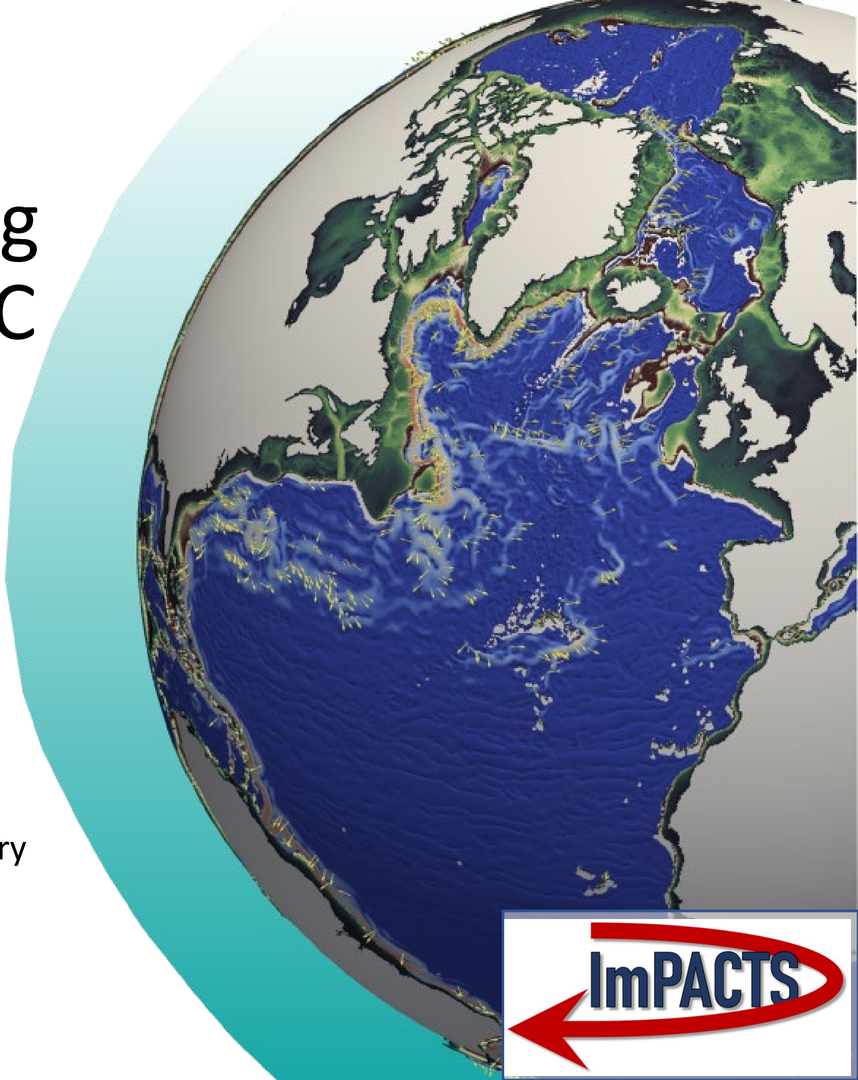
Improving the Understanding and Representation of AMOC using Advanced ML/AI Techniques

Office of Biological and Environment Research (BER)

Office of Advanced Scientific Computing Research (ASCR)

Wei Xu[1], Xihaier Luo[1], Derek Freeman DeSantis[2], Shinjae Yoo[1]

[1]Brookhaven National Laboratory, [2]Los Alamos National Laboratory

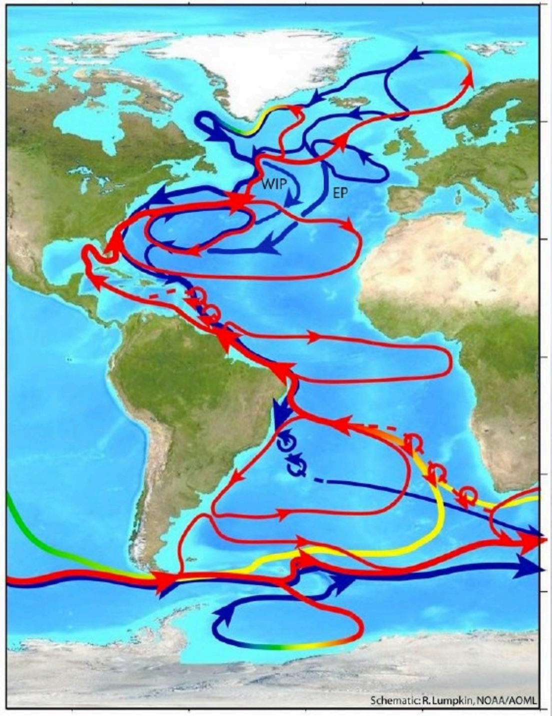


Topic Area: Data Intensive Computing   DOE Data Days 2023

ImPACTS: Improving Projections of AMOC and its Collapse Through advanced Simulations

The Atlantic Meridional Overturning Circulation (AMOC):

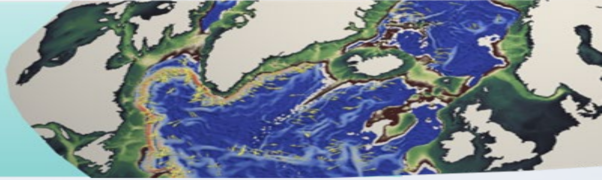
- An essential system of ocean currents in the Atlantic. Significantly contribute to global climate regulation. The variability of AMOC tied to numerous global scale impacts.
- Utilize ESMs to study climate phenomena that use complex mathematical models to simulate the numerous processes and interactions within Earth's system .
- AMOC results from a complex interplay between numerous processes, many are poorly understood or ill-represented in ESMs. Advancing our understanding and ability to simulate AMOC is required.
- Understand the weak AMOC simulated by E3SM and improve the representation. Leverage traditional oceanographic analyses in addition to advanced AI analysis.



PI: Luke Van Roekel, Los Alamos National Laboratory

SciDAC-5 BER/ASCR 2022-2027

ML/AI algorithms for AMOC



Three major efforts leveraging the cutting-edge Machine Learning/Artificial Intelligence (ML/AI) algorithms to analyze and understand AMOC processes beyond traditional oceanographic analyses.

Tackle data challenges including *complexity*, *heterogeneity*, *sparsity*, and *variability*.

- Unsupervised ML – tensor factorizations
- Representation learning – implicit neural networks
- Deep causal analysis – Granger causality and causal anomaly detection

Matrix and Tensor Factorizations

Matrix and Tensor Factorizations: Decomposition of high-dimensional data structures (tensors) into simpler constituent parts (vectors, matrices).

Benefits: Capturing multi-way interactions to reveal hidden patterns and simplify complex data

Methods:

- **SVD/PCA** - Most well known example for matrices
- **Tucker Decomposition** - Decomposition representing tensor as *mode matrices* multiplied by smaller *mixing core tensor*.
- **Others -** Many other methods for compression/feature extraction:
  - Polyadic decomposition
  - Tensor train
  - Tensor networks

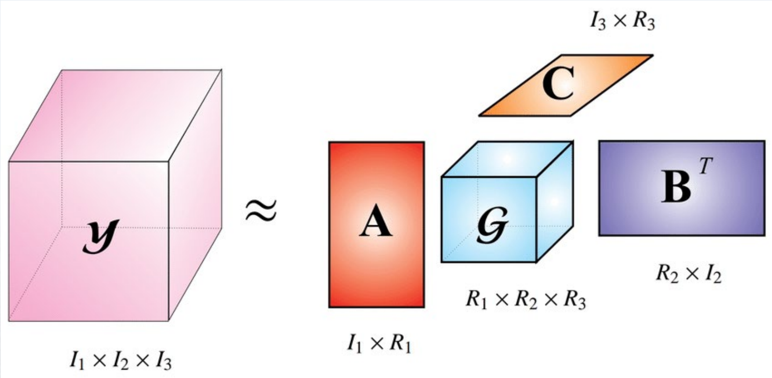
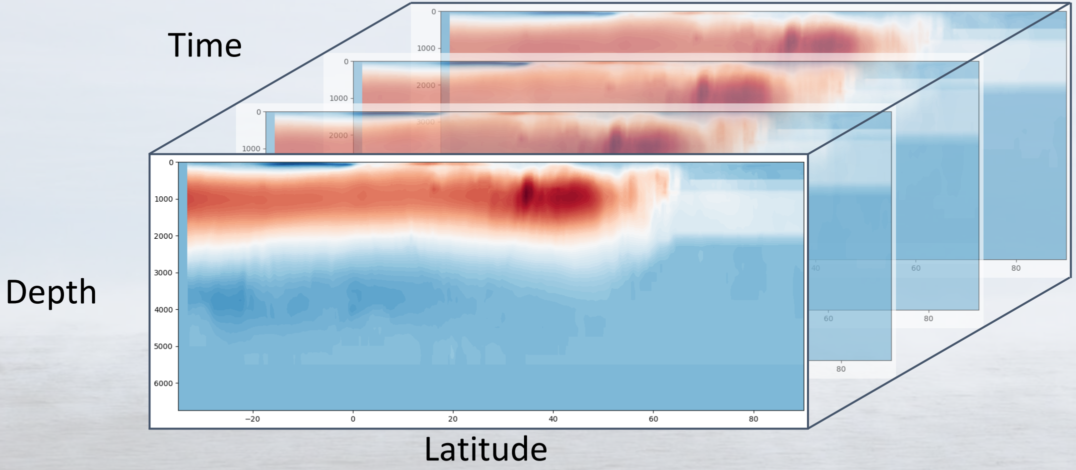


Diagram of the Tucker decomposition

Tucker Application to AMOC



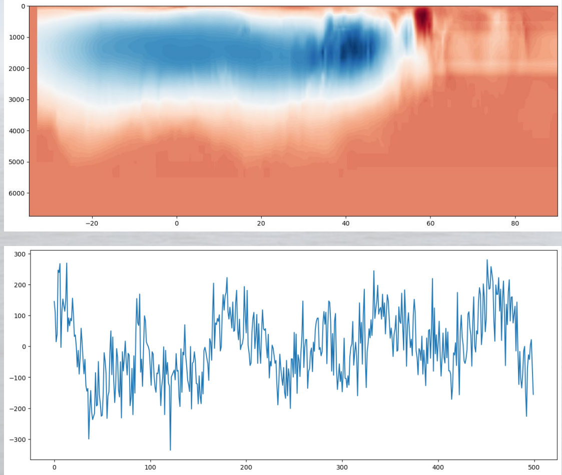
- The Atlantic Meridional Overturning Circulation (AMOC) is a key component of the global ocean conveyor belt, transporting warm surface waters northward and cooler deep waters southward
- The annually averaged overturning streamfunction in Atlantic Ocean is a 3-way data tensor: (Depth) x (Latitude) x (Time)
- Data investigated is pre-industrial control simulations (CESM2)

The Tucker decomposition extracts interactions between depth, latitude and time

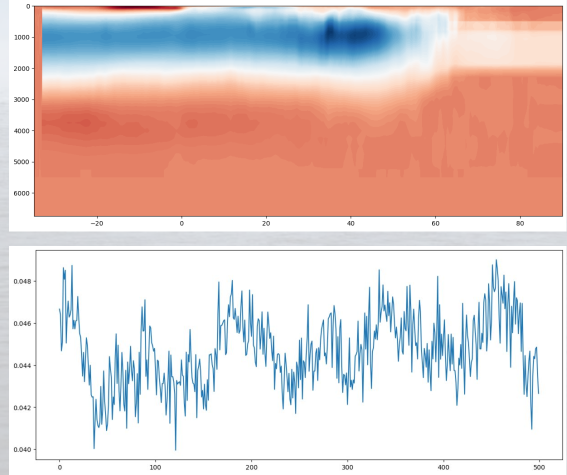
PCA vs Tucker

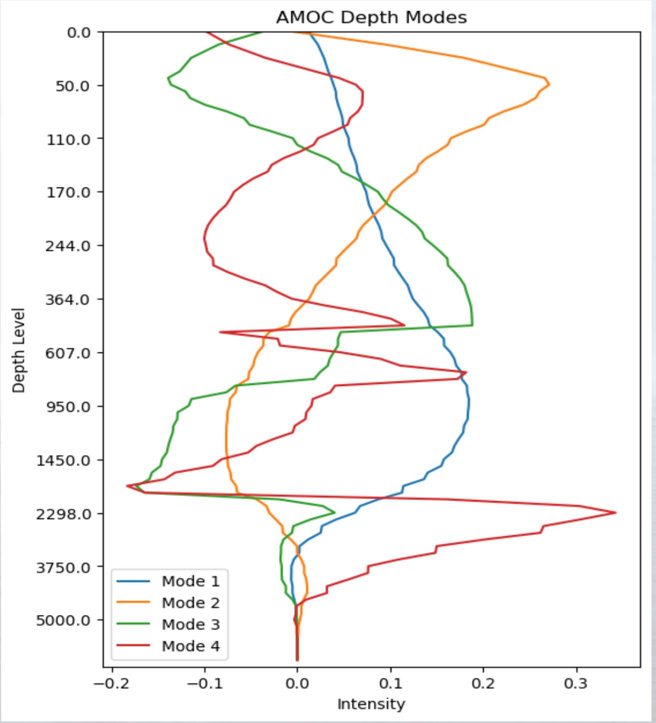
By <u>combining depth and latitude modes</u>, Tucker decomposition recovers dominant modes of spatial variability (PCA):

PCA - First spatial and temporal modes

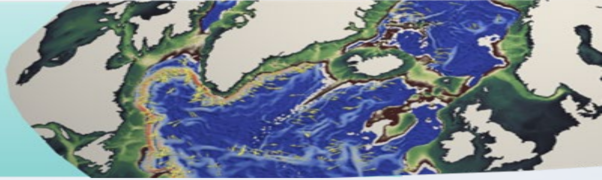


Tucker - First depth/latitude and temporal mode





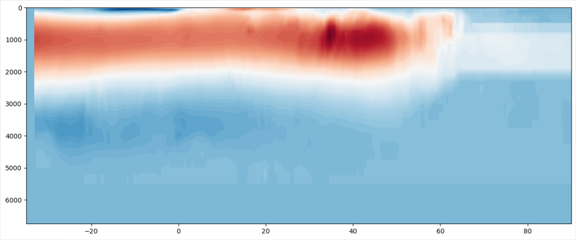
Tucker can extract **additional information** - e.g. depth mode strength



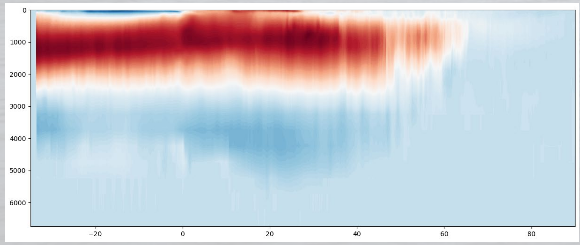
Model Comparison via Tensor Factorization

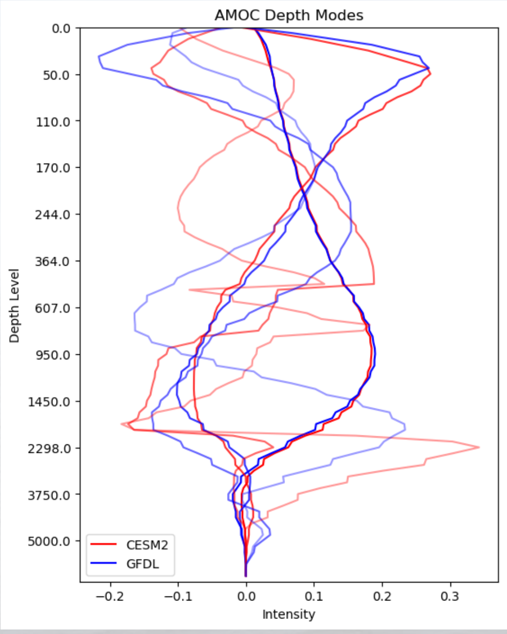
Pick two models → Discover modes → Compare Modes

CESM2



GFDL-CM4

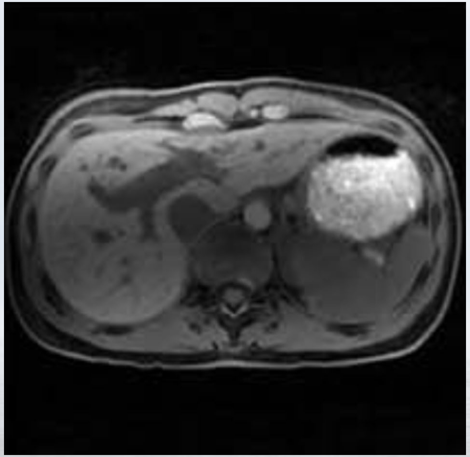


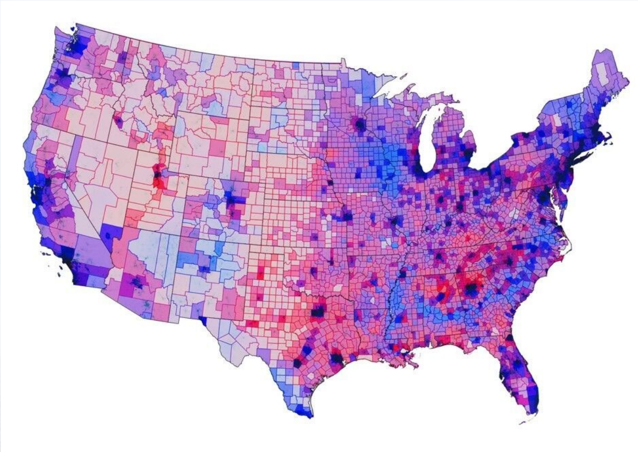


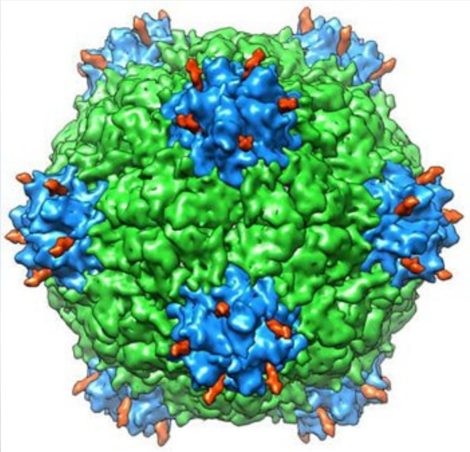
Comparing modes gives insights into the differences in the variability of AMOC between models

- **Procrustes Analysis:** Compare similarity of latent spaces by rotating, translating, and scaling one to another
- **Canonical Correlation Analysis:** Measure similarity by finding linear combination of modes that are maximally correlated
- **Distributional Comparison:** Compare distribution of latent features extracted

Representation Learning





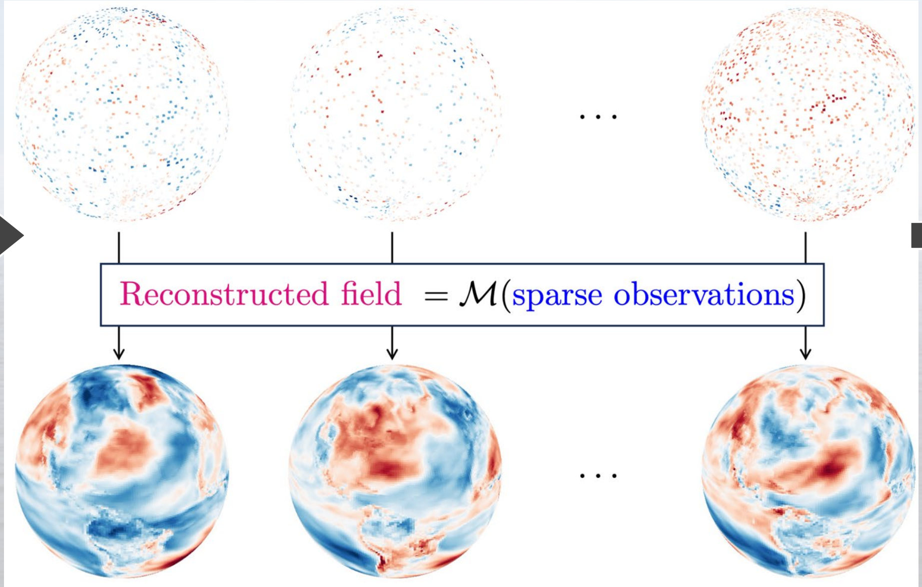


- Medical MRI data: MRI scans inherently provide sparse spatial measurements due to the acquisition process.
- Environmental air quality data: climate data collected from various sources may have uneven coverage, resulting in sparse data points.
- Material crystallography data: sparse data sets can arise in X-ray crystallography experiments, where diffraction measurements are made at discrete points on a crystal.

Challenge

Field reconstruction of complex physical time-evolving field from sparse measurements has been a longstanding challenge.

- ❖ Local
- ❖ Discrete
- ❖ Limited

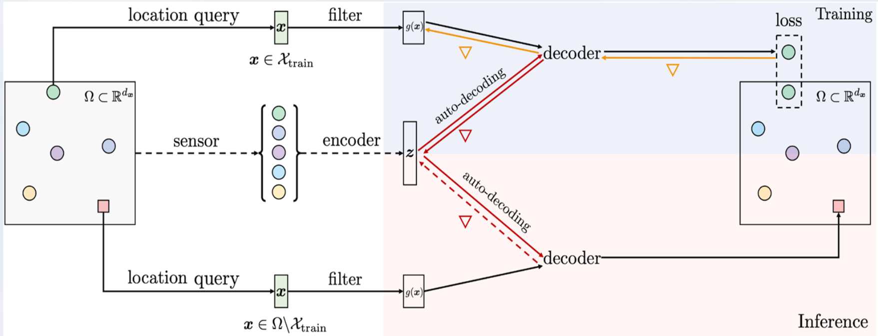


- ❖ Global
- ❖ Continuous
- ❖ Infinite

MMGN Model

We propose **Multiplicative and Modulated Gabor Network (MMGN)**

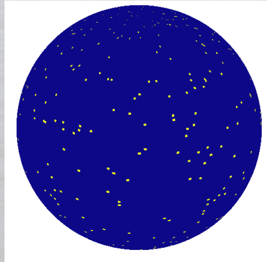
- Encoder: auto-decoding
- Decoder: Gabor filter + multiplicative neural network

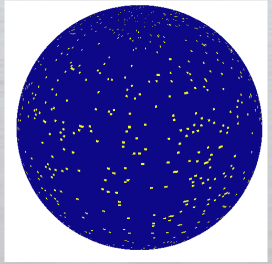


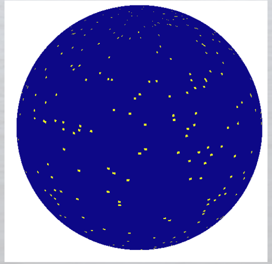
Our model is able to handle a variety of **uncertainties**:

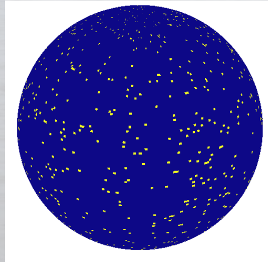
- Uncertainty in the *number* of sensors (possibly malfunctioning)
- Uncertainty in the *position* of sensors due to unknown external forces, like those encountered in oceanic conditions

Four scenarios of sparsity:

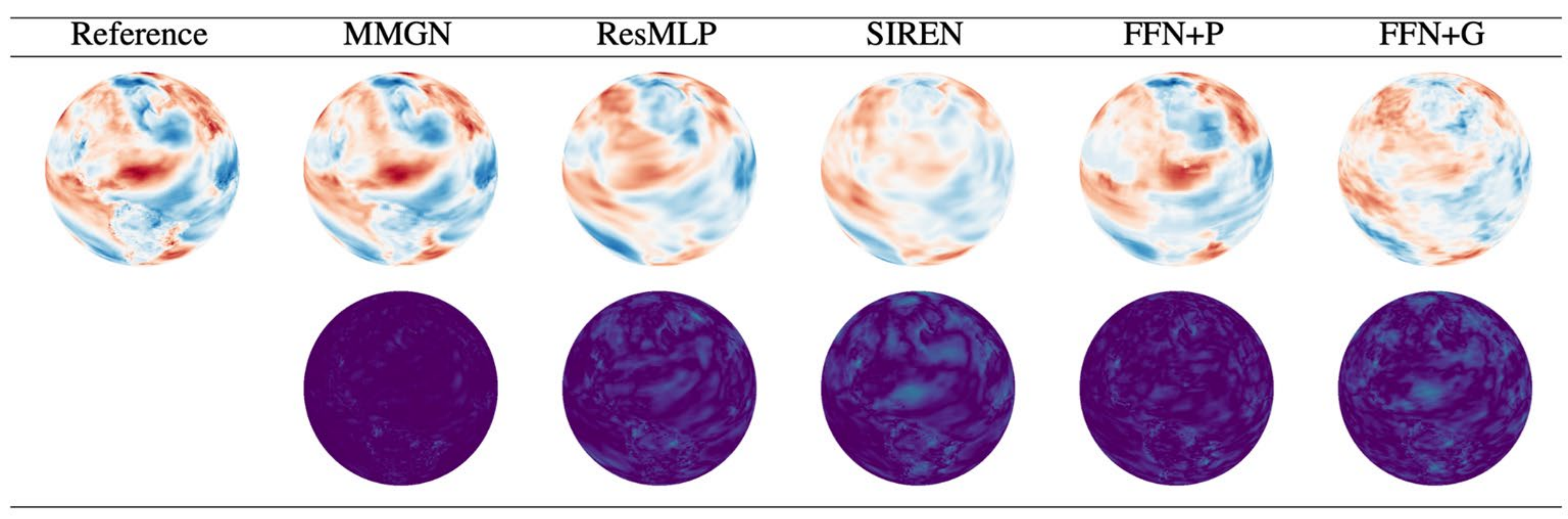






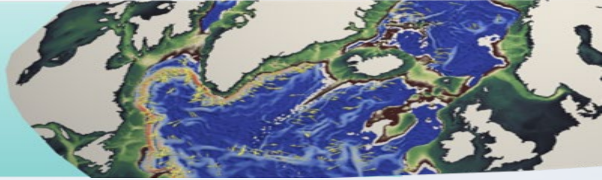


Qualitative Results



Visualizations of true and reconstructed fields. The first column displays the ground truth, the first row showcases predictions from different models, and the second row presents corresponding error maps relative to the reference data. In the error maps, darker pixels indicate lower error levels.

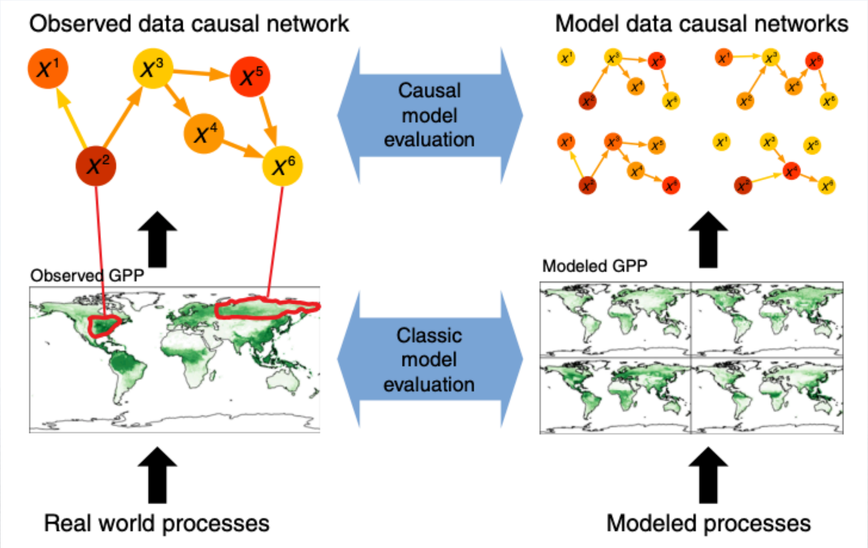
Quantitative Results



Model	Simulation-based Data Task 1	Simulation-based Data Task 2	Simulation-based Data Task 3	Simulation-based Data Task 4	Satellite-based Data Task 1	Satellite-based Data Task 2	Satellite-based Data Task 3	Satellite-based Data Task 4
	Sampling ratio $s = 5\%$				Sampling ratio $s = 0.1\%$			
ResMLP	1.951e-2	1.672e-2	1.901e-2	1.468e-2	1.717e-3	1.601e-3	1.179e-3	1.282e-3
SIREN	2.483e-2	2.457e-2	2.730e-1	2.455e-2	3.129e-1	4.398e-2	1.304e-2	9.338e-2
FFN+P	2.974e-2	1.121e-2	1.495e-2	8.927e-3	2.917e-3	2.392e-3	7.912e-4	7.565e-4
FFN+G	2.943e-2	1.948e-2	1.980e-2	1.426e-2	4.904e-3	7.969e-3	1.005e-3	1.044e-3
MMGN	**4.244e-3**	**4.731e-3**	**3.148e-3**	**3.927e-3**	**1.073e-3**	**1.131e-3**	**6.309e-4**	**6.298e-4**
Promotion	78.24%	57.79%	78.94%	56.01%	37.51%	29.35%	20.26%	16.74%

MSE (mean squared error) is recorded. A smaller MSE denotes superior performance. For clarity, we highlight the best result in bold and underline the second-best. We have also included the promotion metric, which indicates the reduction in relative error compared to the second-best model for each task.

Deep Causal Analysis

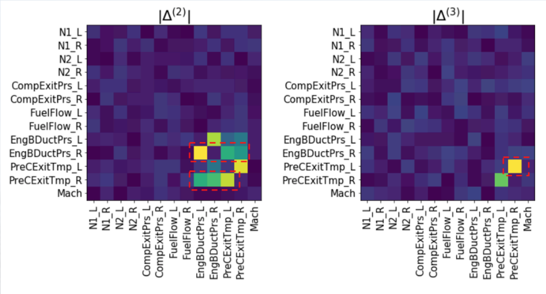


Picture credit: Runge, J., Bathiany, S., Bollt, E. *et al.* Inferring causation from time series in Earth system sciences. *Nat Commun* **10**, 2553 (2019).

- Complex interactions between potential AMOC drivers that occur over a wide range of spatio-temporal scales
- Identify and understand direct and indirect causal pathways for improving the simulation of AMOC in E3SM.
- Scalable and non-linear causal analysis is required.

Deep Causal Analysis

- Granger causality (GC): looking at the causation based on the effects of predictability.
- Our solutions: deep learning based Granger causal inference frameworks capable of exploring the underlying causal graph (variable associating graph) of multivariate time series data.



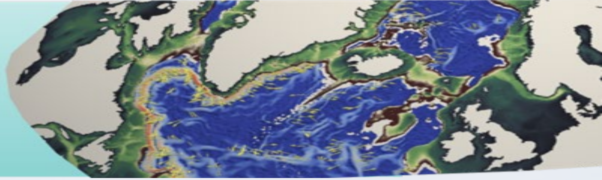
TCCL: A Time Series Classification with Nonlinear Granger Causality Learning.

DAVAC: Detection and Diagnosis of Anomaly with Variable Association.

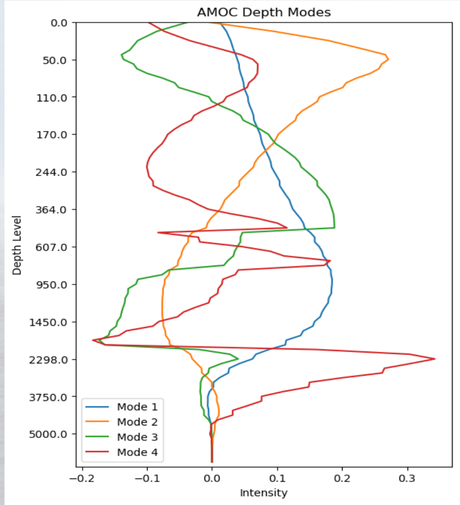
TCCL: 1) outperforms existing classification methods on *imbalanced* time series learning on full flight data; 2) learns the nonlinear Granger causal graph of each *class*, without any predefined kernel; and 3) interpret the difference between time series classes with the *change* of Granger causal graphs.

DAVAC: 1) learns the variable association in normal time series; 2) detect any anomaly in inference mode; and 3) discover how association changes in the detected anomaly.

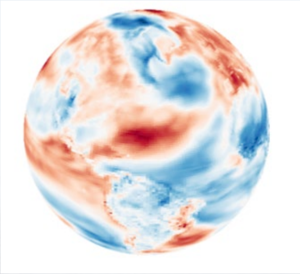
Conclusion

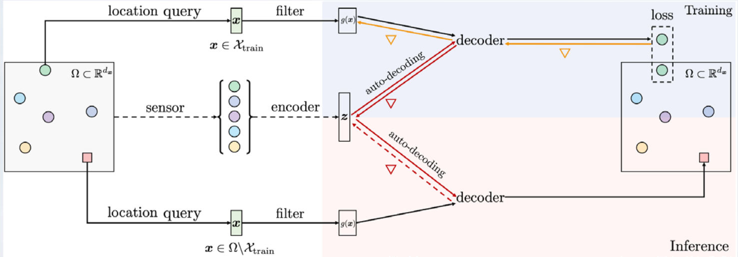


Demonstrate *three* major techniques to tackle data complexity, uncertainty, and variability for data intensive computing.



Tensor Factorizations





Representation Learning using Implicit Neural Networks

Granger Causal Analysis And Anomaly Detection

Wei Xu
Brookhaven NL
xuw@bnl.gov