

COMPUTATIONAL
CLIMATE & OCEAN GROUP
at UC Davis

UCDAVIS
COMPUTER SCIENCE

OceanCBM: A Concept Bottleneck Model for Mechanistic Interpretability in Ocean Forecasting

Sanah Suri*†, Kieran Ringel*, Maike Sonnewald*
***University of California, Davis**
† sansuri@ucdavis.edu

NeurIPS 2026

Motivation

Extreme ocean phenomena are challenging not only to predict but also ***diagnose*** as accurate forecasts along do not reveal ***underlying physical drivers***.

While recent machine learning approaches achieve strong predictive skill, they remain largely ***opaque*** and provide limited guarantees of ***fidelity to ground-truth physics***.

Concept bottleneck models

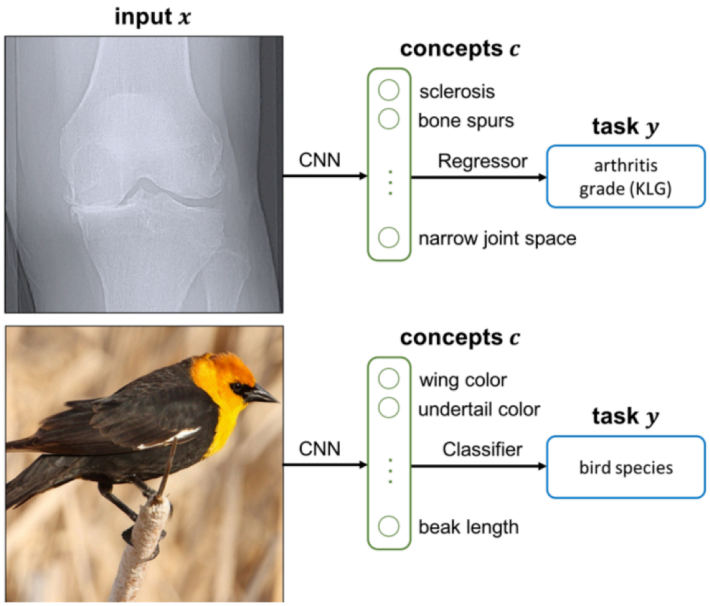


Figure 1 [Koh et al. 2020]

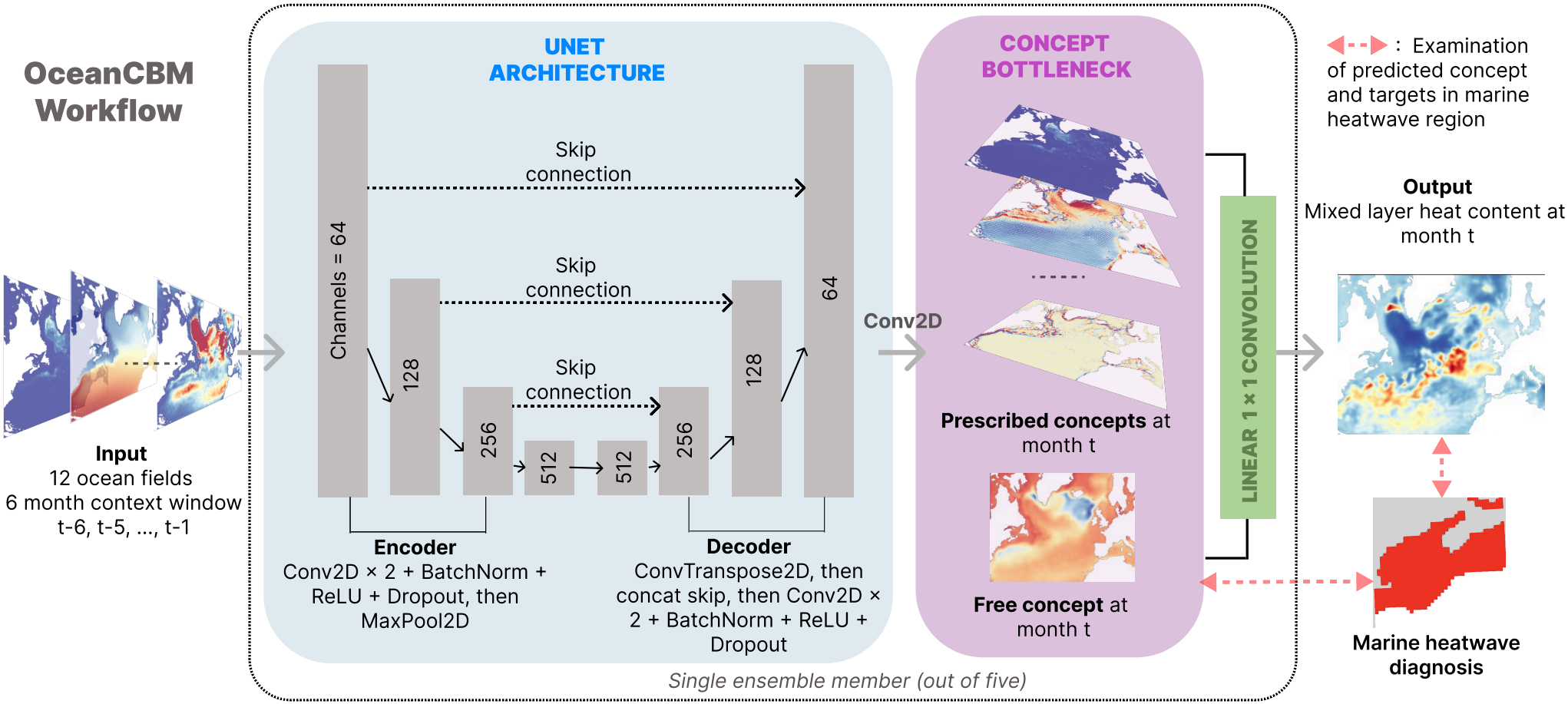
- *Concept bottleneck models (CBMs) route predictions through an intermediate layer of interpretable concepts.*

- *Generally, a combined loss function is utilized for training*

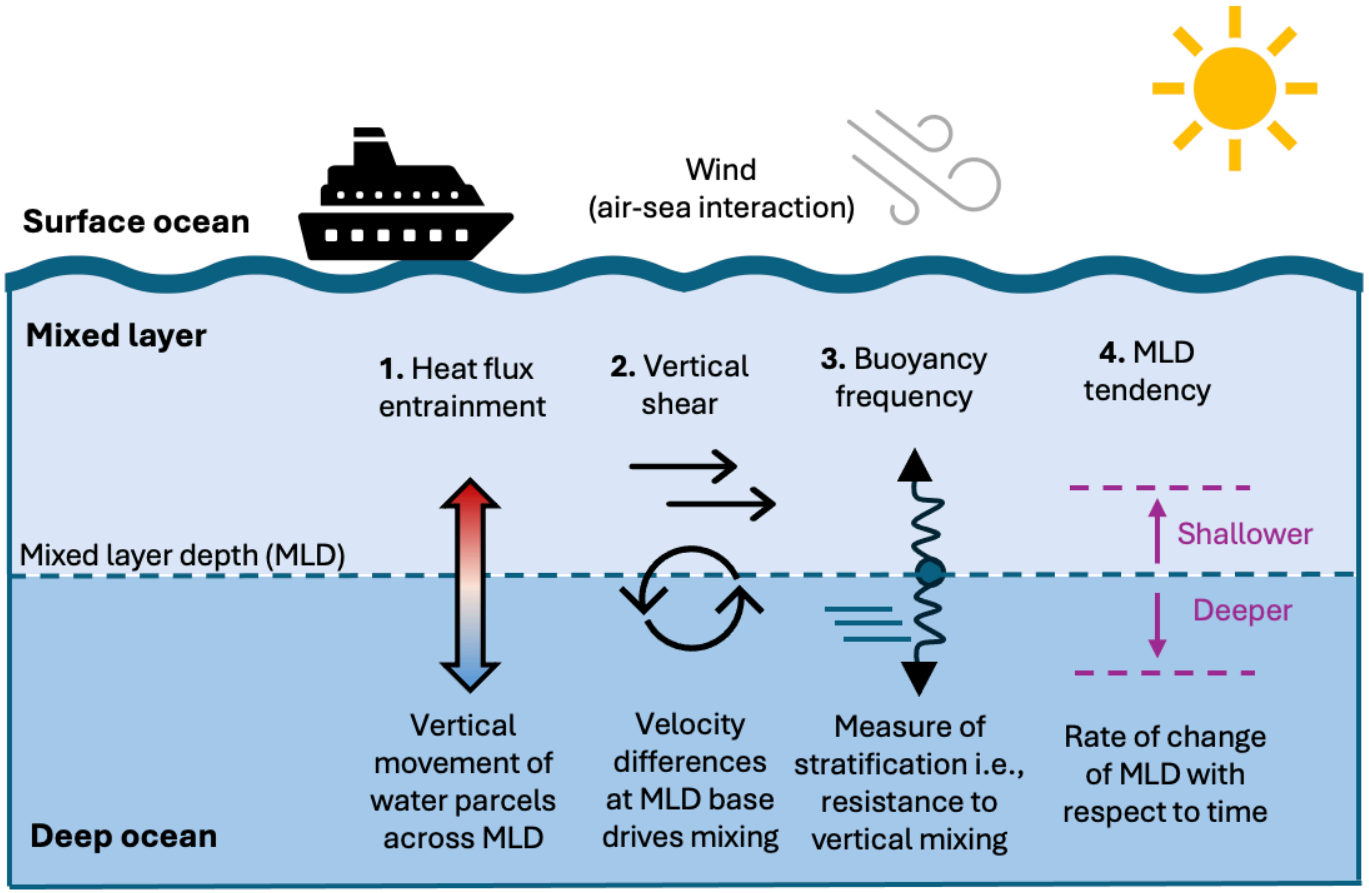
$$\mathcal{L}_{\text{combined}} = \lambda \mathcal{L}_{\text{concept}} + (1 - \lambda)\mathcal{L}_{\text{pred}}$$

- *CBMs have successfully been applied to several tasks, but not to ocean physics*

Presenting OceanCBM



The Concepts

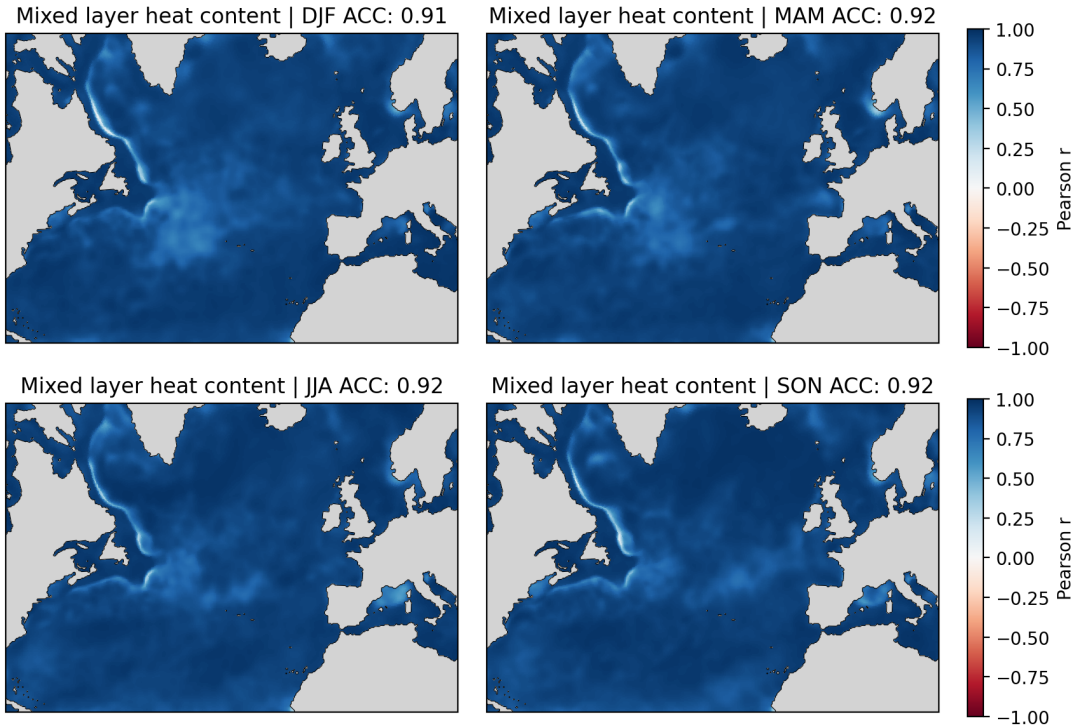


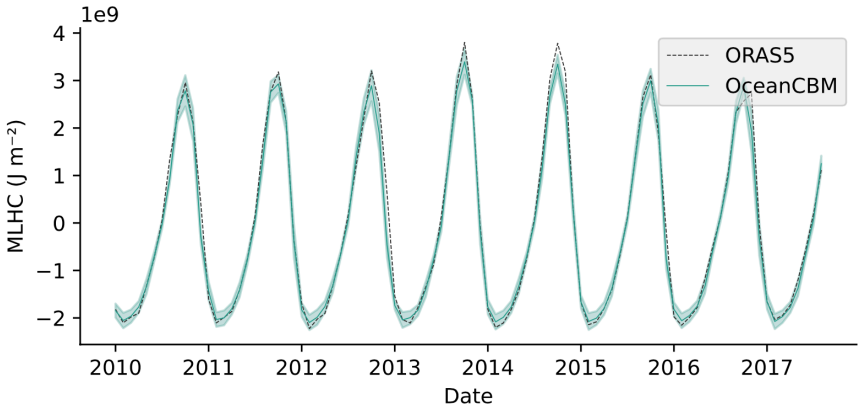
- *Carefully selected by physical oceanography expert*
- *The mixed layer is where the ocean interacts with the atmosphere, making it dynamically rich and important*
- *Calculated using input data from ECMWF's ocean reanalysis 5 (ORAS5)*
- *ORAS5 is an underutilized dataset in the field*

Evaluating OceanCBM

1. Baselines
    A. **Prediction-only model:** *Same architecture without the concept bottleneck layer but with matching latent dimensions*
    B. **Prescription-only model:** *Retains the concept bottleneck but removes the free concept*
2. Metrics
    A. **Spatial:** *Anomaly correlation coefficient (ACC)*
    B. **Temporal:** *Time series comparison*
    C. **Learning mechanisms:** *Weights of concept -> prediction linear layer*

Mixed layer heat content (MLHC) prediction

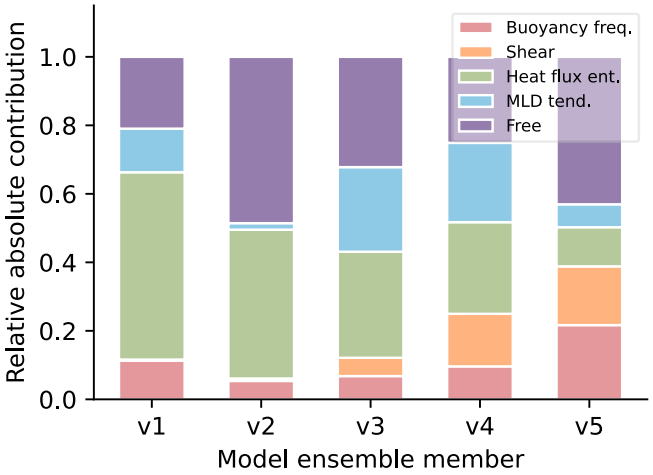




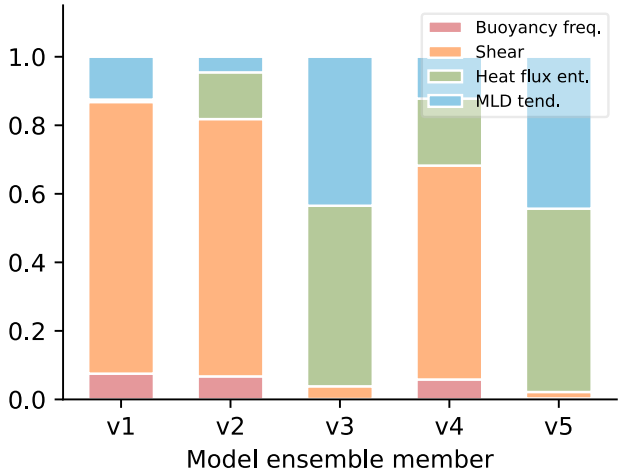
Model	DJF	MAM	JJA	SON
OceanCBM	0.9116	0.9174	0.9191	0.9207
Prediction-only	0.9131	0.9186	0.9180	0.9216
Prescription-only	0.9129	0.9187	0.9188	0.9212

Learning mechanisms

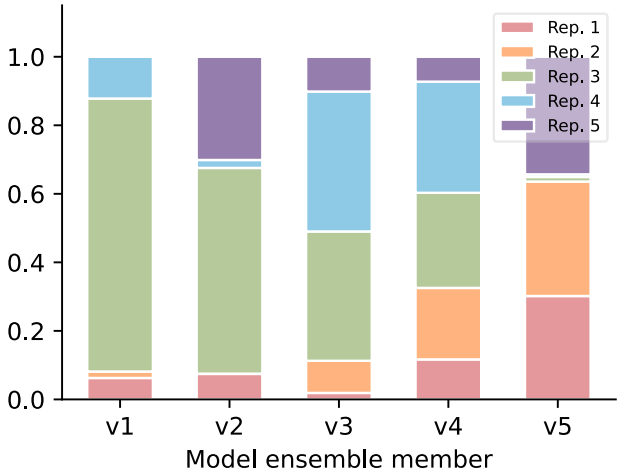
Respective ensemble members have the same random weight initialization (v1 matches v1 matches v1)



OceanCBM model



Prescription-only model



Prediction-only model

Free concept dual role

(1) Residual physical signal

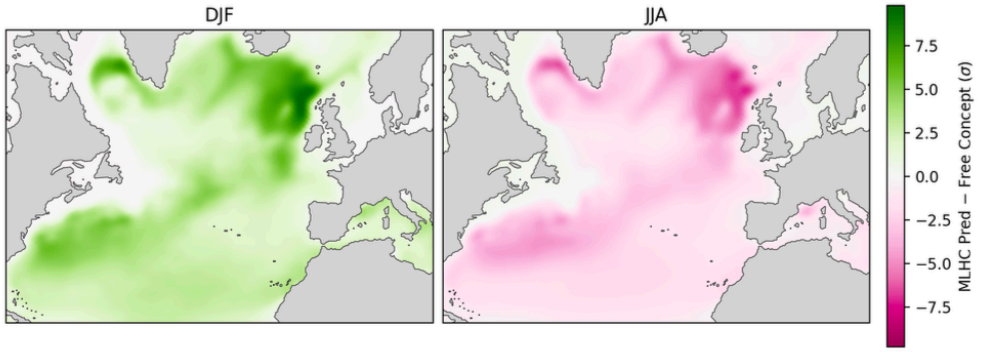


Figure 6: **Seasonal differences between MLHC and free concept.** The difference between the predicted MLHC and free concept, taken in normalized space, over the out-of-sample time period. Results are shown for DJF and JJA, representing winter and summer conditions respectively.

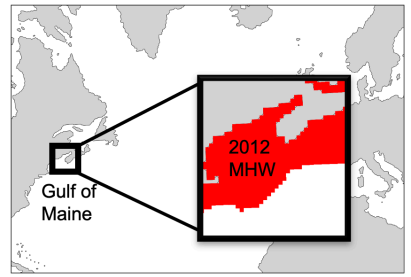
(2) Concept regularizer

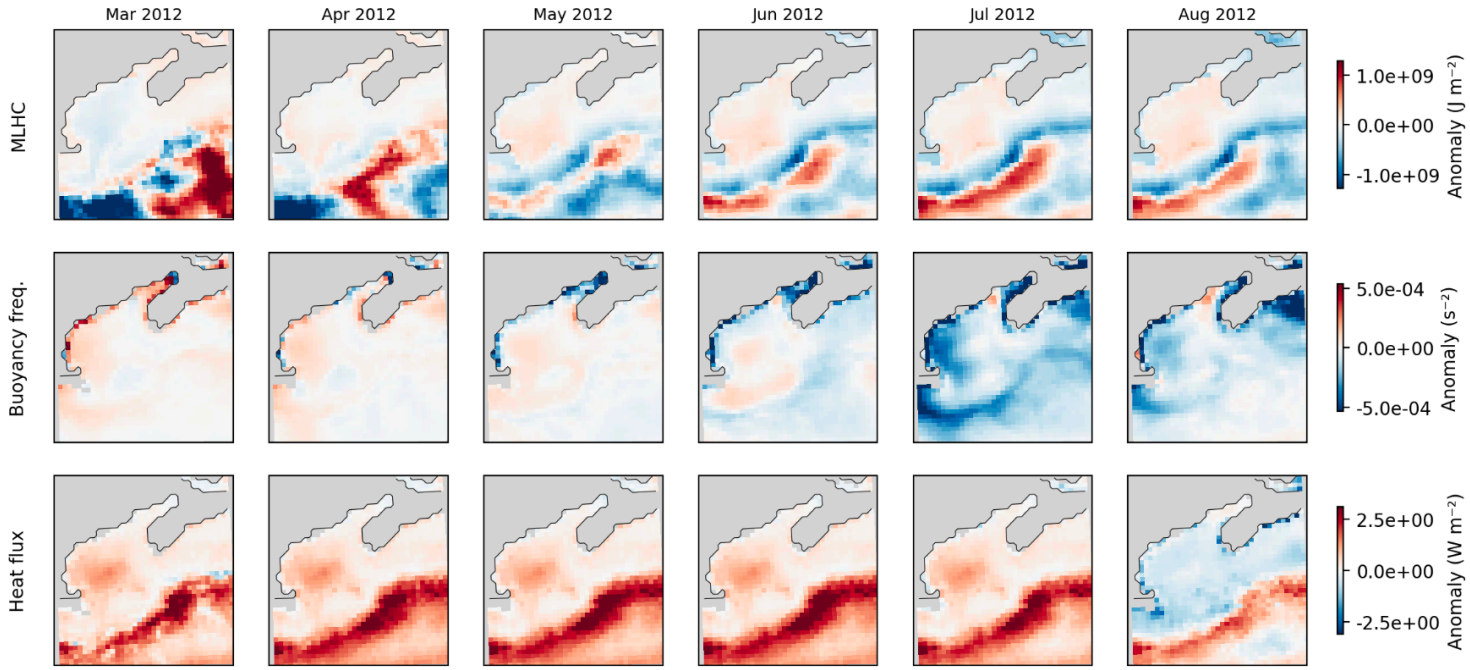
Vertical Shear ACC

	OceanCBM	Prescription-only
DJF	**0.60**	0.53
MAM	**0.61**	0.57
JJA	**0.67**	0.57
SON	**0.58**	0.57

Table 1: **Comparing concept ACC:** Seasonal mean ACC of vertical shear for OceanCBM and prescription-only model, with ACC being consistently lower in the prescription-only configuration.

Case study: 2012 Gulf of Maine marine heatwave





Our analysis agrees with previous atmospheric and ecological findings

Analysis

- *Positive anomaly in heat flux entrainment implying persistent atmospheric forcing attempting to penetrate downward into deeper ocean.*
- *Positive anomaly in buoyancy frequency implying increased stratification making this mixing hard, allowing warmer waters to accumulate in the mixed layer*
- *Closer to onset, behavior switches meaning ocean advection was working against atmospheric forcing but it was not "enough"*