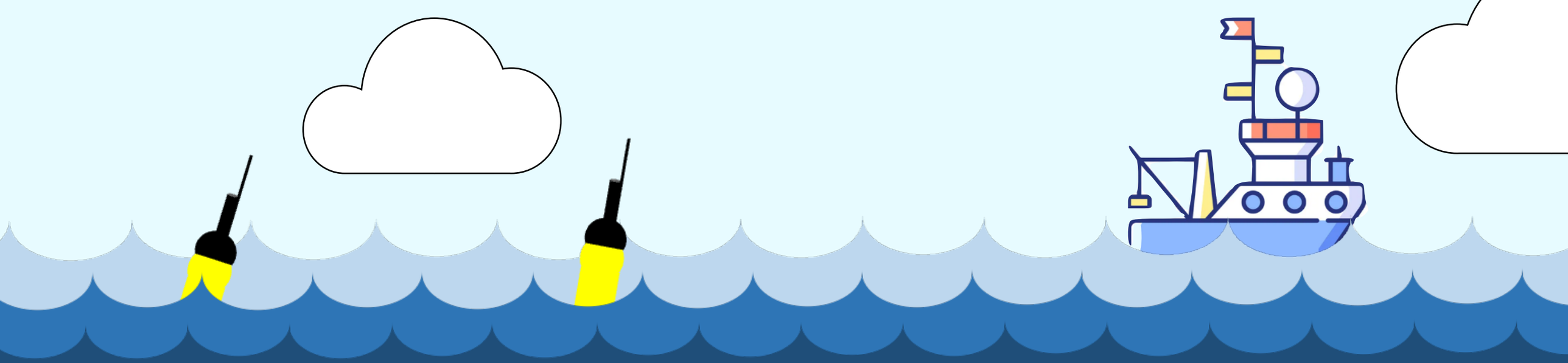


Sparse data & explainable ML to evaluate Earth System Models

Abby P. Shaum, Galen A. McKinley

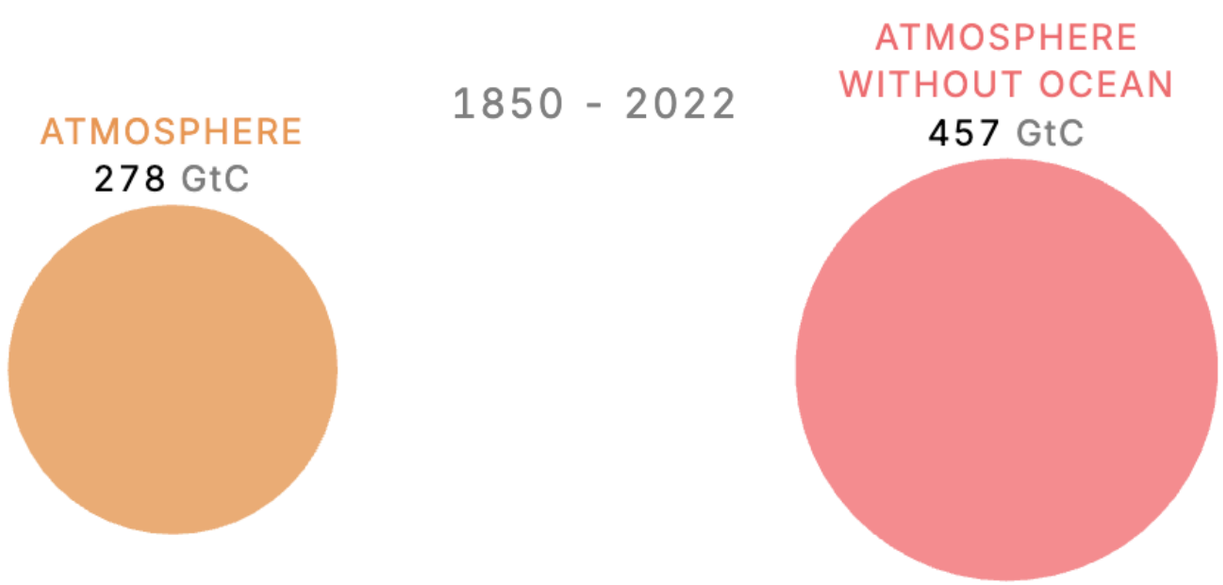
Columbia University and Lamont-Doherty Earth Observatory

CESM Land & Biogeochemistry Working Group Meeting | NSF NCAR, February 25, 2025

We can measure the strength of the ocean carbon sink using pCO_2 measurements

- Oceans take up carbon from the atmosphere
- So far, oceans have absorbed ~40% of anthropogenic carbon emissions

pCO_2 : "partial pressure" of CO_2 in the surface ocean

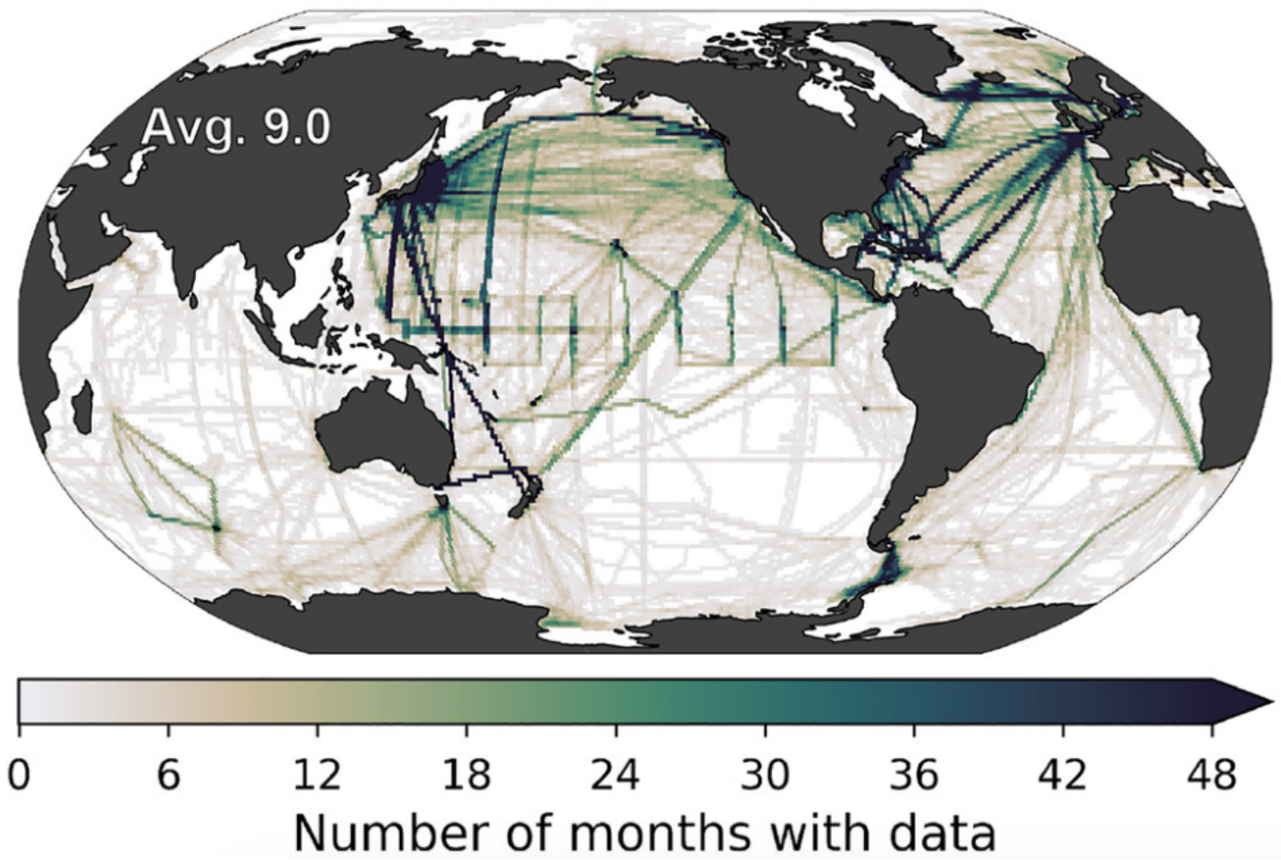


See full article at https://stories.leap.carbonplan.org/ocean-sink

However, pCO_2 observations are sparse

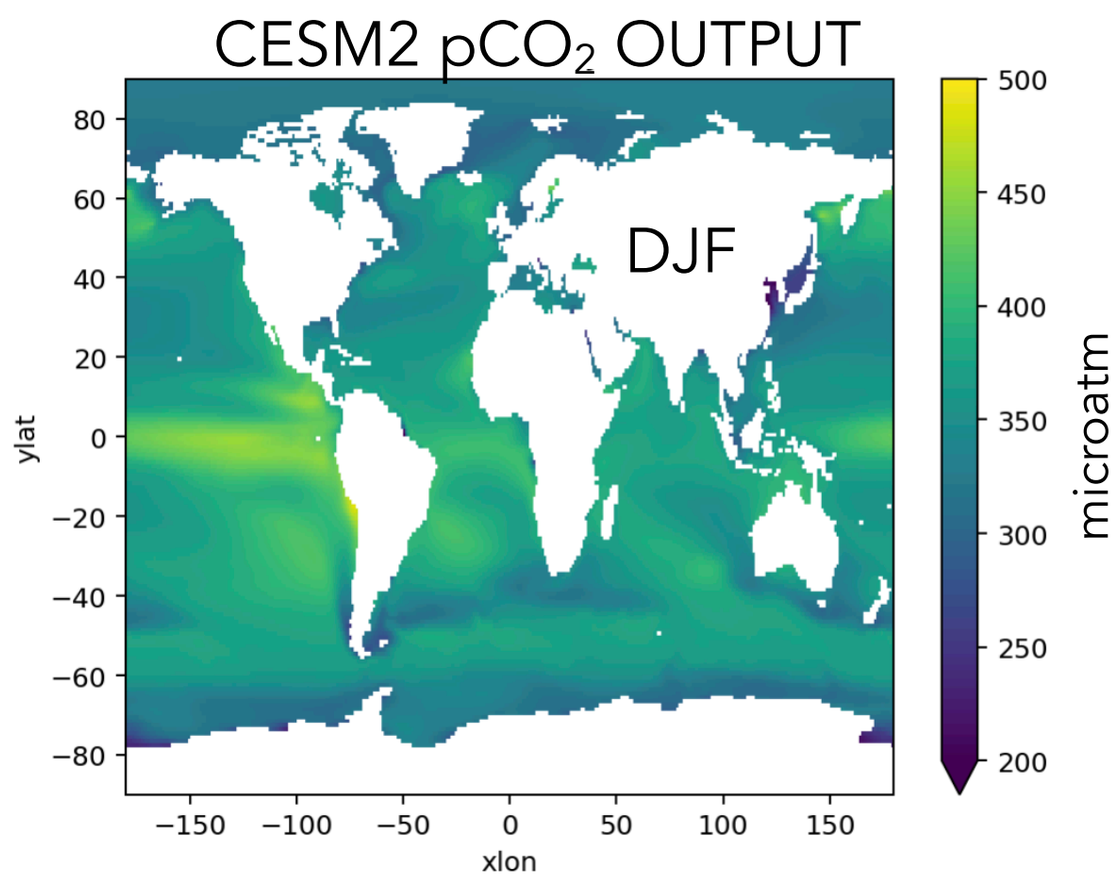
- Measurements cover < 2% of the ocean in space/time
- Sampling biased for Northern Hemisphere, and each hemisphere's respective summer months

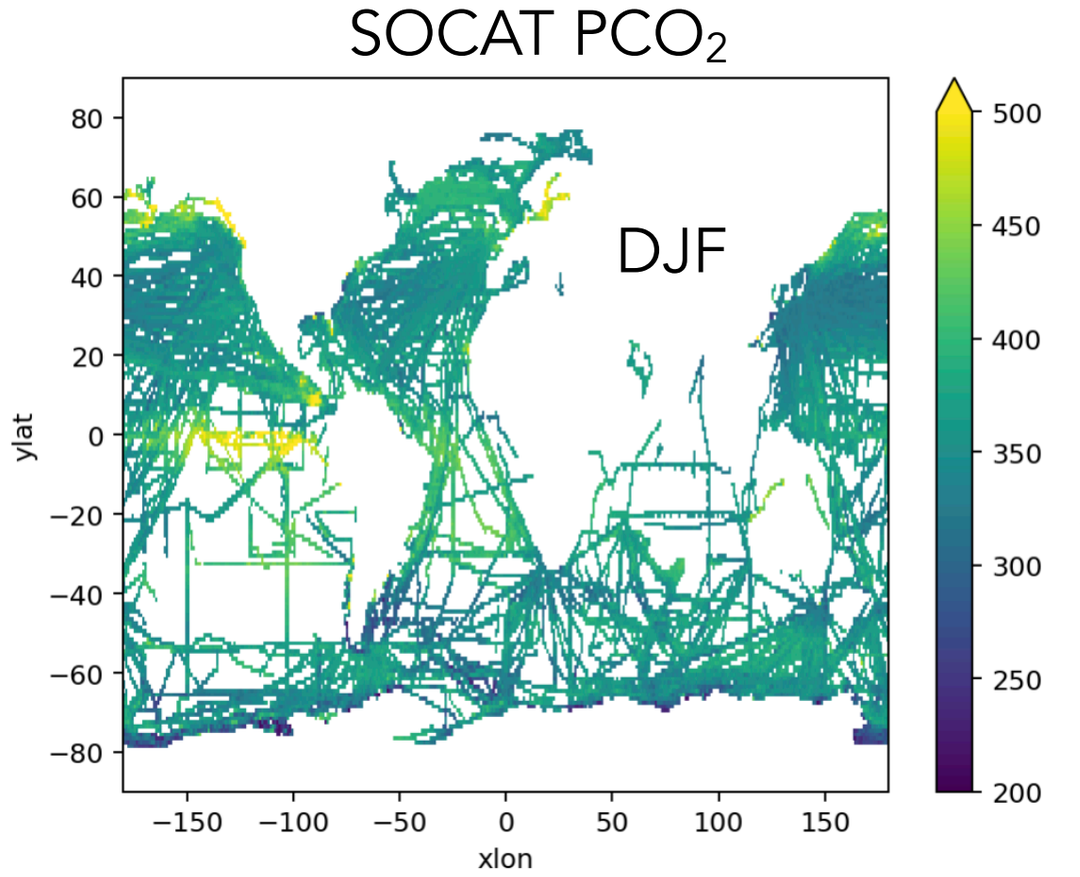
SOCAT ship tracks from 1982-2016



CESM2 provides full coverage pCO_2

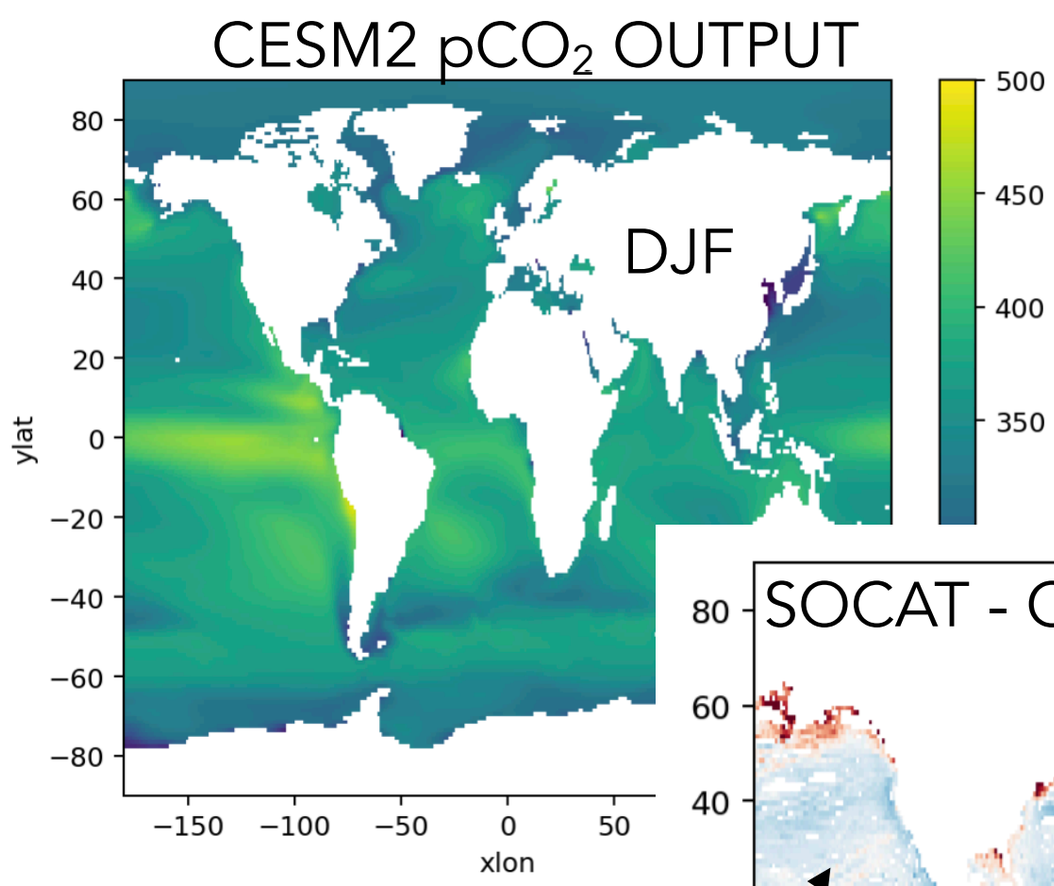
But model output does not match real-life observations!

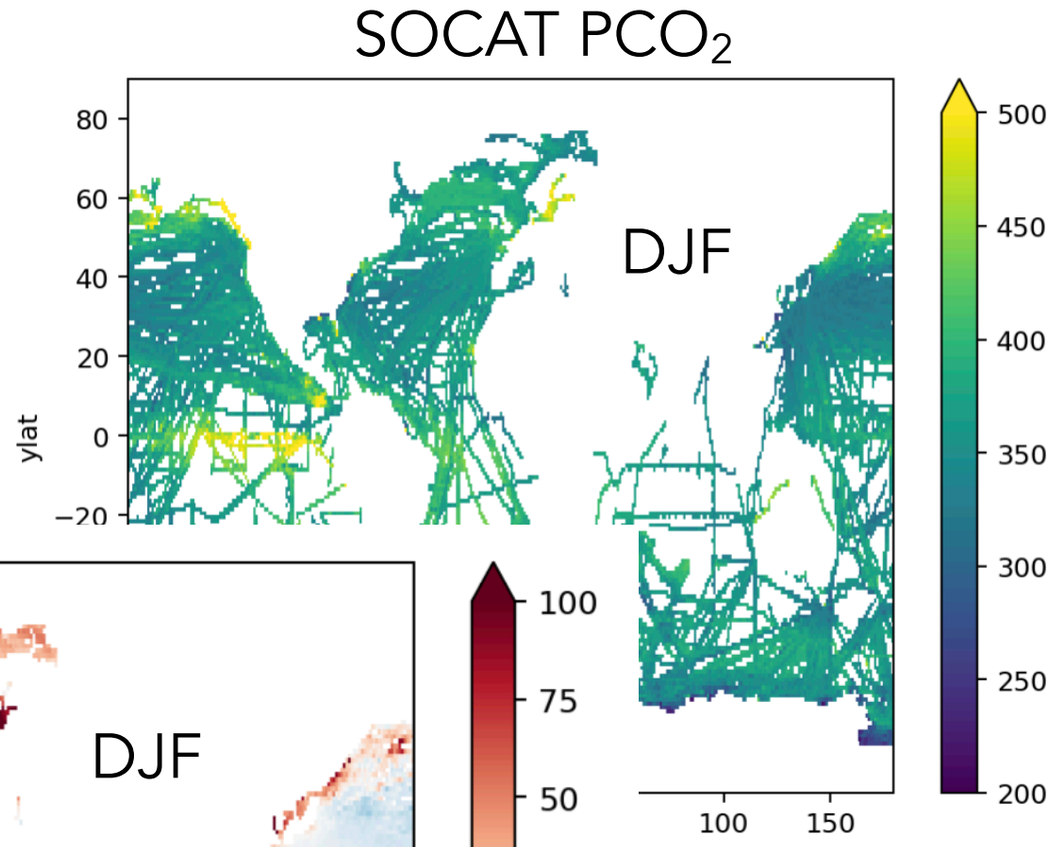


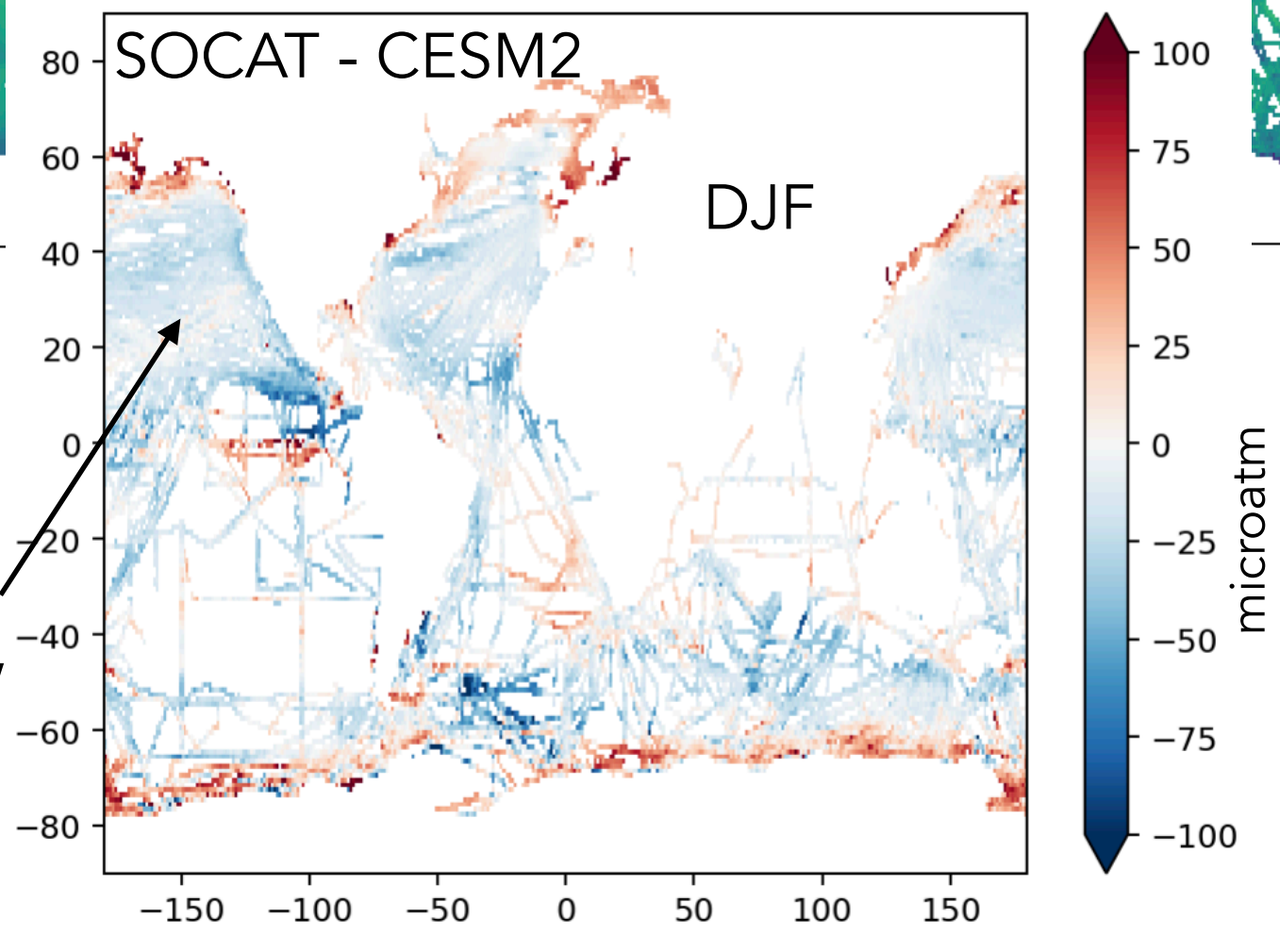


CESM2 provides full coverage pCO_2

But model output does not match real-life observations!







We call this pCO_2 "misfit"

Can we understand the underlying drivers behind pCO_2 misfit?

1. What impacts pCO_2 misfit more: CESM2 climatology or inter-annual variability?

2. If CESM2 climatology IS the predominant source of pCO_2 misfit, how does each biogeochemical driver impact pCO_2 misfit?

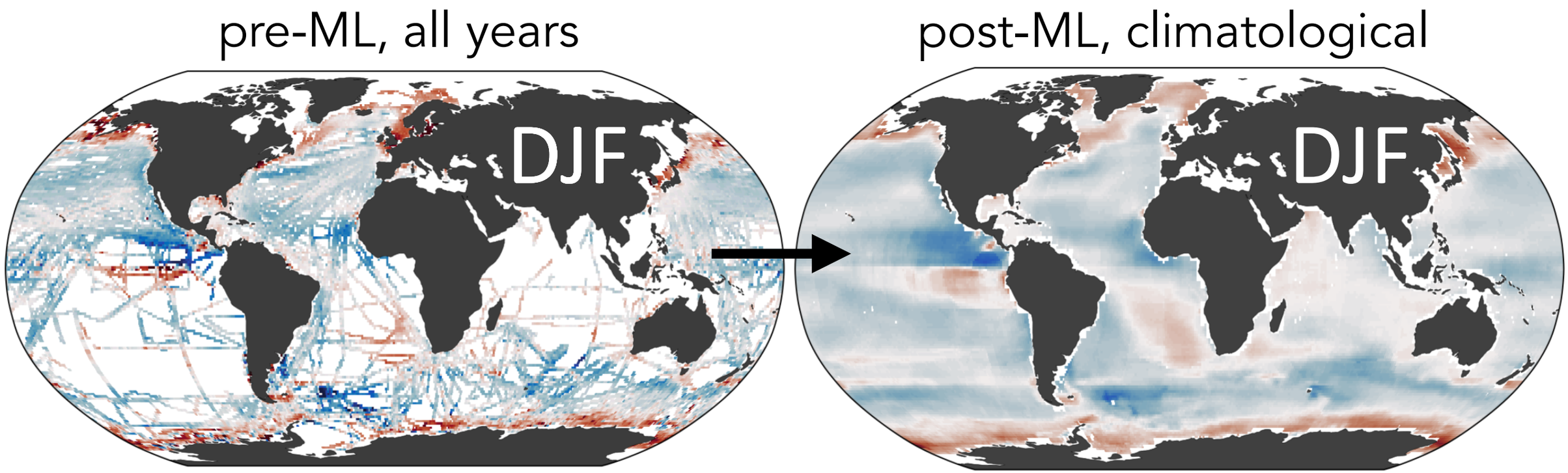
Using ML to reconstruct pCO_2 misfit with LDEO-HPD

ESMs, GOBMs, and ML Data Products estimate the ocean carbon sink

LDEO-HPD:

- Combines Earth System Models **and** Machine Learning
- **Uses sparse data to understand model errors over broad space-time scales**

Using ML to reconstruct pCO_2 misfit

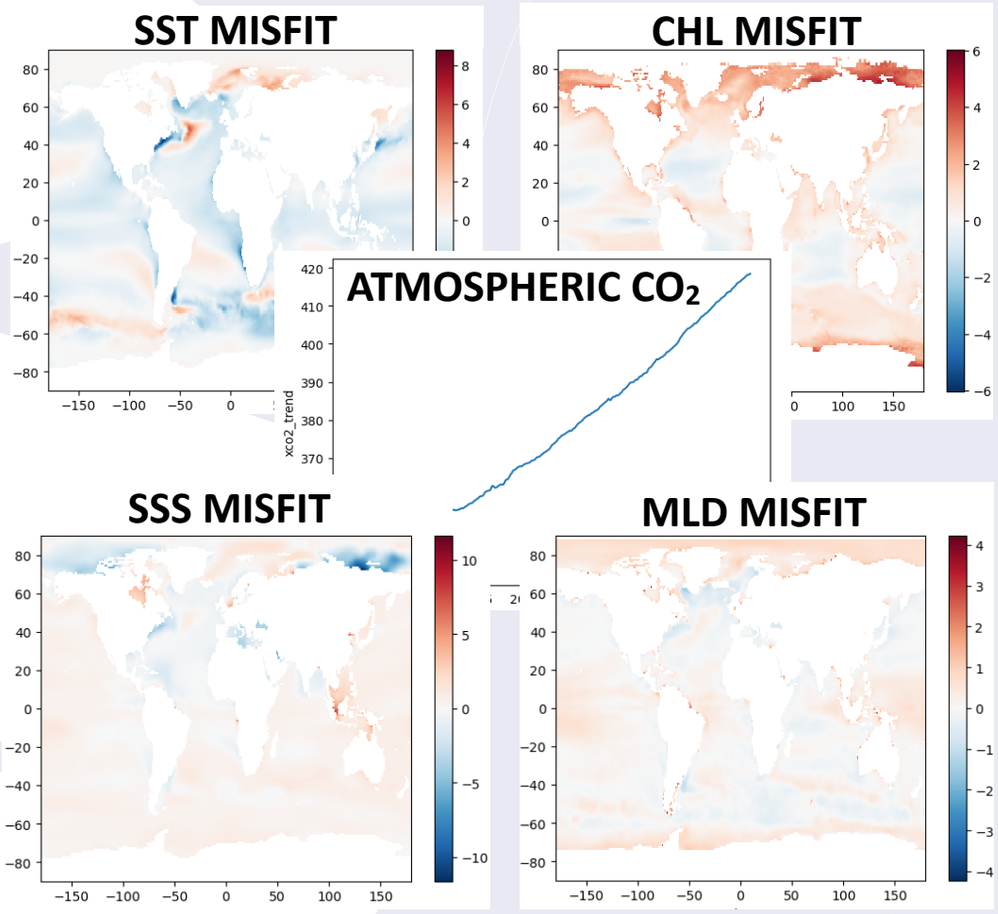


ML Target: CESM2 pCO_2 Misfit

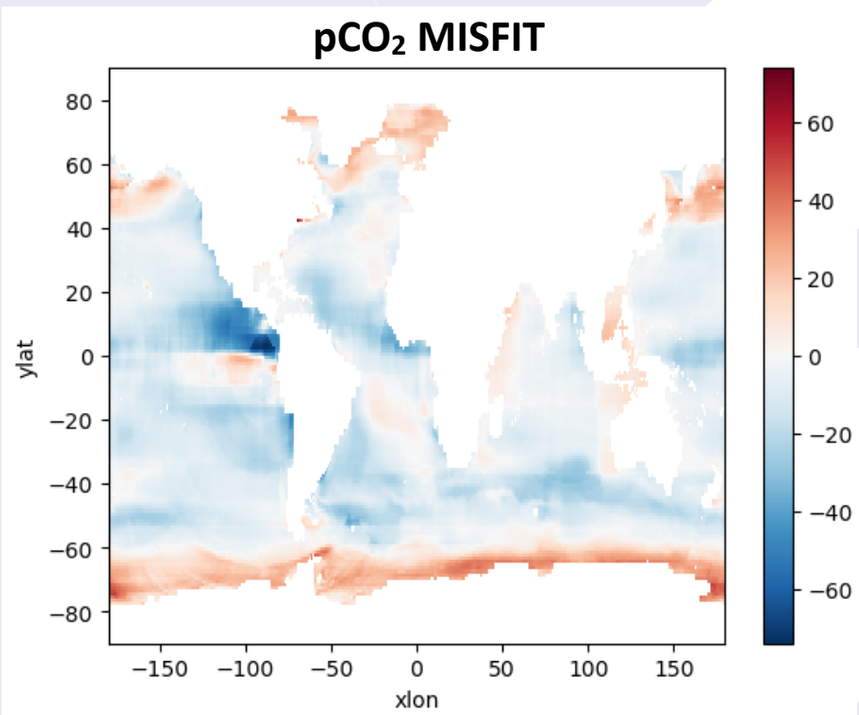
Method:

We train on misfits of biogeochemical drivers!

Biogeochemical drivers



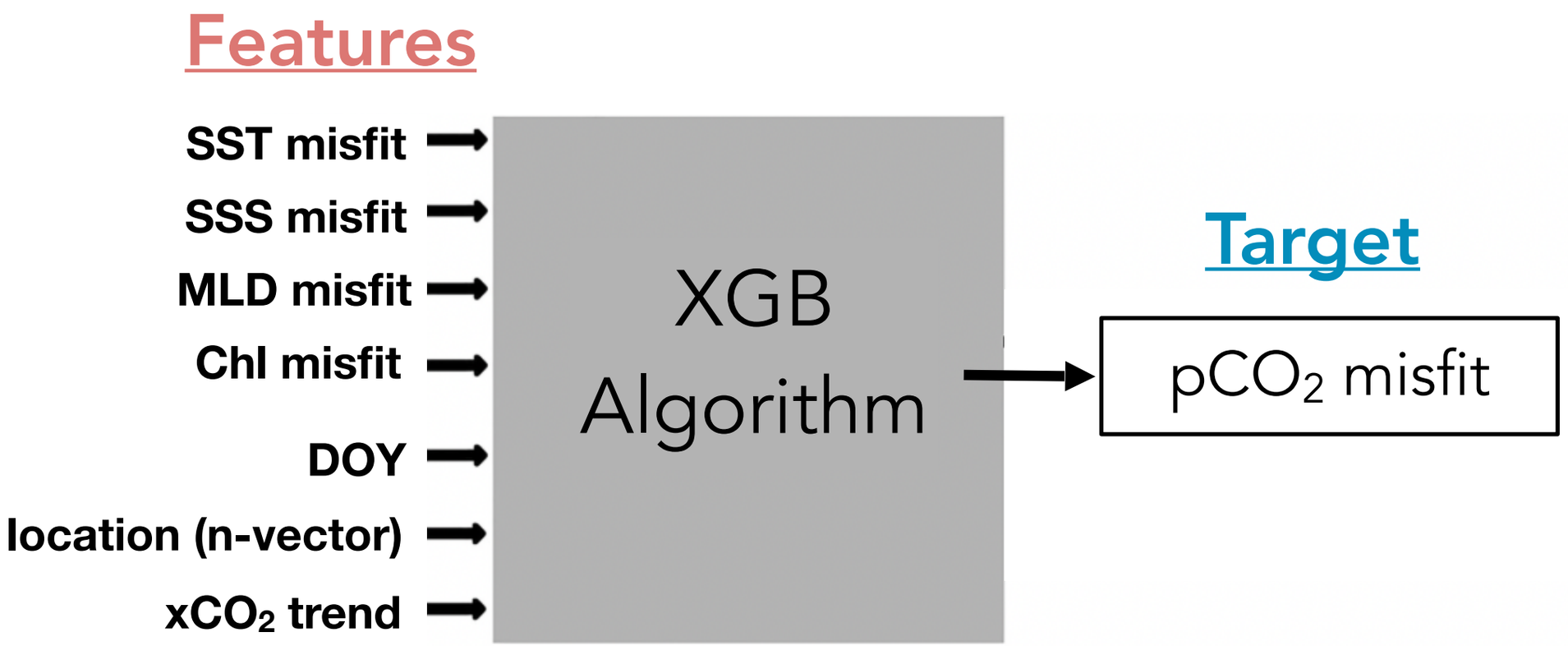
Target



Following Gloege et al., 2022 , **JAMES** (doi: 10.1029/2021MS002620)

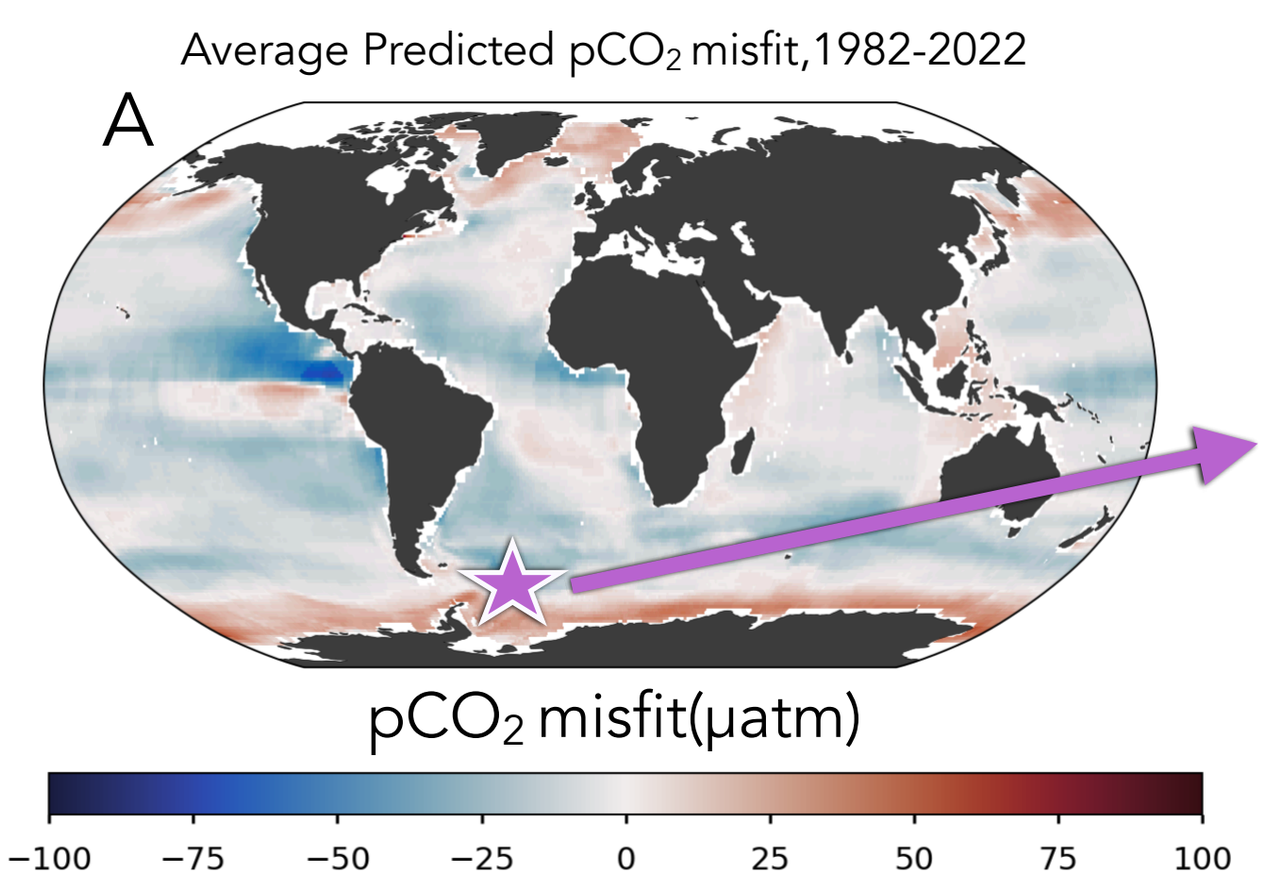
Method:

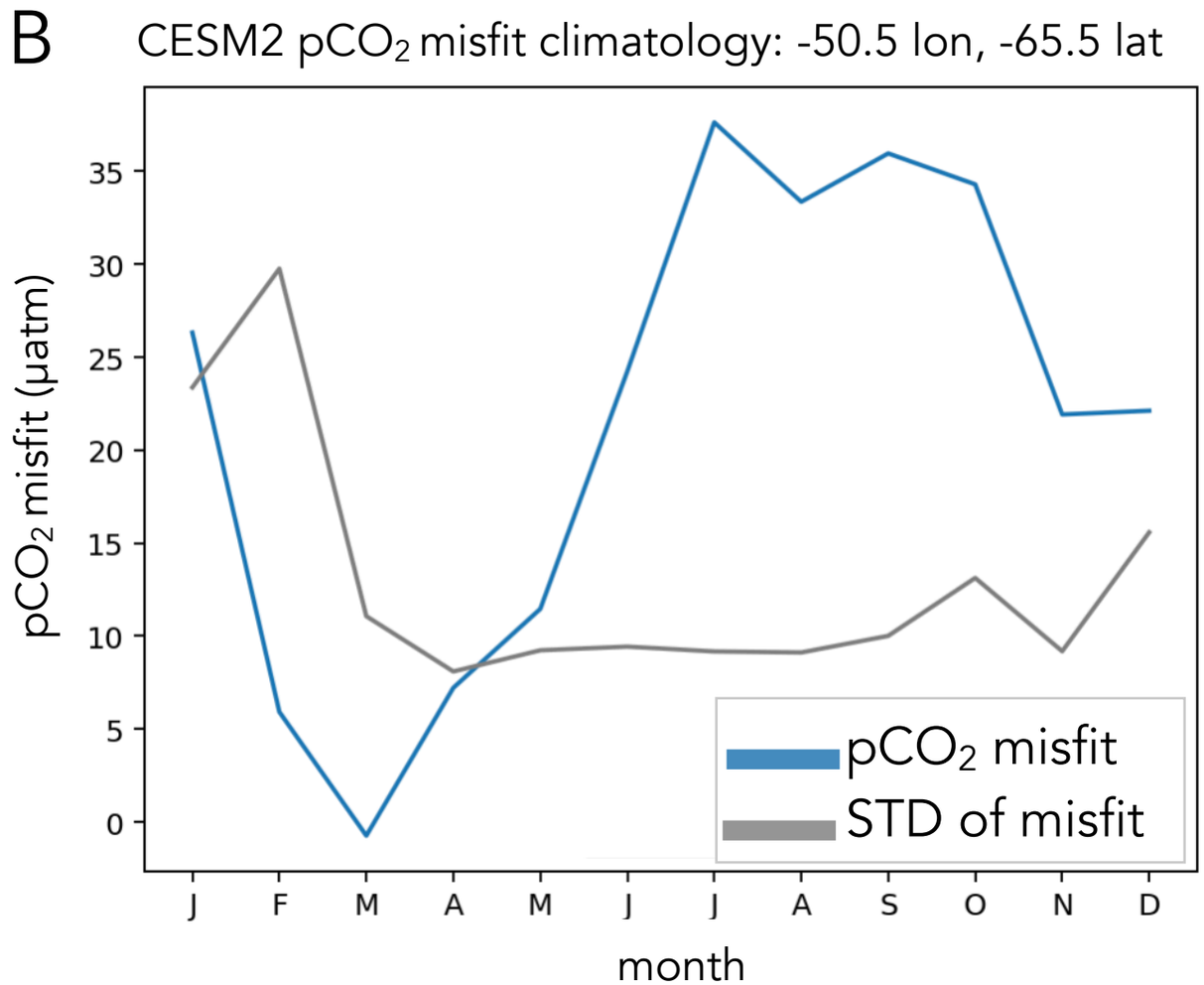
1. Calculate misfits (observation - CESM2) for SST, SSS, MLD, Chl, and pCO_2
2. Train XGB algorithm to learn relationships
3. Use trained ML model to predict pCO_2 misfit everywhere in space/time
4. Use SHAP (XAI) to determine spatial and temporal feature importance



Question 1 Results:

What impacts pCO_2 misfit more: CESM2 climatology or inter-annual variability?



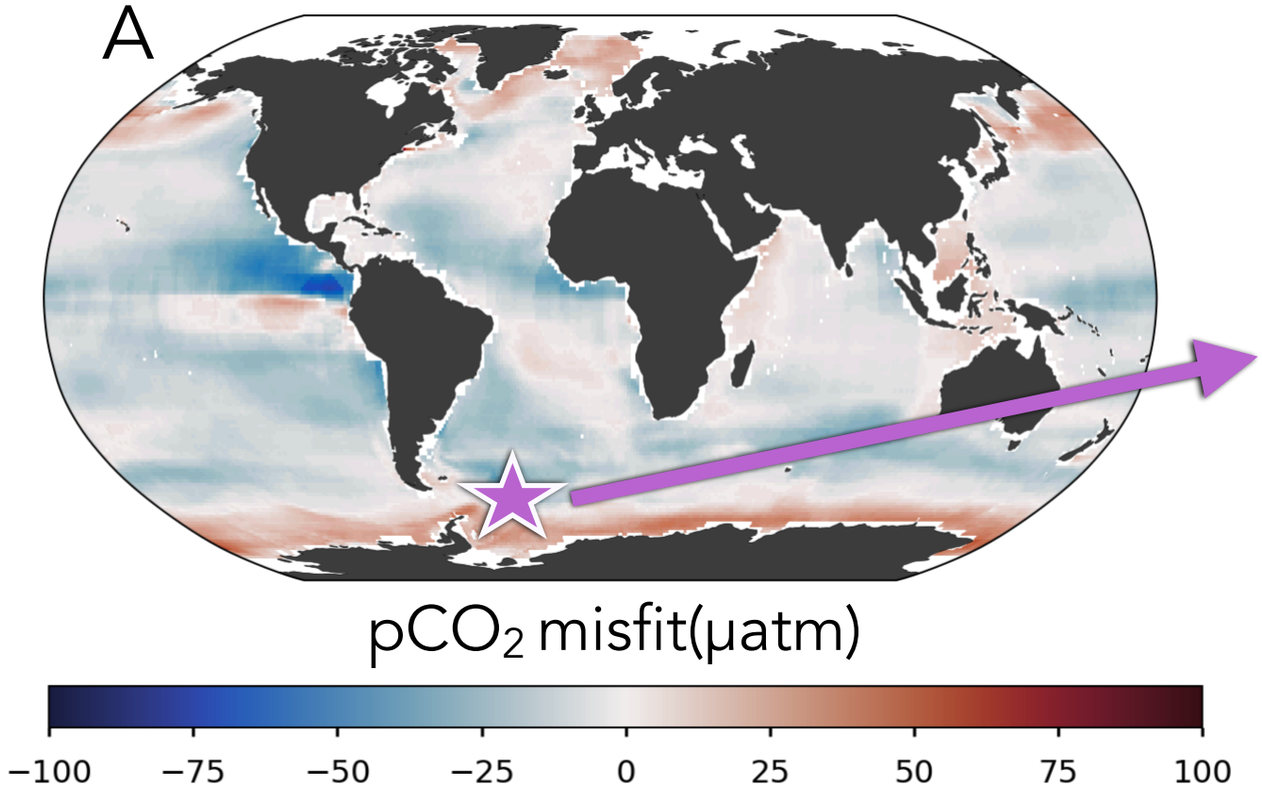


<u>Take away:</u> CESM2 climatology contributes more to pCO_2 misfit than inter-annual variability does

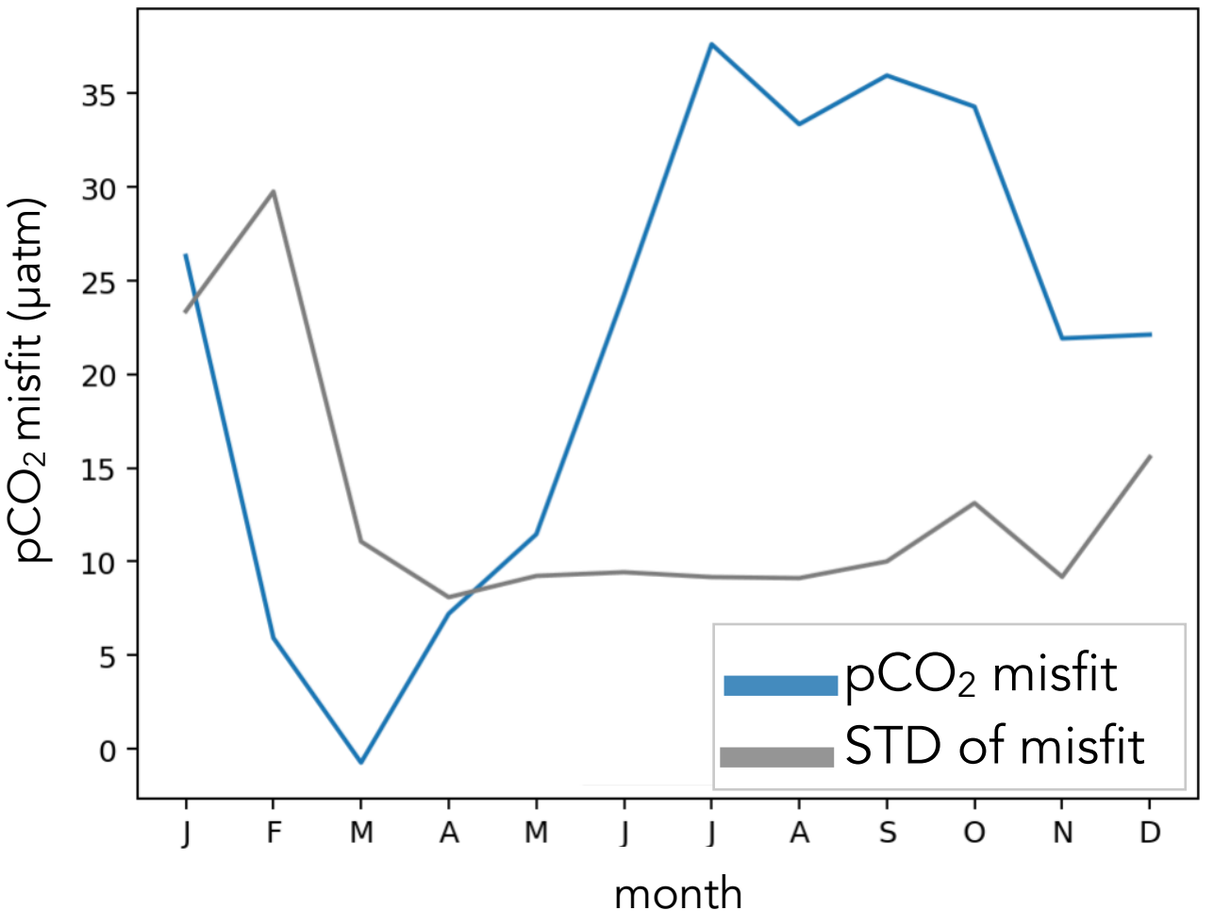
Question 1 Results:

What impacts pCO_2 misfit more: CESM2 climatology or inter-annual variability?

Average Predicted pCO_2 misfit,1982-2022



B CESM2 pCO_2 misfit climatology: -50.5 lon, -65.5 lat



How does each biogeochemical driver impact pCO_2 misfit?

DETAILS ON SHAP

```
explainer = shap.TreeExplainer(trained_ML_model,feature_data)
shap_values = explainer.shap_values(feature_data)
```

SHAP is easy to use!

ML Feature Variables: our **players**

(misfits of sst, sss, chl, mld, as well as xCO_2 and loc/doy)

ML Task: our **game**

(predicting pCO_2 misfit)

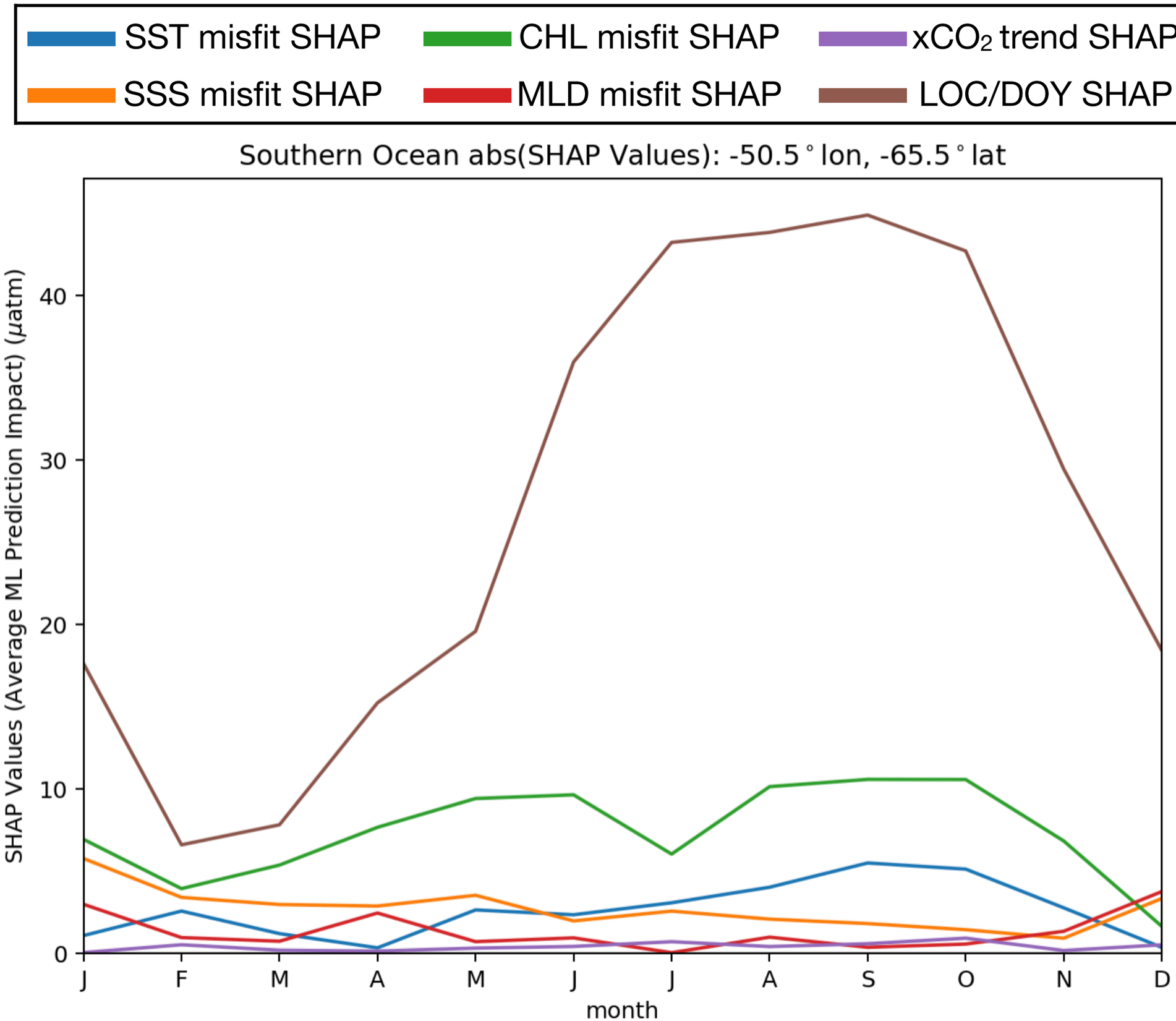
Feature Importance: the **payout**

(larger feature importance -> larger contribution to the ML task)

Great interpretable AI source here: https://christophm.github.io/interpretable-ml-book/shapley.html

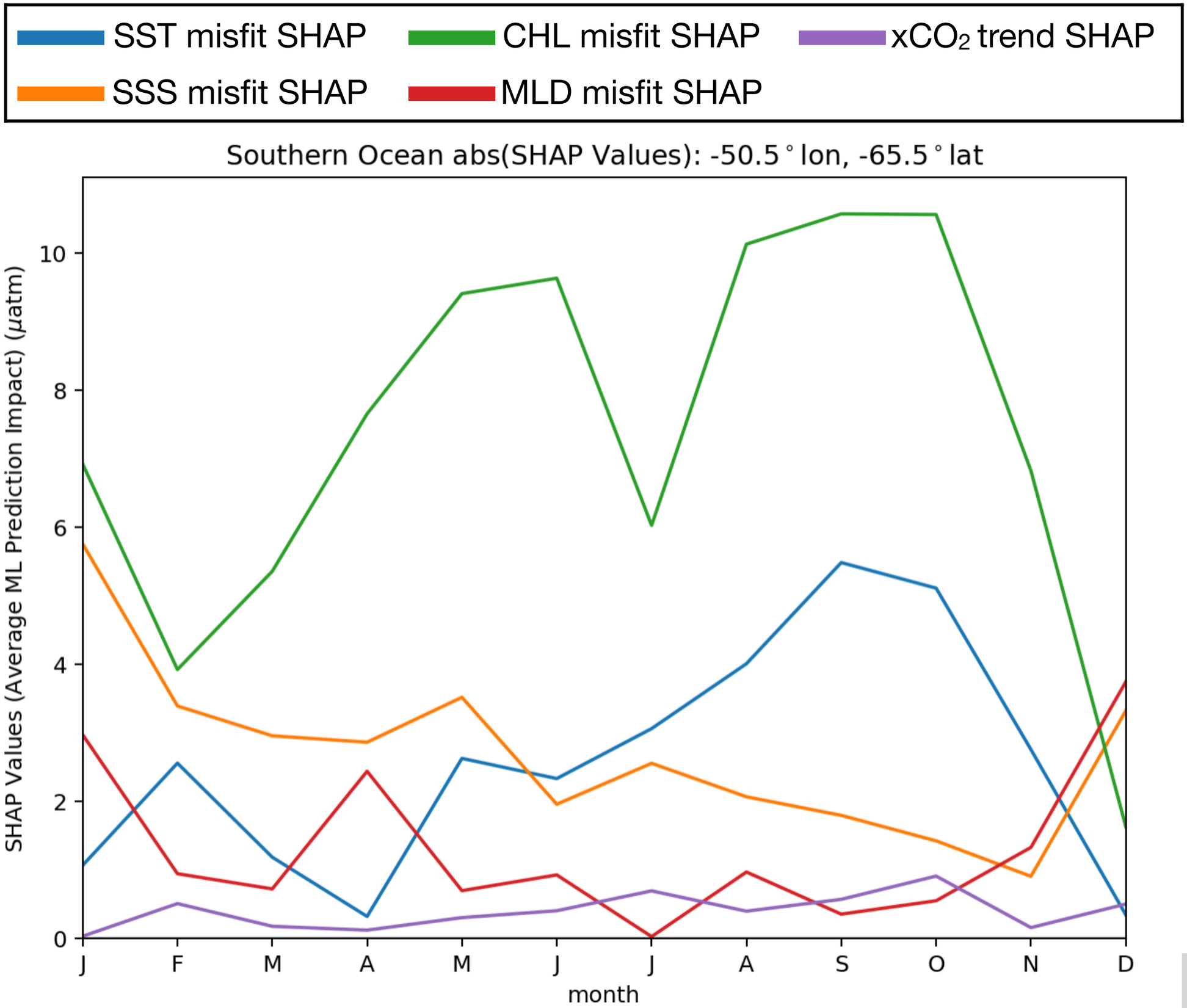
Preliminary Question 2 Results:

How does each biogeochemical driver impact pCO_2 misfit?



Preliminary Question 2 Results:

How does each biogeochemical driver impact pCO_2 misfit?



Method expansions

1. Perform global analysis for this CESM2 member

2. Develop an analysis pipeline

3. Extend method to more CESM2 members (beyond r10i1p1f1)

4. Expand methodology of reconstructing misfit (model-observation difference) to land or atmosphere problems, where useful

5. Continue exploring SHAP and more interpretable AI tools

Questions? Input?

Contact me at at
cs4430@columbia.edu

McKinley Ocean Carbon Group

McKinley OCG Page

Ocean Carbon Sink Page

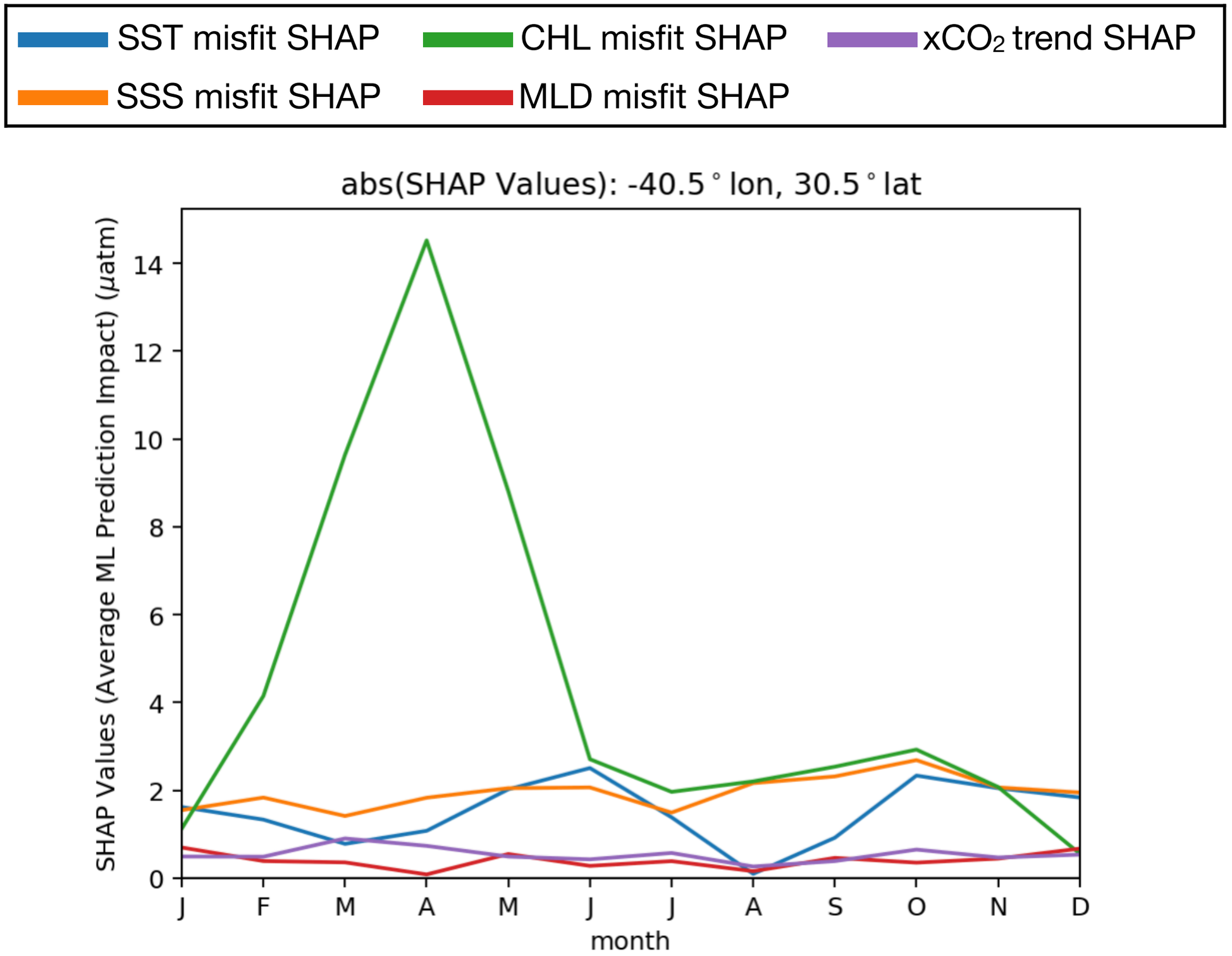
References:
1. Gloege et al., 2022 , **JAMES** (doi: 10.1029/2021MS002620);
2. Bennington et al., 2022, **GRL** (doi: 10.1029/2022GL098632);
3. Gloege et al., 2021, **GBC** (doi:10.1029/2020GB006788);
4. McKinley et al., 2023, **ERL** (doi: 10.1088/1748-9326/acc195)

Resources:
1. SHAP and more explainable AI/ML: https://christophm.github.io/interpretable-ml-book/shapley.html

Preliminary Question 2 Results:

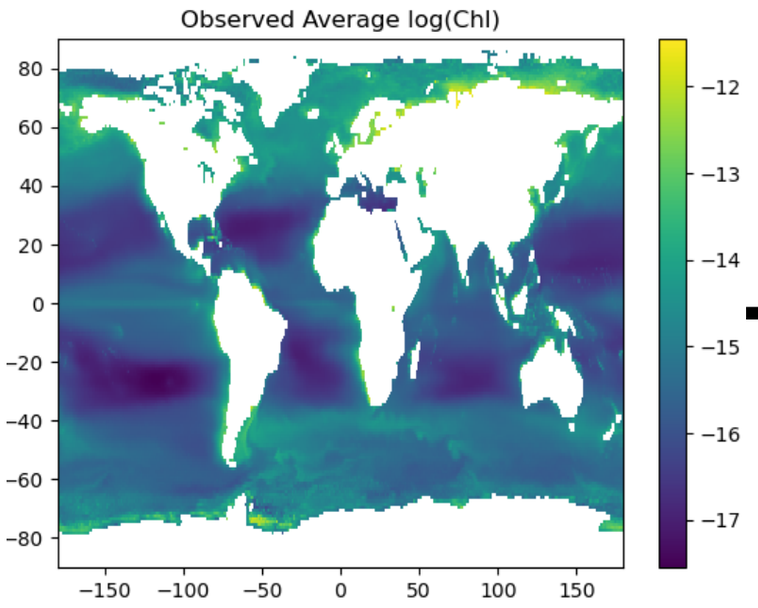
How does each biogeochemical driver impact pCO_2 misfit?

Mid-Atlantic

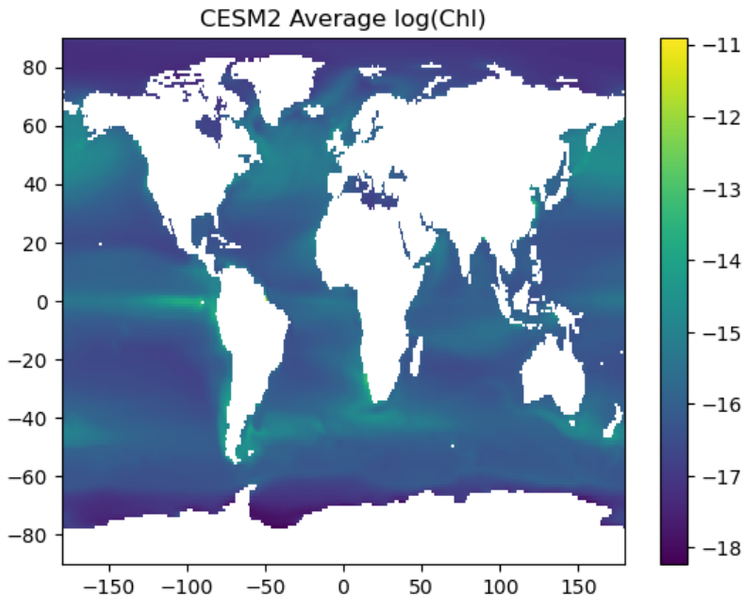


log(CHL)

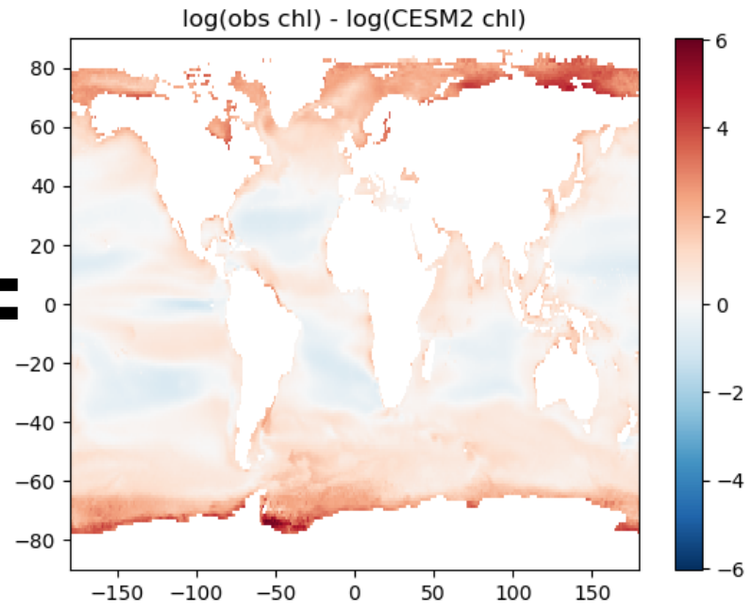
OBSERVED



CESM2 OUTPUT

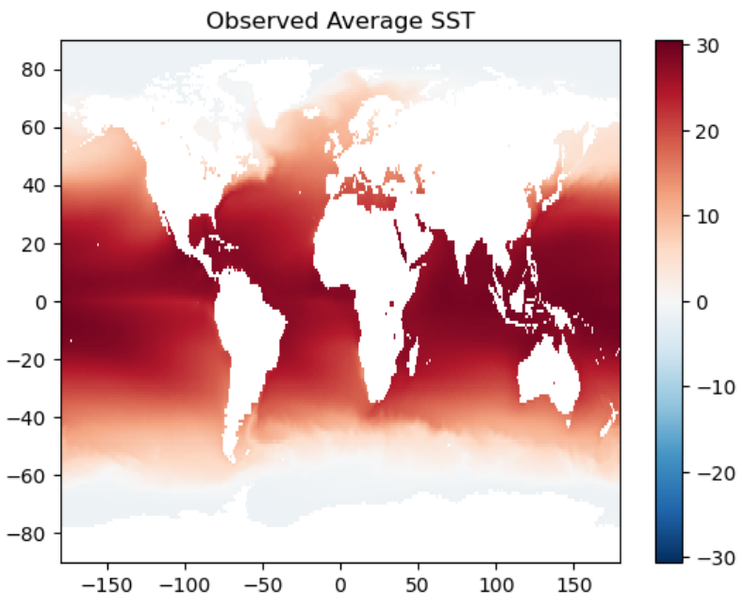


DIFFERENCE

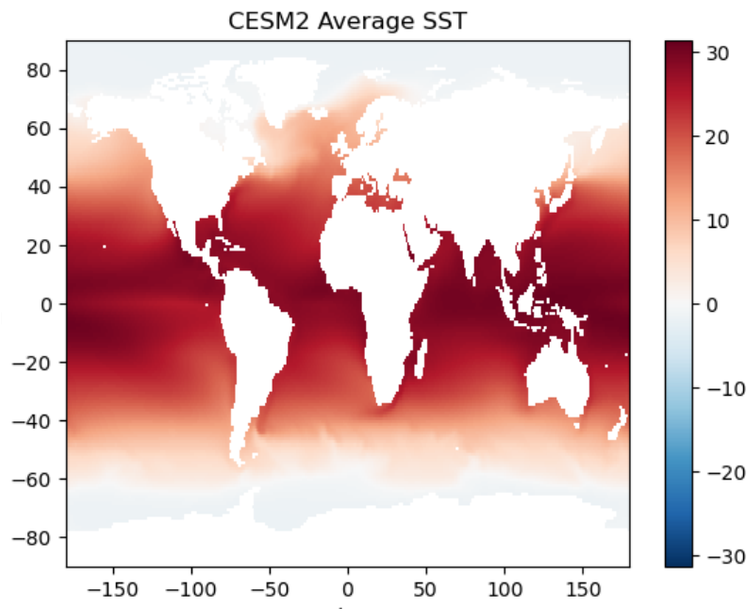


SST (degC)

OBSERVED



CESM2 OUTPUT



DIFFERENCE