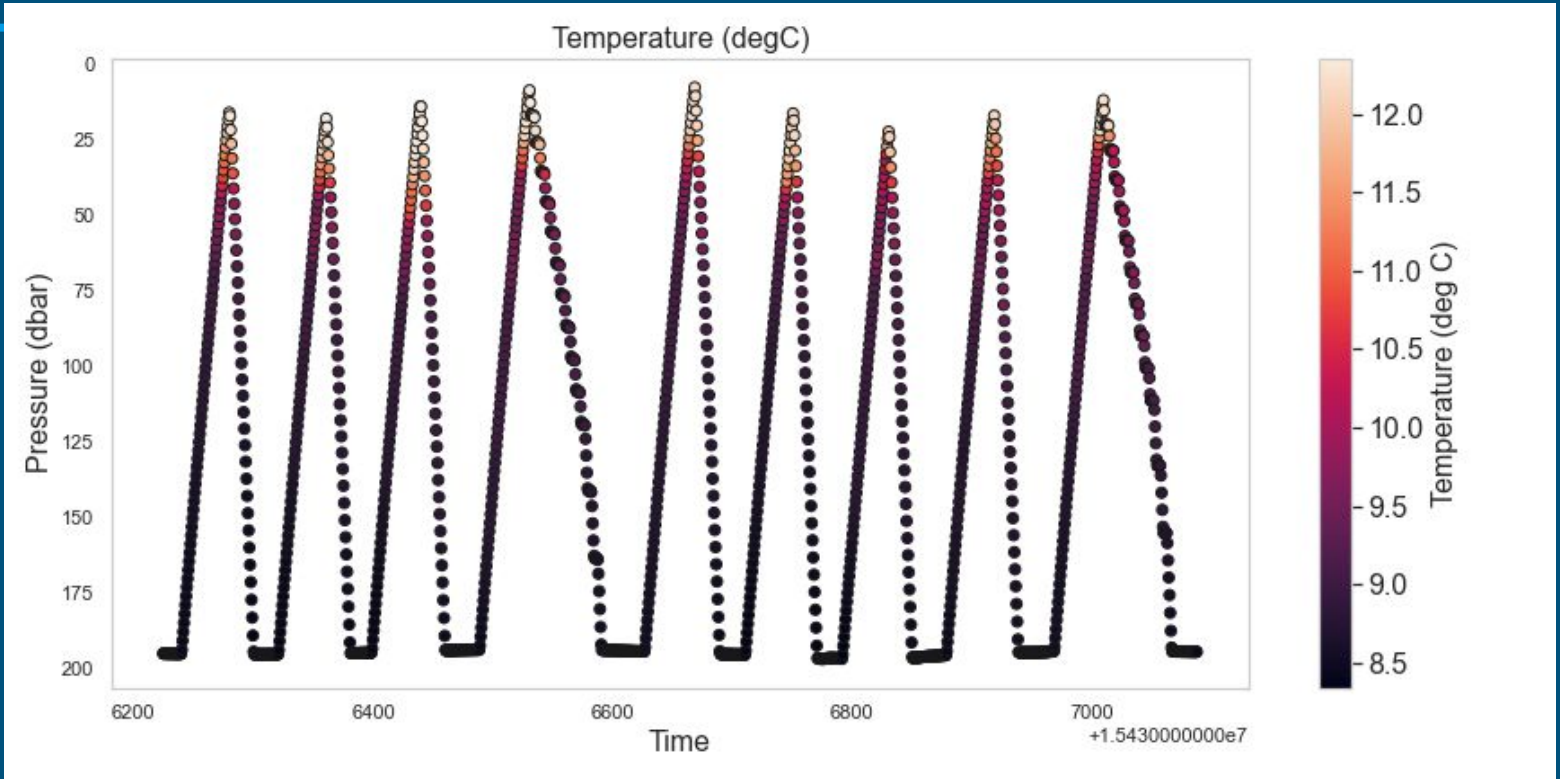
- Max_depth = 9
- Min_samples_leaf = 5
- Min_samples_split = 20

Feature importance:

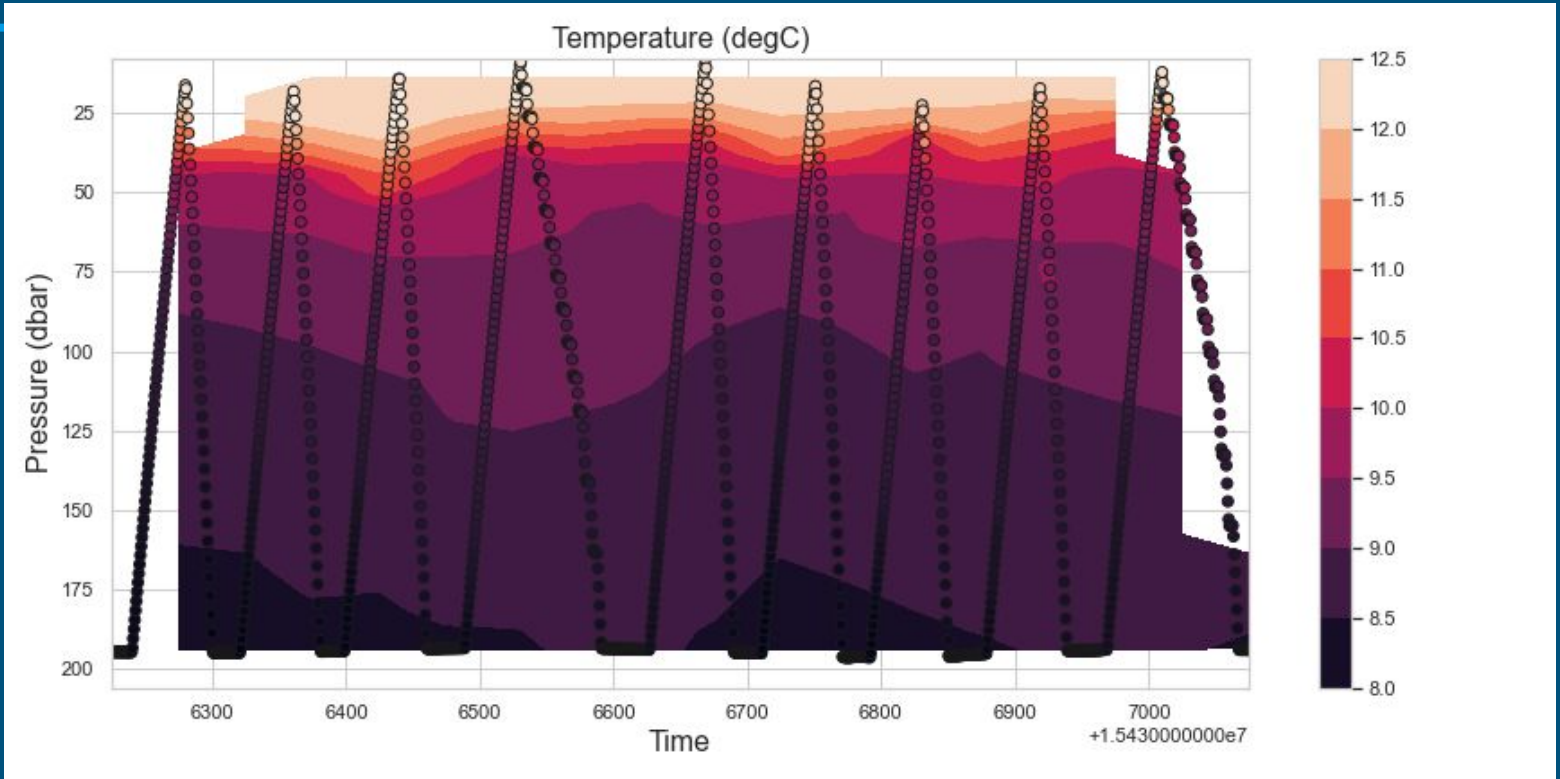
- Practical_salinity = 0.13799343
- Seawater_temperature = 0.14348944
- Sea_surface_temperature = 0.71851714

Interpolating Profiler Data to a Regular 2D Grid

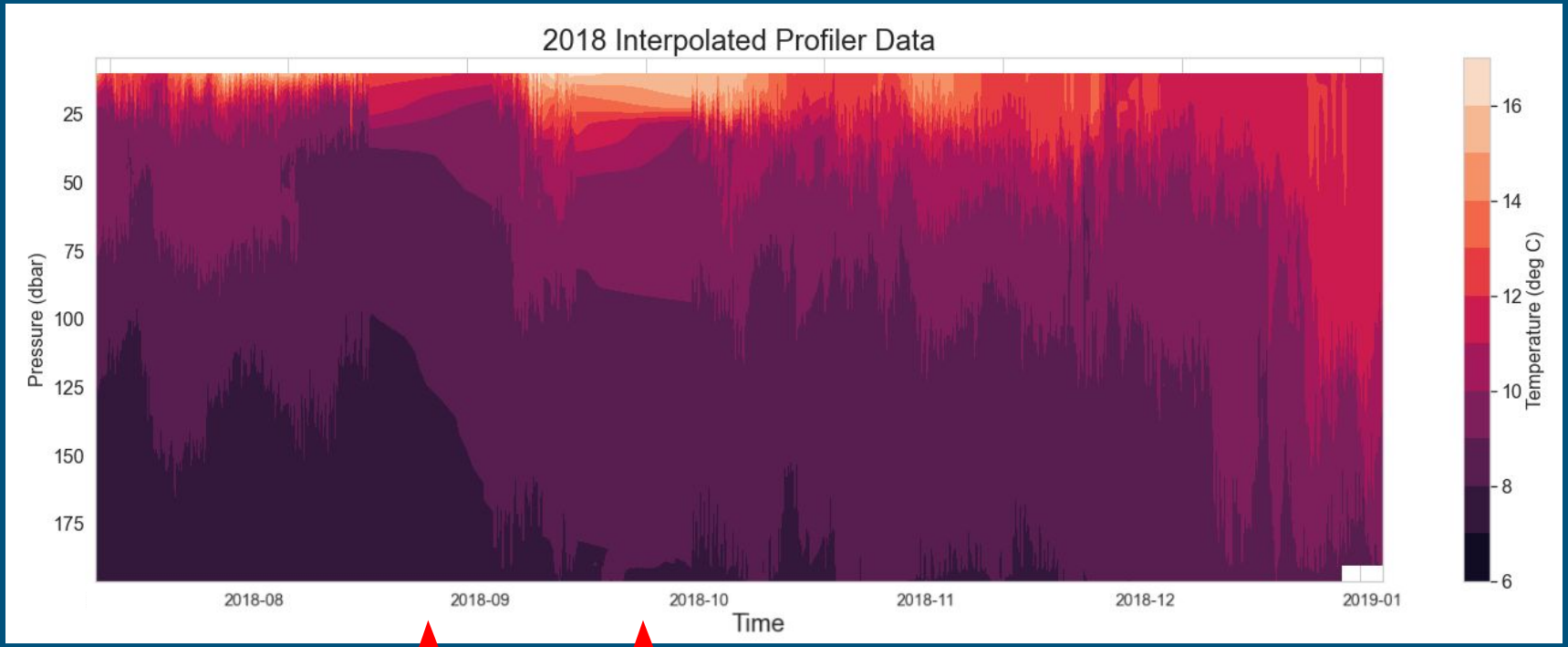
Raw Profiler Data



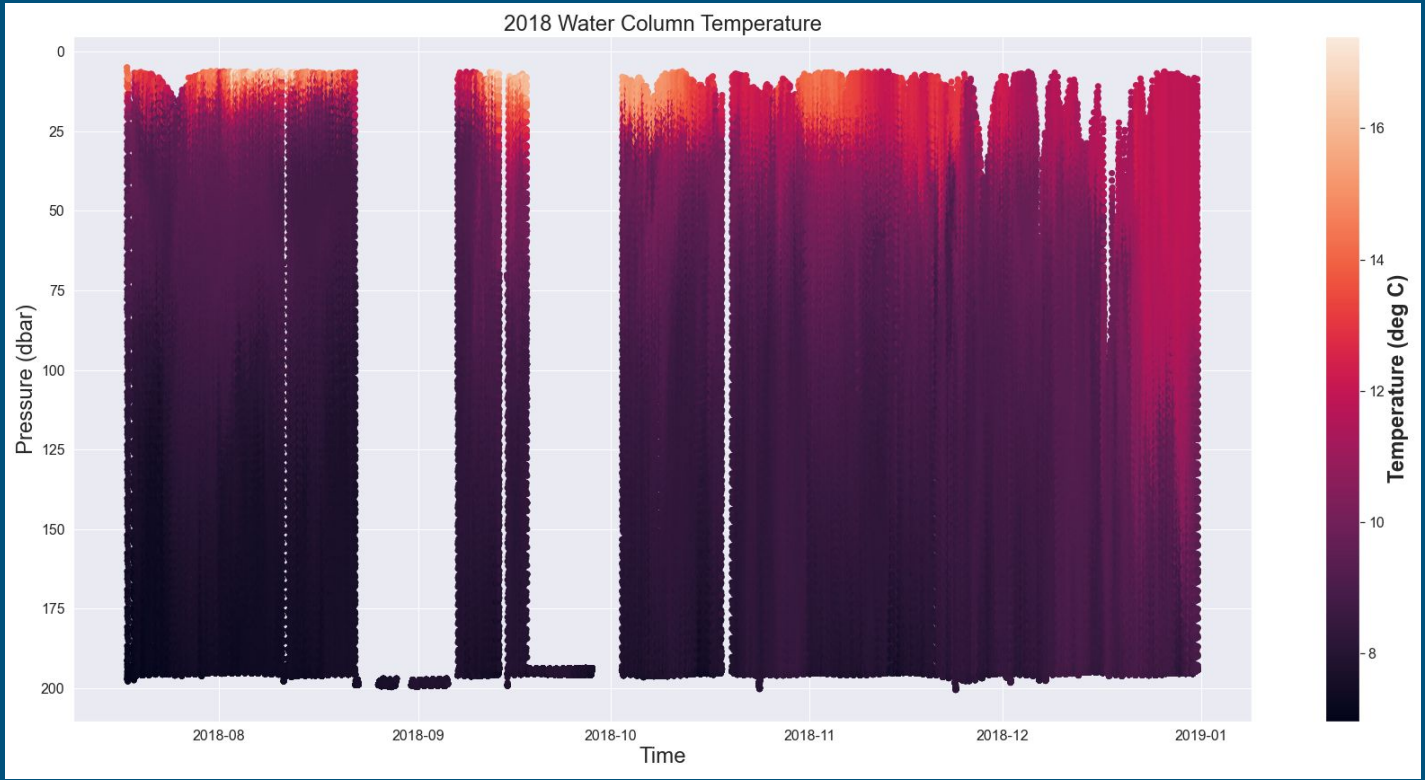
Scipy's griddata() - 1 Day



Scipy's griddata() - 2018



Temperature of the Water Column



The Point of Interpolating

Interpolating the data turned 1 column of irregularly spaced data into a functional 2D grid!

Each column is now a different depth, and each depth can be passed into the same models as before to see if the models are more informed by this full view of the water column.

Depth →

Time ↓

10	16	22	28	34	40	46	52	58	...
14.050275	11.959311	10.073221	9.767676	9.482159	9.105380	9.056556	9.055258	9.038180	...
13.957127	10.870573	9.901940	9.695652	9.342295	9.117574	9.069495	9.021141	9.080058	...
14.042208	10.949701	9.850174	9.776190	9.300516	9.257225	9.193214	9.130008	9.150563	...
14.140387	11.095460	9.932039	9.728931	9.258736	9.346245	9.331689	9.174076	9.217667	...
14.116346	11.205226	9.967389	9.708485	9.617728	9.247863	9.456468	9.246932	9.266785	...
...	...	...	...	...	...	...	...	...	...
11.958360	11.958135	11.958165	11.952446	11.940669	11.822778	11.618385	11.532611	11.528637	...
11.966183	11.966942	11.967257	11.958619	11.929104	11.848490	11.649076	11.534315	11.527931	...
11.974130	11.973730	11.974008	11.968621	11.933196	11.913816	11.740736	11.675417	11.565127	...
11.982202	11.979041	11.978485	11.968747	11.949102	11.948089	11.841979	11.765422	11.649837	...
11.990295	11.986075	11.982824	11.968874	11.965074	11.965405	11.936267	11.844511	11.732570	...

Logistic Regression Classification

Baseline accuracy: 60.42%

Train accuracy: 75.23%

Test accuracy: 76.36%

Best parameters:

- C = 0.778
- Penalty = l2

Greatest coefficients:

- 88 meters: 3.23
- 190 meters: 2.90
- 106 meters: 2.695
- 100 meters: 2.03

Decision Tree Classification

Baseline accuracy: 60.42%

Train accuracy: 91.15%

Test accuracy: 90.40%

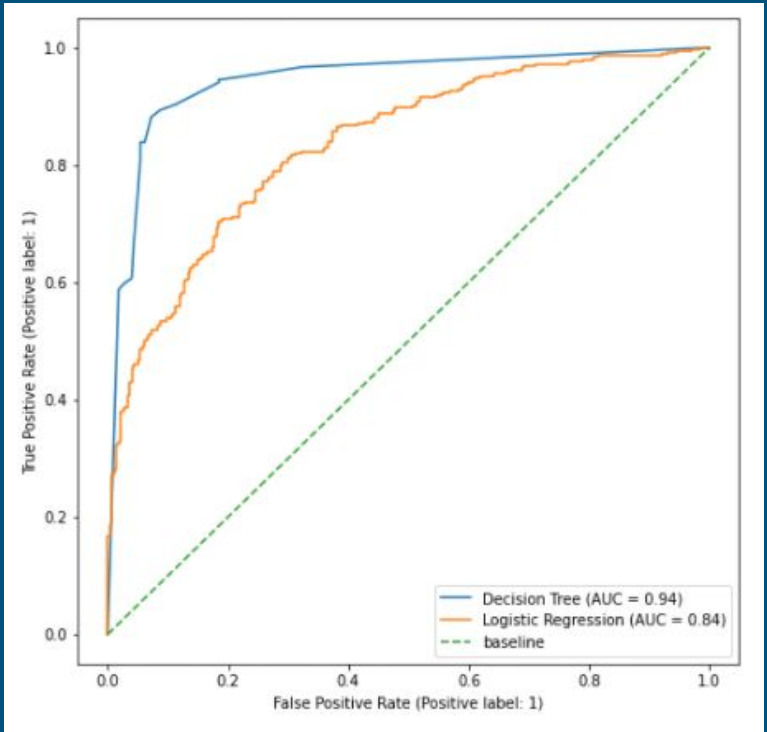
Best parameters:

- Max_depth = 9
- Min_samples_leaf = 2
- Min_samples_split = 5

Feature importance:

- 136 meters: 0.3069
- 196 meters: 0.1253
- 10 meters: 0.1202
- 40 meters: 0.0849

Model Comparison



Baseline accuracy: 60.42%

Logistic regression accuracy

Train: 75.23% Test: 76.36%

Decision tree accuracy

Train: 91.15% Test: 90.40%

Future Work and Additions

- Scale up: add more variables, or perform the same or similar analysis at different locations
- Turn it into a multiclass problem - I did this, it was bad
- Turn it into a regression problem to predict the CUTI index using wind or environmental variables, because everything should be proportional
- Use other classification models (will lose interpretability)
- Perform PCA on interpolated profiler data (will lose interpretability)
- Apply unsupervised clustering methods to T-S plots (see Extra Content slides)

Conclusions

Conclusions

As oceanographers, we know processes like upwelling and downwelling are happening because we study ocean physics, but it's exciting to apply analytical methods to the data and see how much can be done automatically!

Using only information from two locations in the water column, the models were able to beat the baseline accuracy.

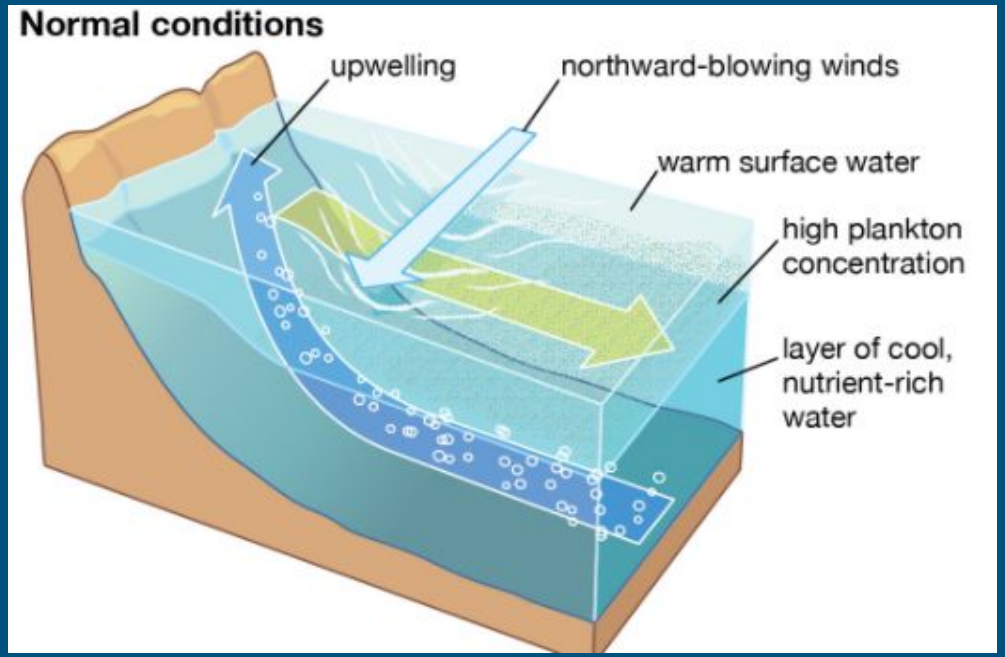
Letting the models make predictions based on the full view of the water column significantly improved accuracy,

Extra Content

What is Coastal Upwelling?

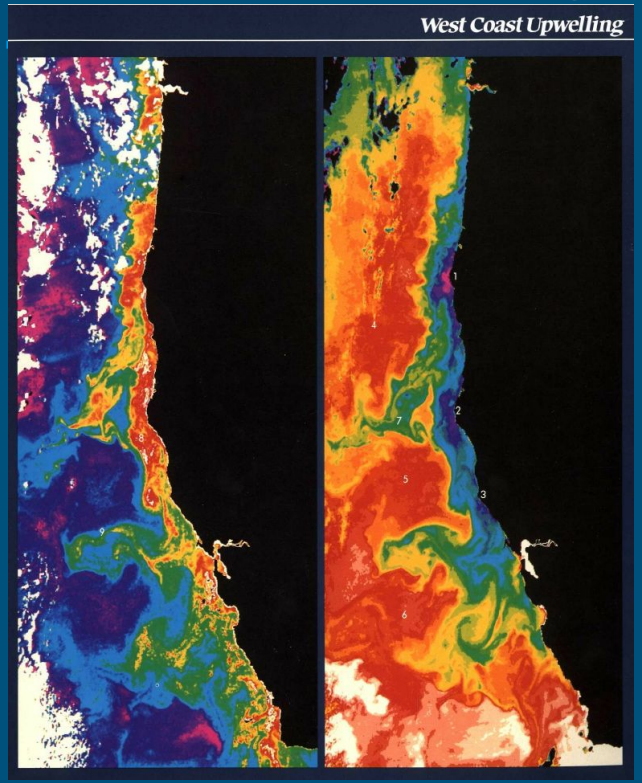
Winds blowing parallel to coastlines push surface water away from shore. Deep water is pulled up to the surface to replace it - this is called upwelling!

Downwelling describes the opposite - surface water is pushed towards shore and sinks when it reaches land.

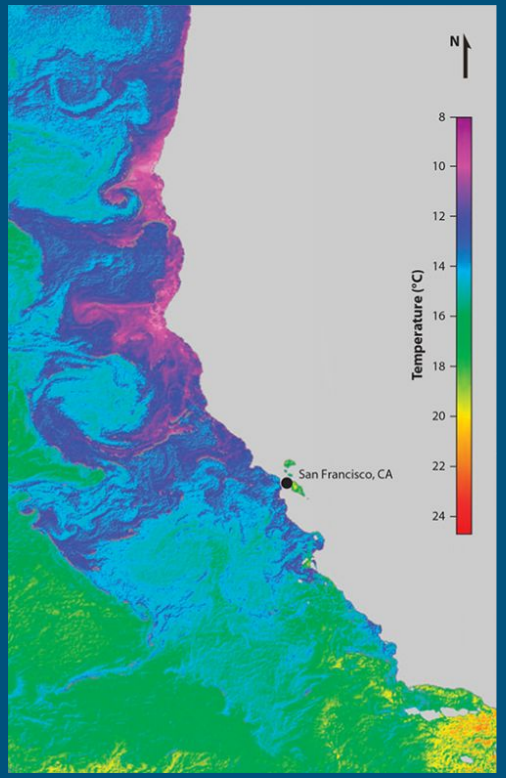


Source: https://www.britannica.com/science/upwelling#/media/1/618923/2739

More upwelling images

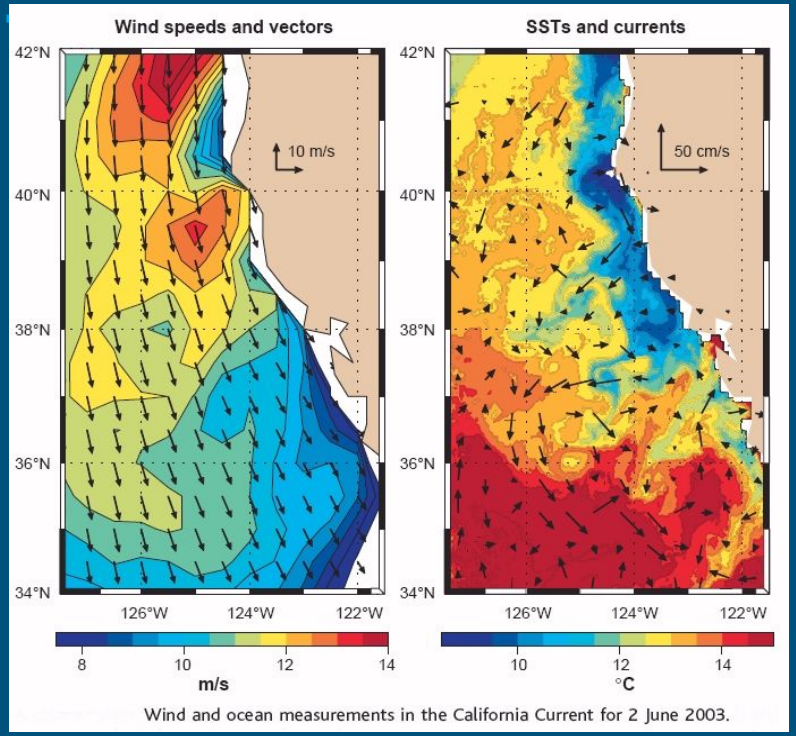


Source: https://www.researchgate.net/figure/Satellite-ocean-color-and-temperature-maps-right-showing-the-upwelling-areas-and_fig1_282646192



Source: 2011 publication by Eric Sanford and Morgan W. Kelly shows how coastal upwelling plumes (shown in purple) create a mosaic of variations in water temperature, nutrients, pH, and other parameters off the coast of Northern California.

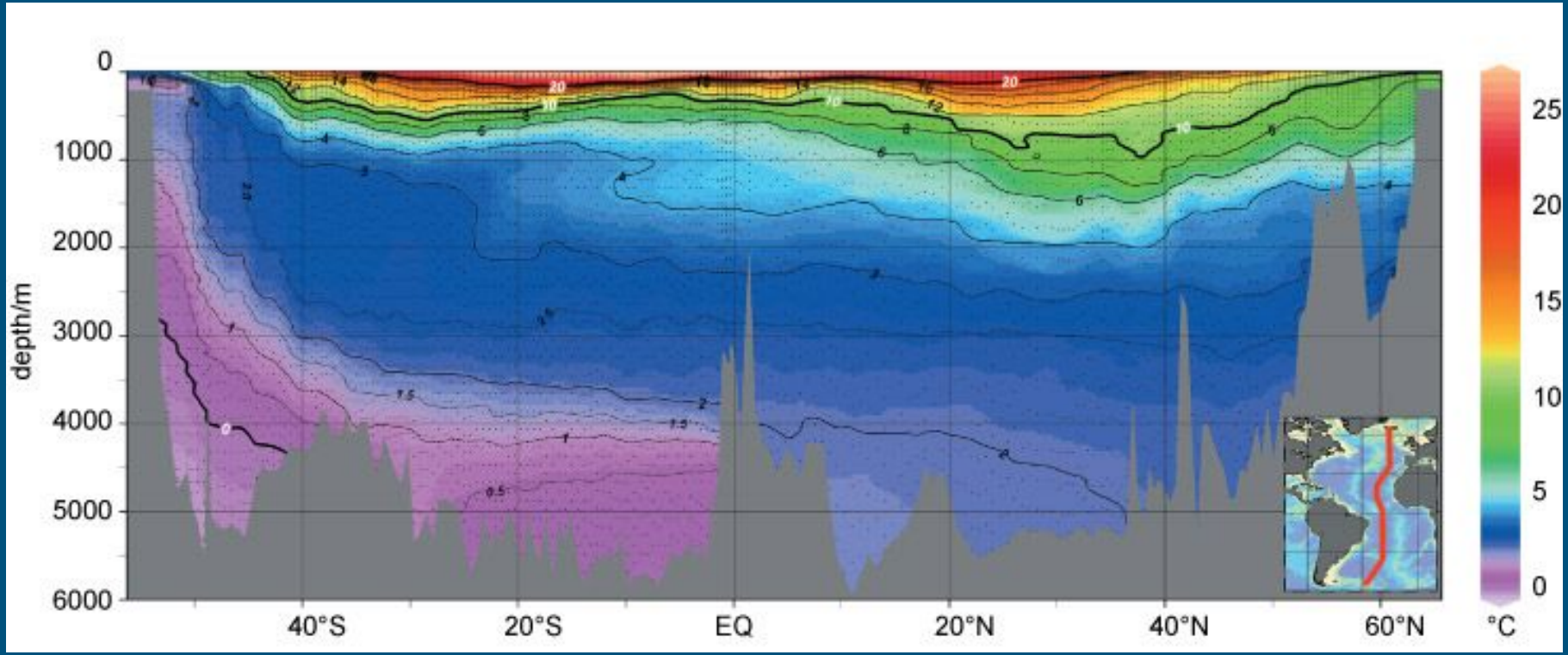
Upwelling is a Wind-Driven Process



Source: https://cimss.ssec.wisc.edu/sage/oceanography/lesson1/activity2.html

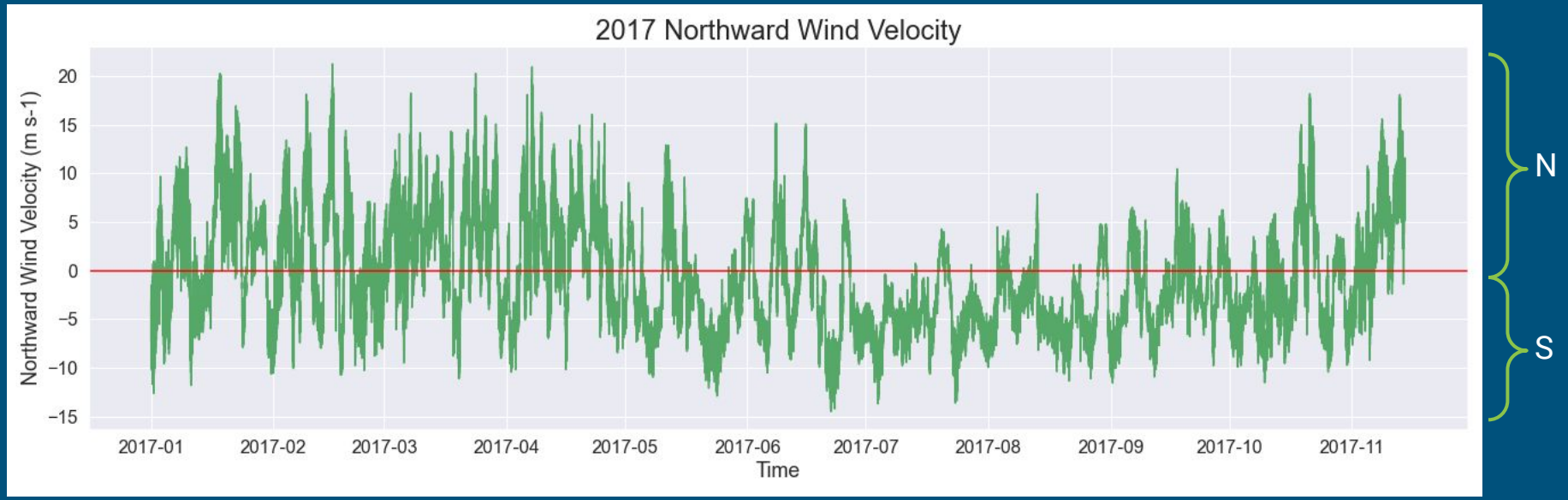
Based on alongshore wind stress characteristics in central and north California, three seasons are defined: Upwelling Season (April-June) with strong upwelling-favorable winds and large standard deviation due to frequent reversals; Relaxation Season (July-September) with weak equatorward winds and low variability; and Storm Season (December-February) characterized by weak mean wind stress but large variability. Source

Ocean structure



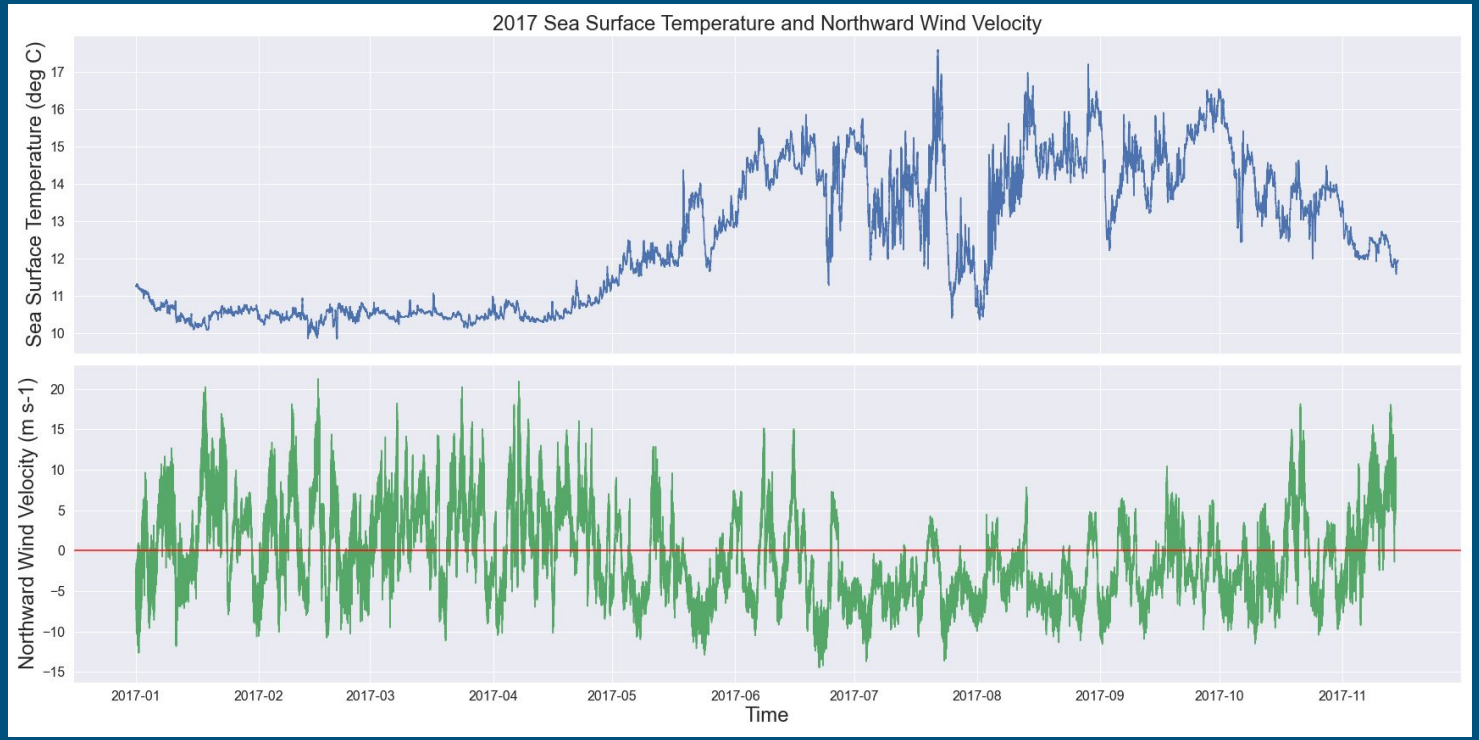
Source: https://www.open.edu/openlearn/ocw/mod/oucontent/view.php?id=20129&extra=thumbnailfigure_idm46317812507168

Surface Mooring Data - Wind Velocity

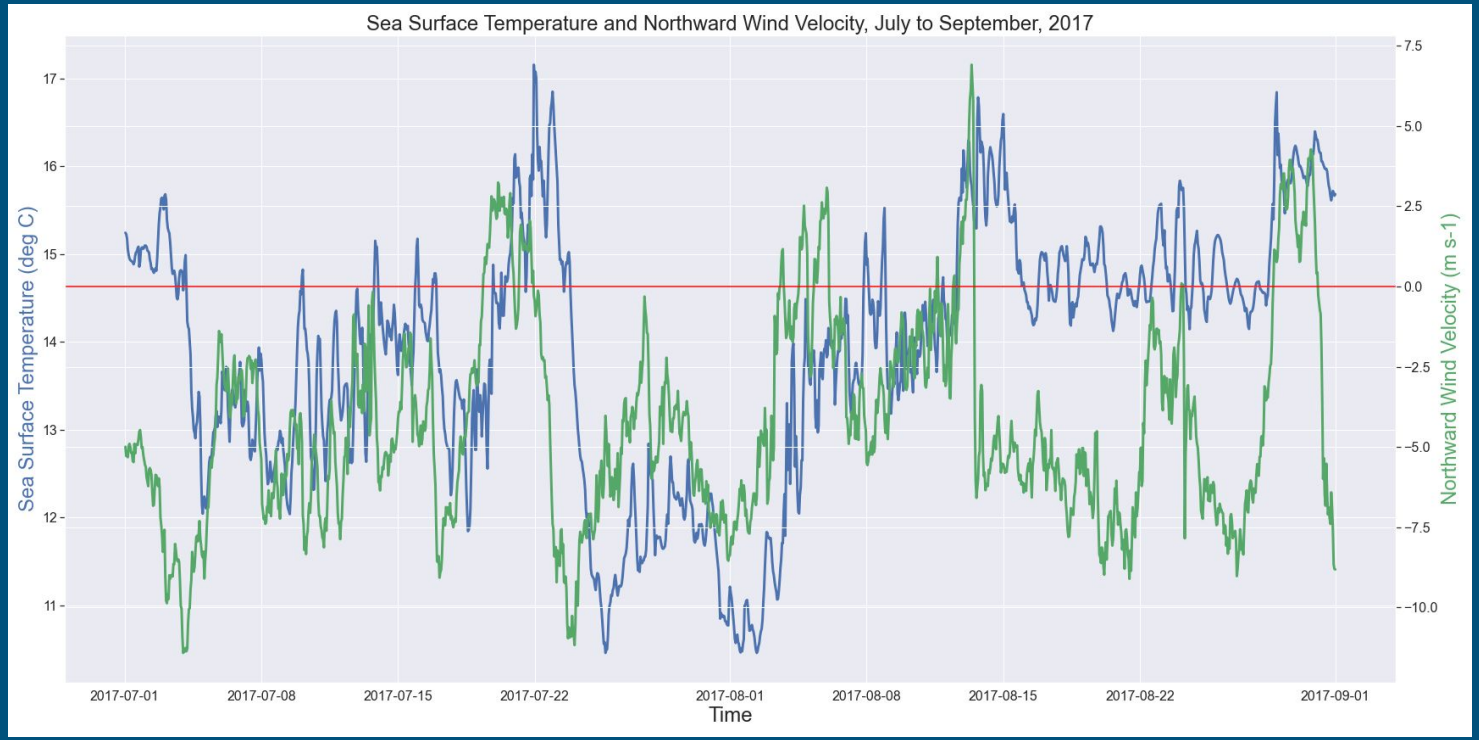


Winds were blowing from North to South for about 58% of the year in 2017, and 63% of the year in 2018.

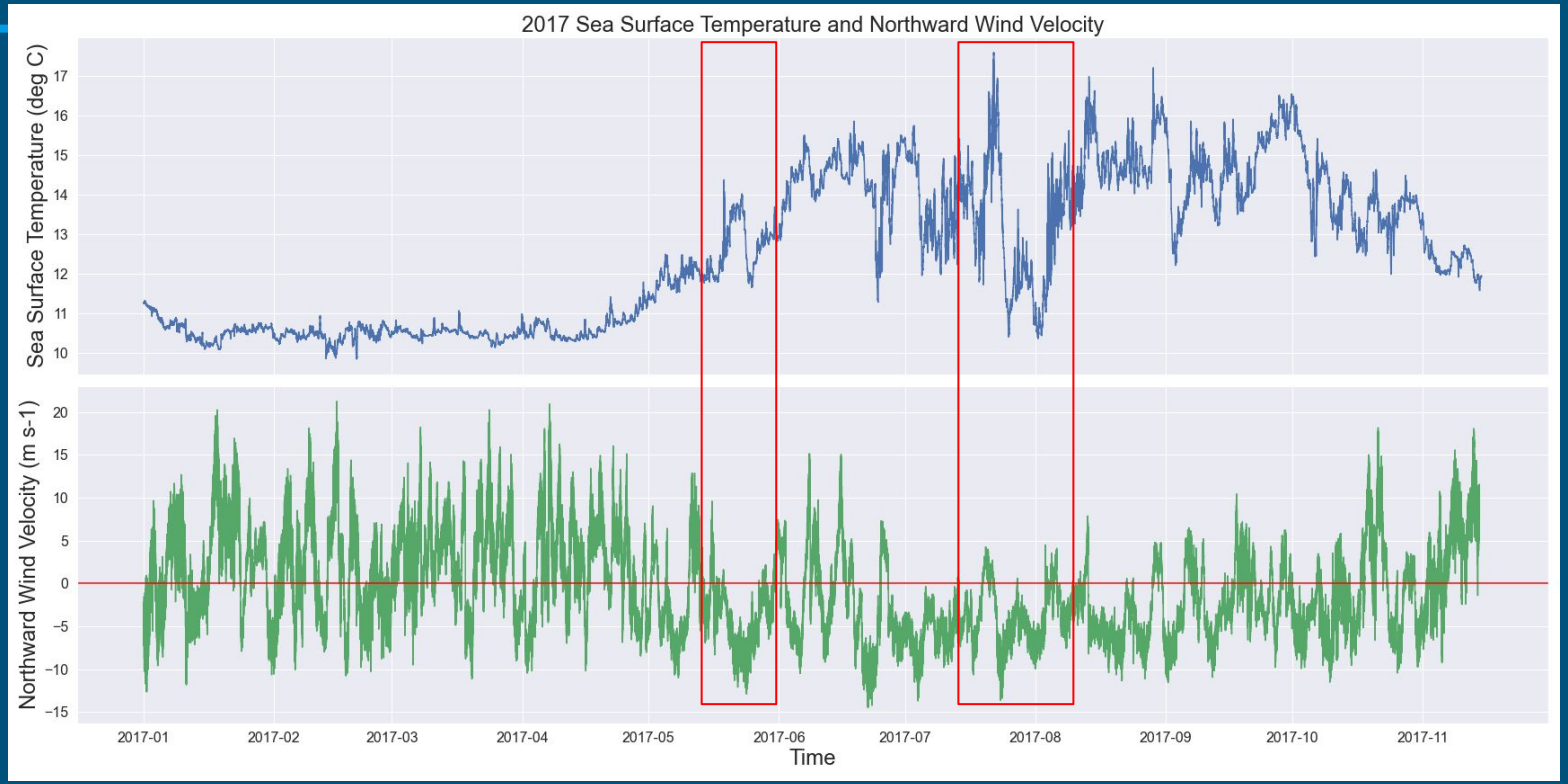
Surface Mooring Data - Wind-Driven Process



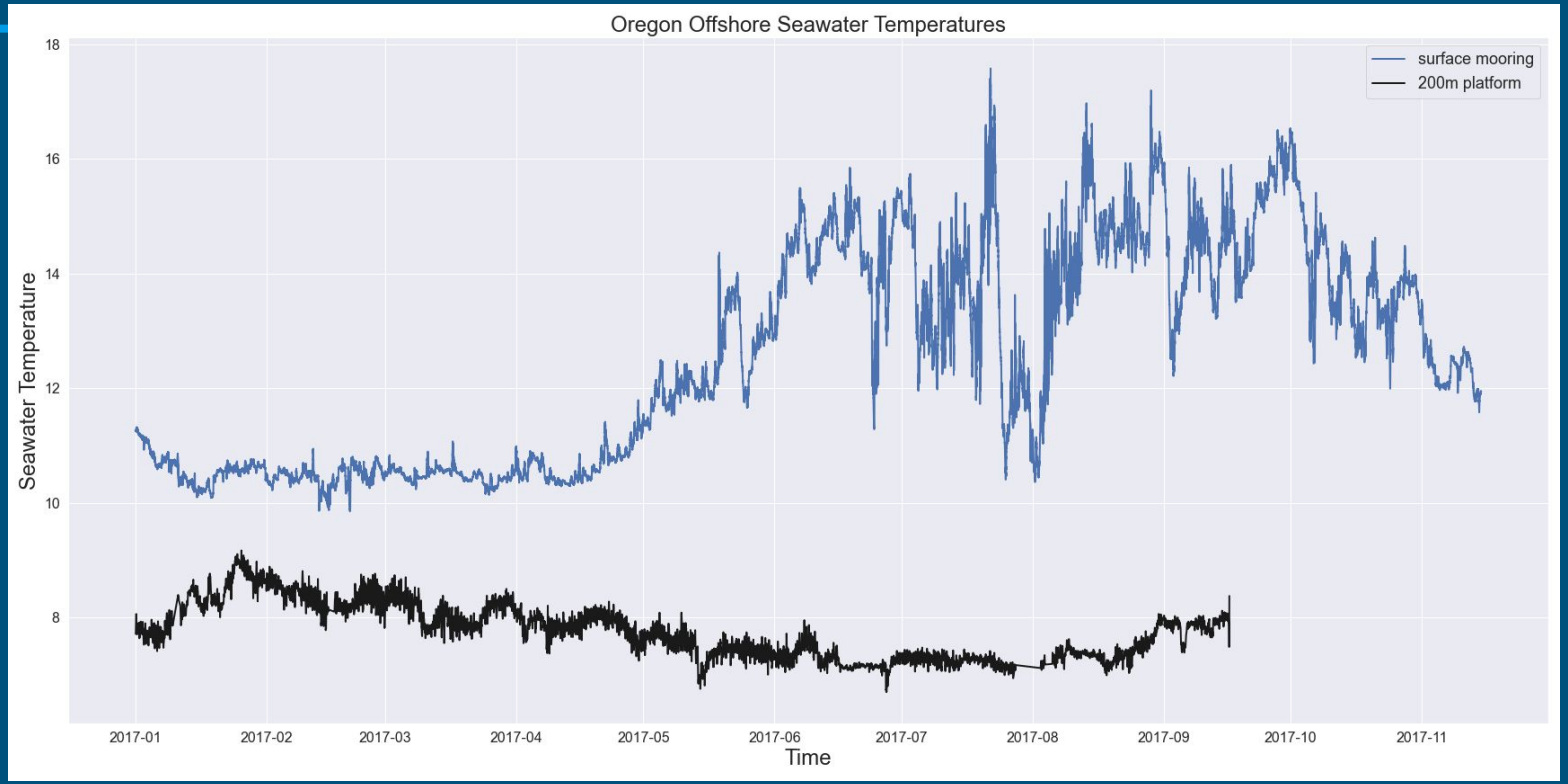
SST Follows Northward Wind Velocity



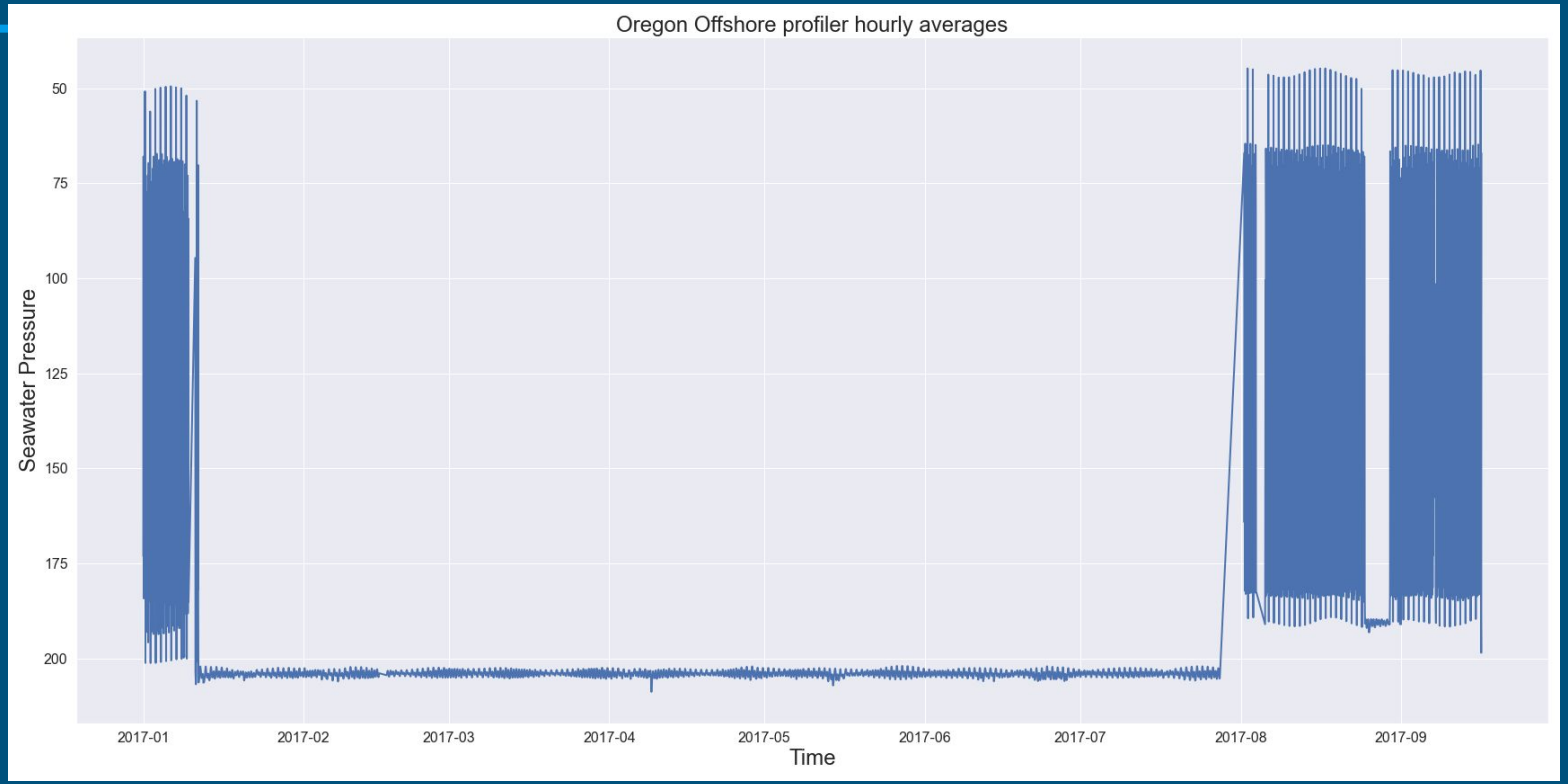
Surface Mooring Data - Wind-Driven Process



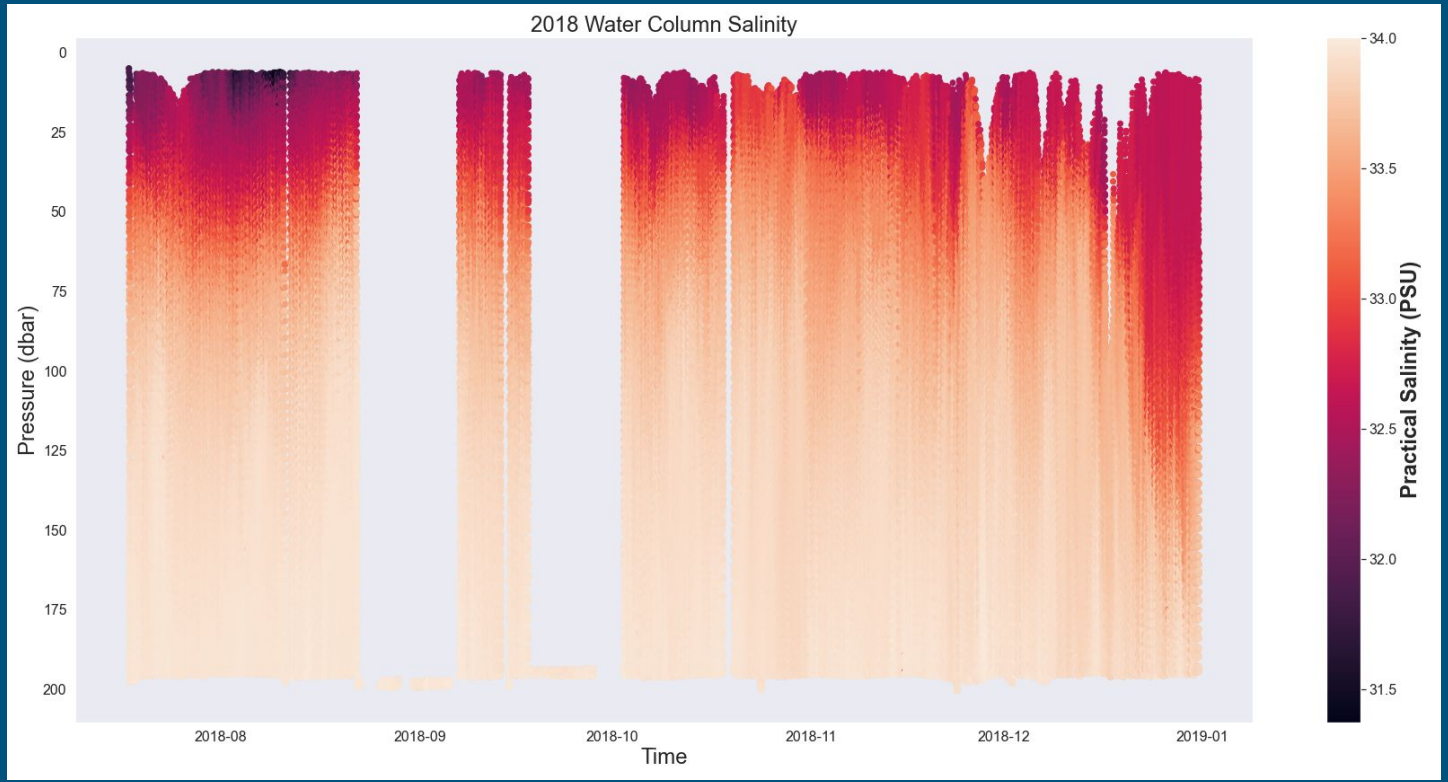
200m Platform - Upwelling of Cold Water



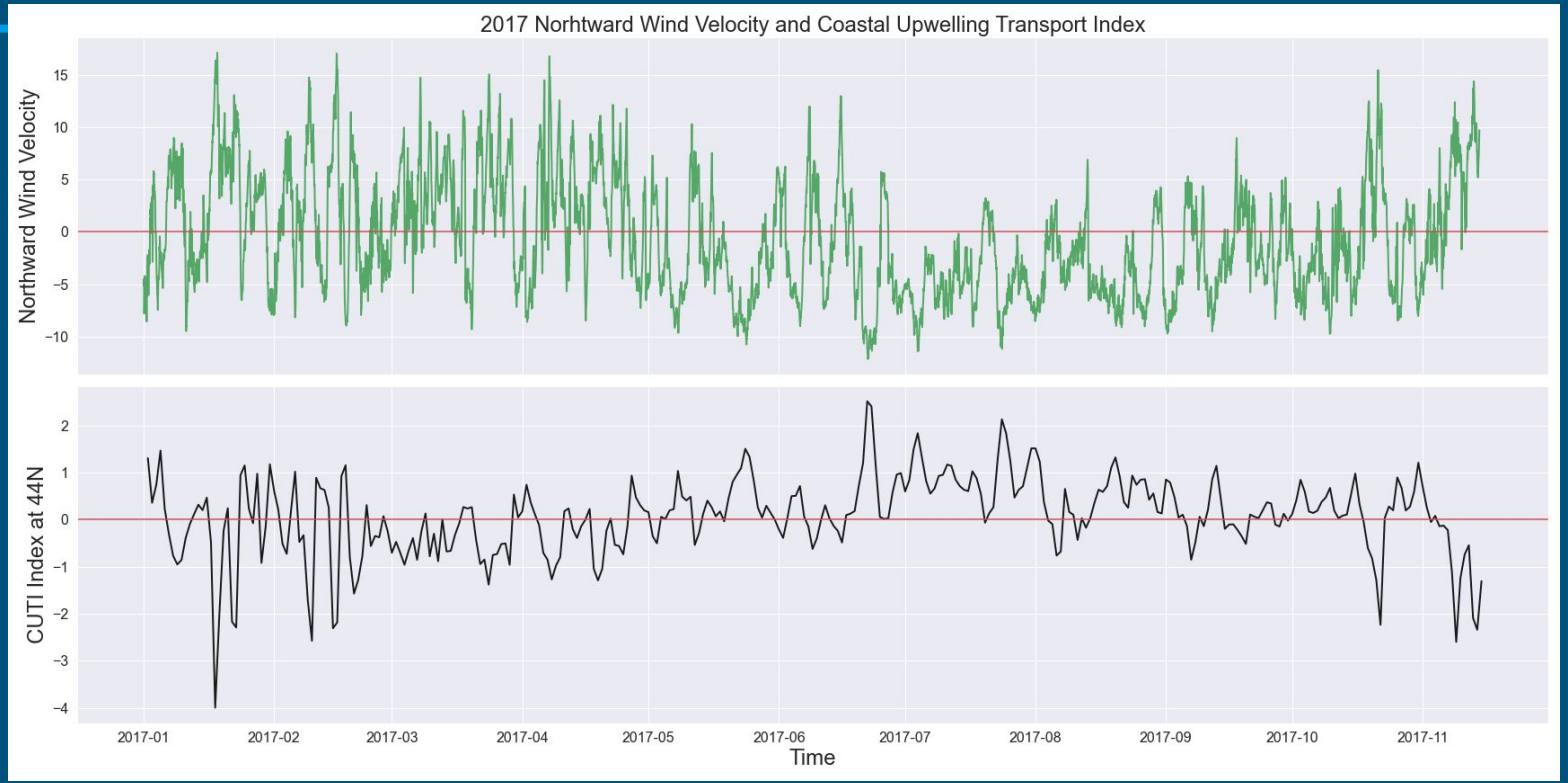
2017 Profiler Data Availability Issues



Salinity of the Water Column

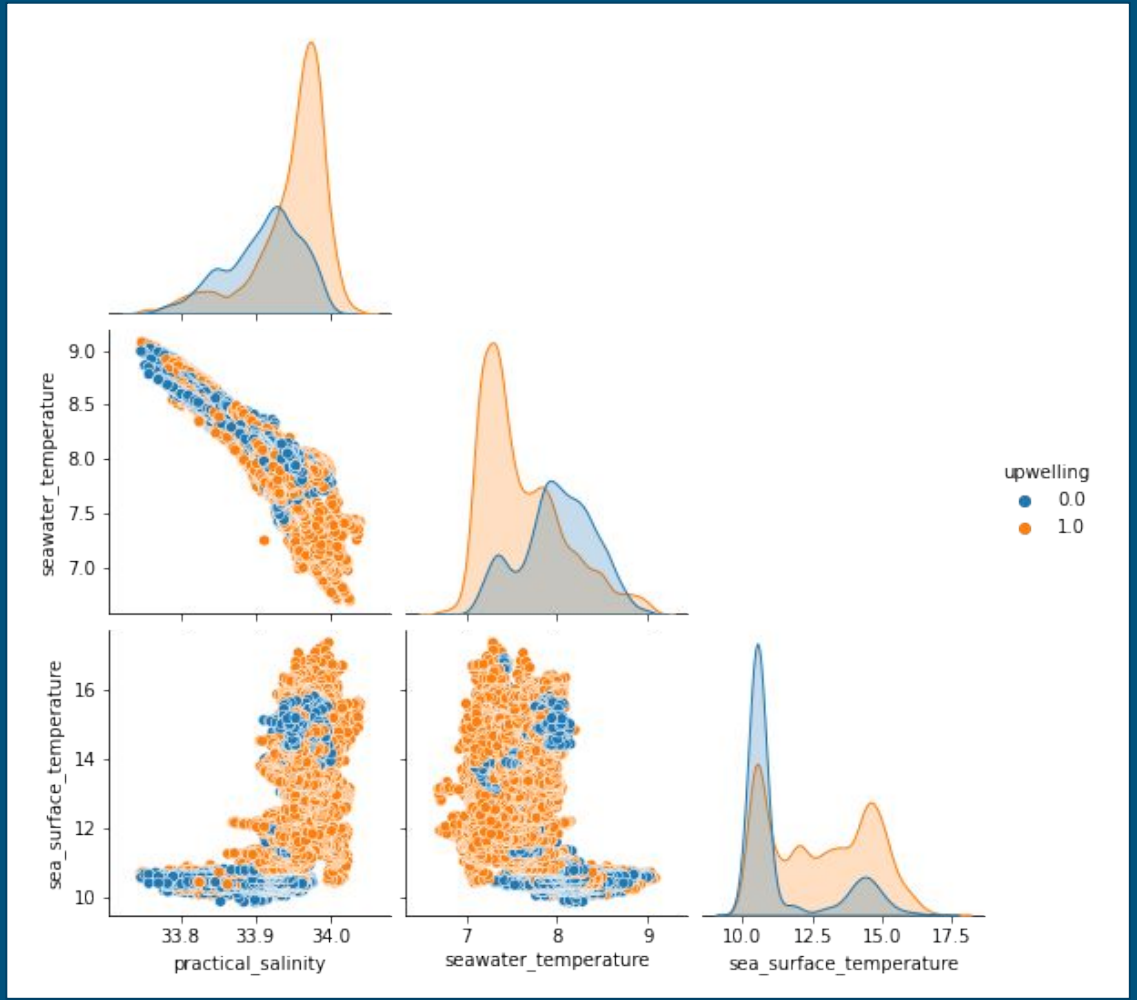


CUTI and Northward Wind Velocity

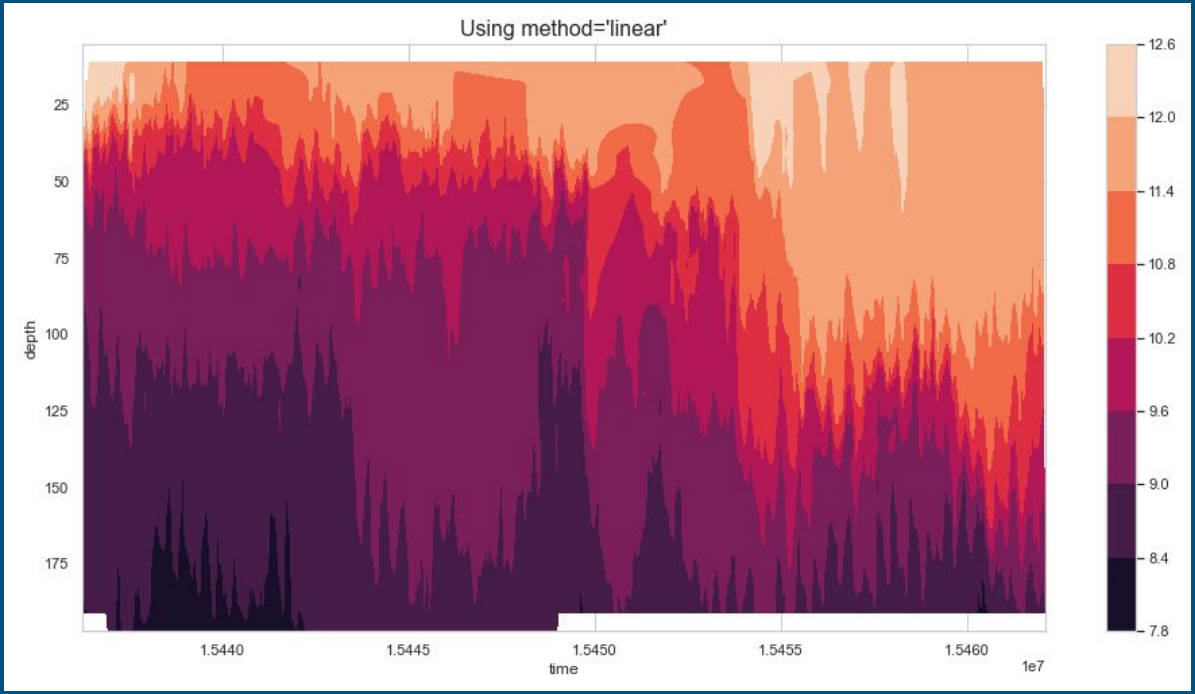


Misclassified Data

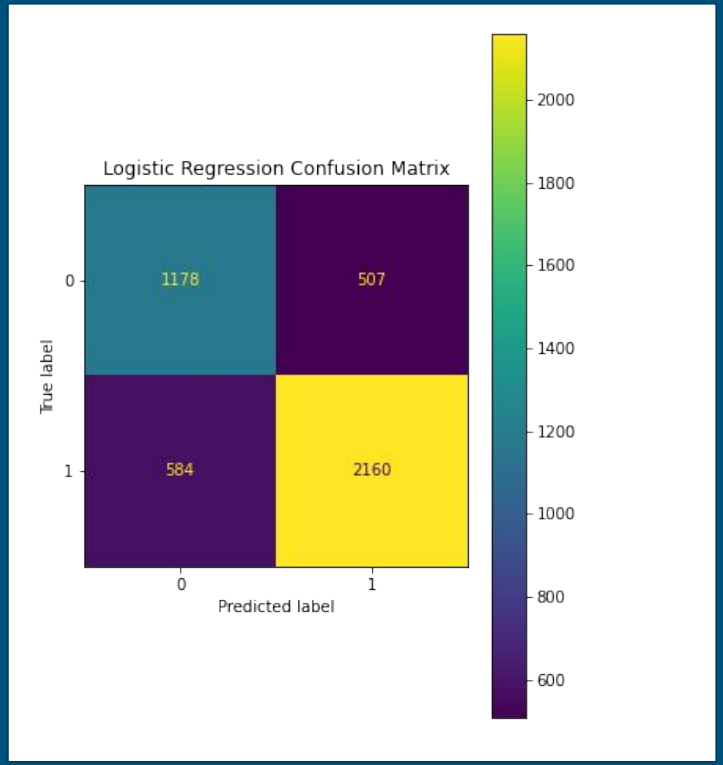
Pairplot

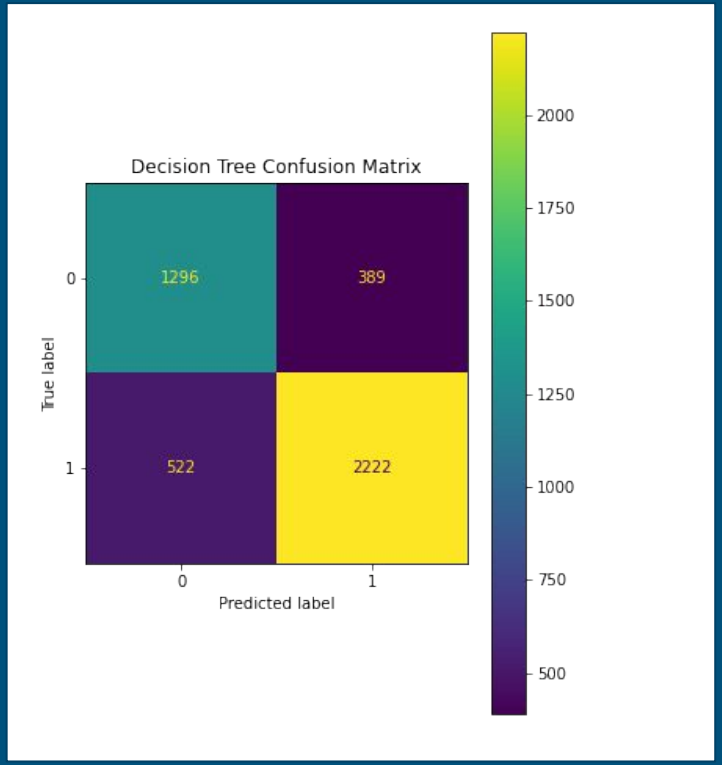

seawater_temperature
9.0
8.5
8.0
7.5
7.0
sea_surface_temperature
16
14
12
10
33.8
33.9
34.0
practical_salinity
7
8
9
seawater_temperature
10.0
12.5
15.0
17.5
sea_surface_temperature
upwelling
0.0
1.0

Scipy’s griddata() - 1 Month



Confusion Matrices





```
print(classification_report(y_test, gs_dt.predict(X_test)))
```

	precision	recall	f1-score	support
0.0	0.71	0.77	0.74	1685
1.0	0.85	0.81	0.83	2744
accuracy			0.79	4429
macro avg	0.78	0.79	0.78	4429
weighted avg	0.80	0.79	0.80	4429

```
print(classification_report(y_test, gs_lr_test_preds))
```

	precision	recall	f1-score	support
0.0	0.67	0.70	0.68	1685
1.0	0.81	0.79	0.80	2744
accuracy			0.75	4429
macro avg	0.74	0.74	0.74	4429
weighted avg	0.76	0.75	0.75	4429