

From the Cosmos to the Climate

Viviana Acquaviva

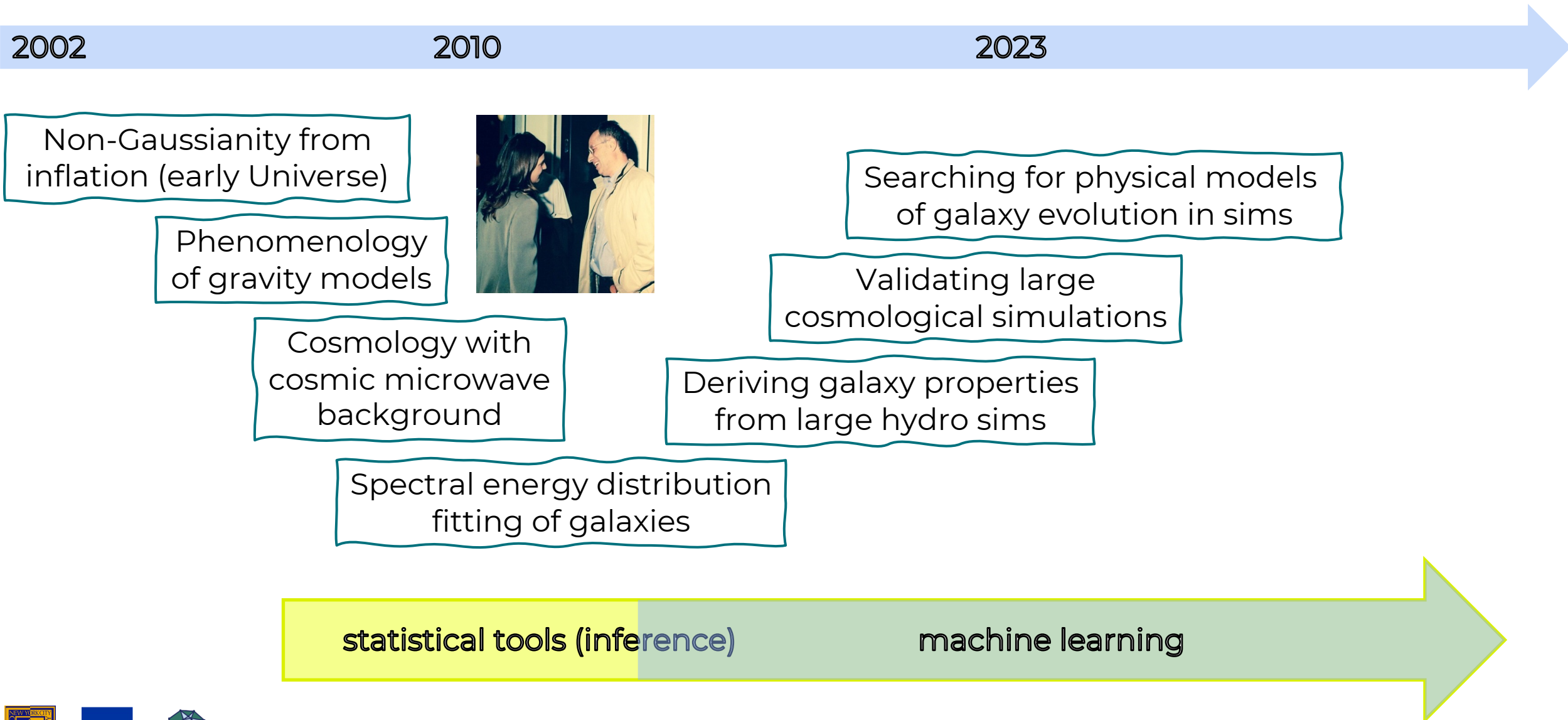
City University of New York
and
Columbia University, Lamont-Doherty Earth Observatory

Sabino for President, Sep 2025



LEAP

I am a former Cosmologist



A striking realization

she really shouldn't
be in Cosmology...
or Astrophysics, just to be safe



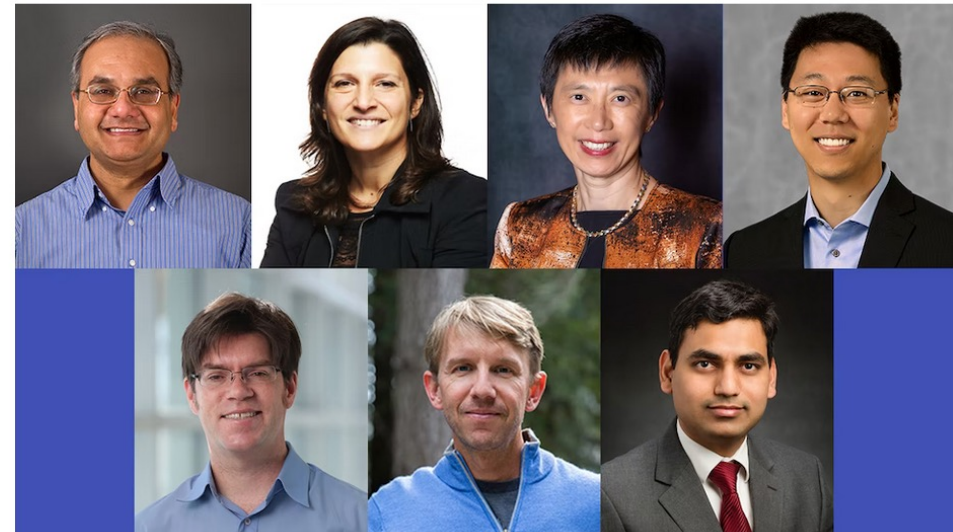
a brilliant
career awaits me

I PIVOTed to Climate Data Science in 2023

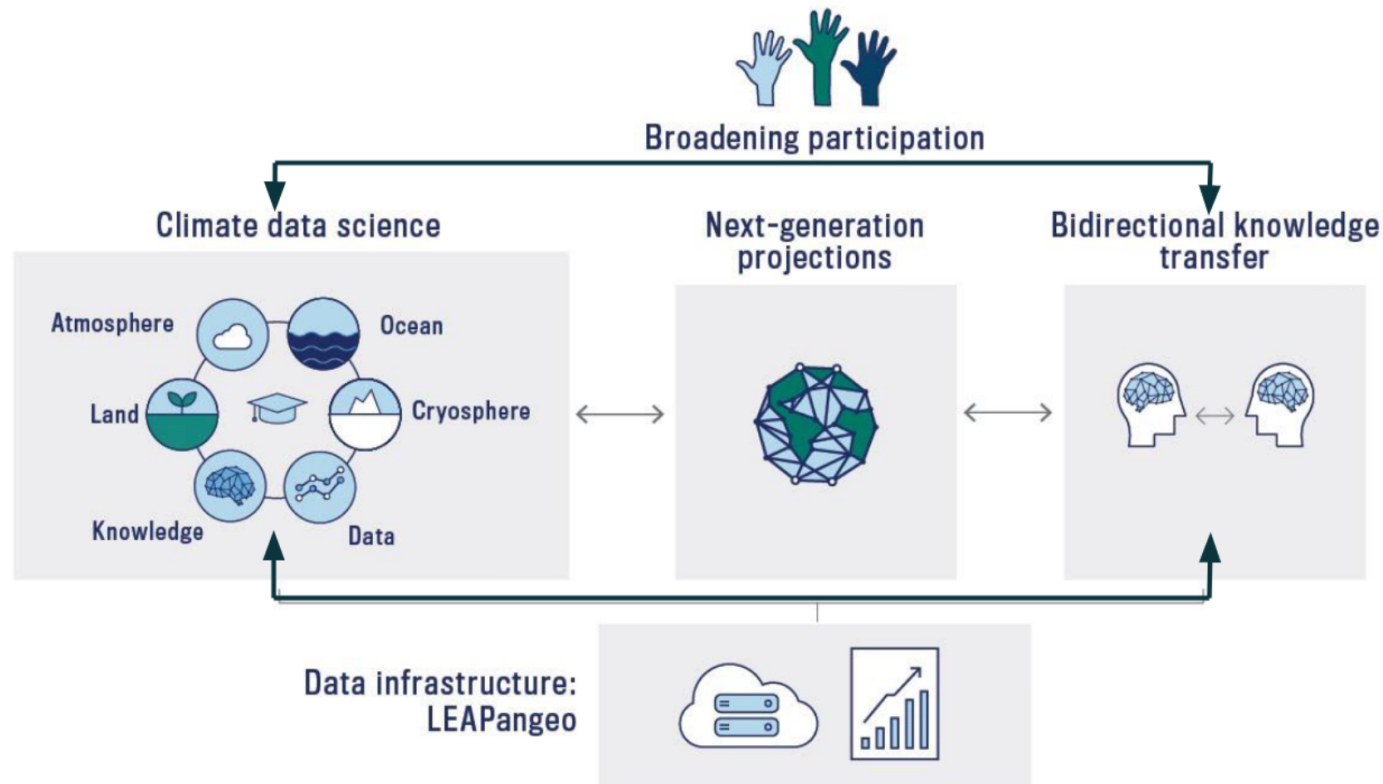


The Simons Foundation is pleased to announce its first class of Pivot Fellows. The program will support the seven accomplished researchers as they apply their talent and expertise to a new field in mathematics or the natural sciences.

Each fellow will receive support for one year of training in their new field under a mentor, followed by the opportunity to apply for up to five years of research funding in the new discipline.

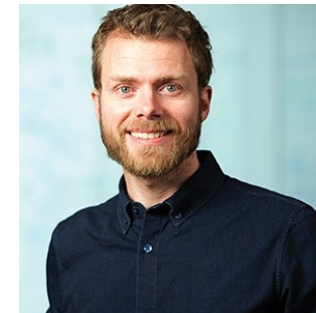


LEAP (Learning the Earth with Artificial Intelligence and Physics)



LEAP is a
NSF-funded (😓)
Science and
Technology Center
at Columbia University

My mentors:



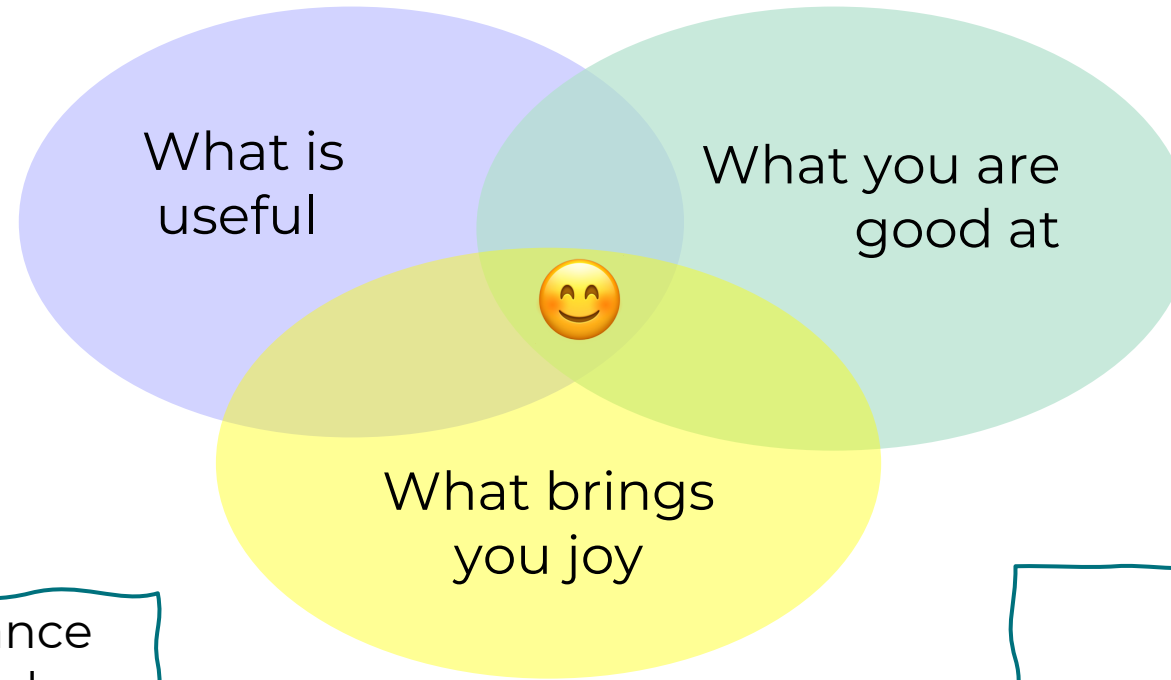
Pierre Gentine



Galen McKinley

A Venn diagram I have spent a lot of time with

(Dr Ayana Elizabeth Johnson,
How to save a planet)



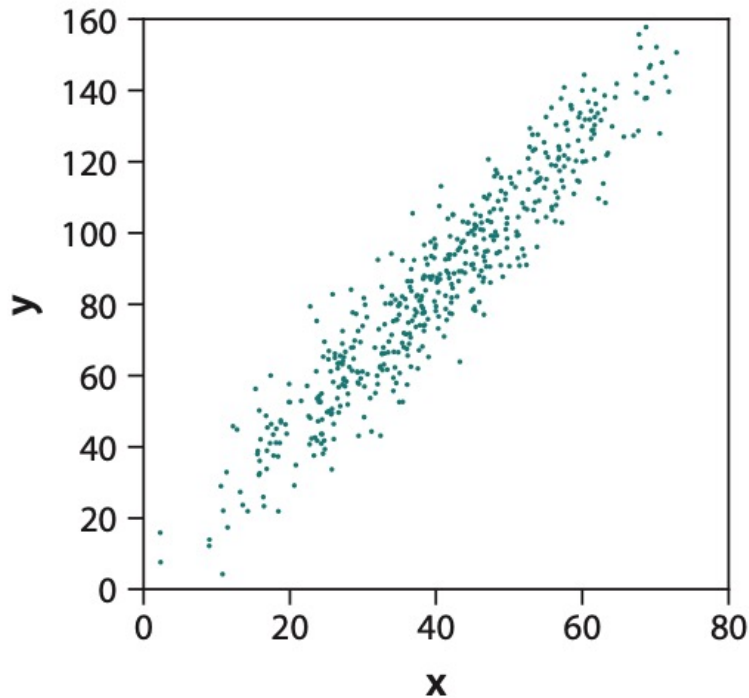
Assessing the performance
of global climate models

AI tools for ethics,
explainability,
and easier workflow

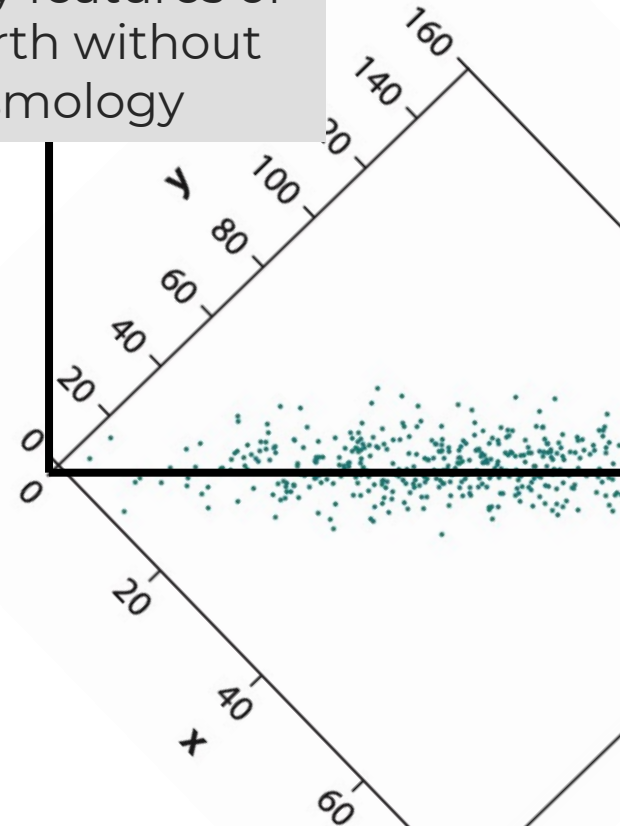
Ocean carbon cycle
from a data perspective

Princ Comp Analysis of my work vs this conference

our work combined



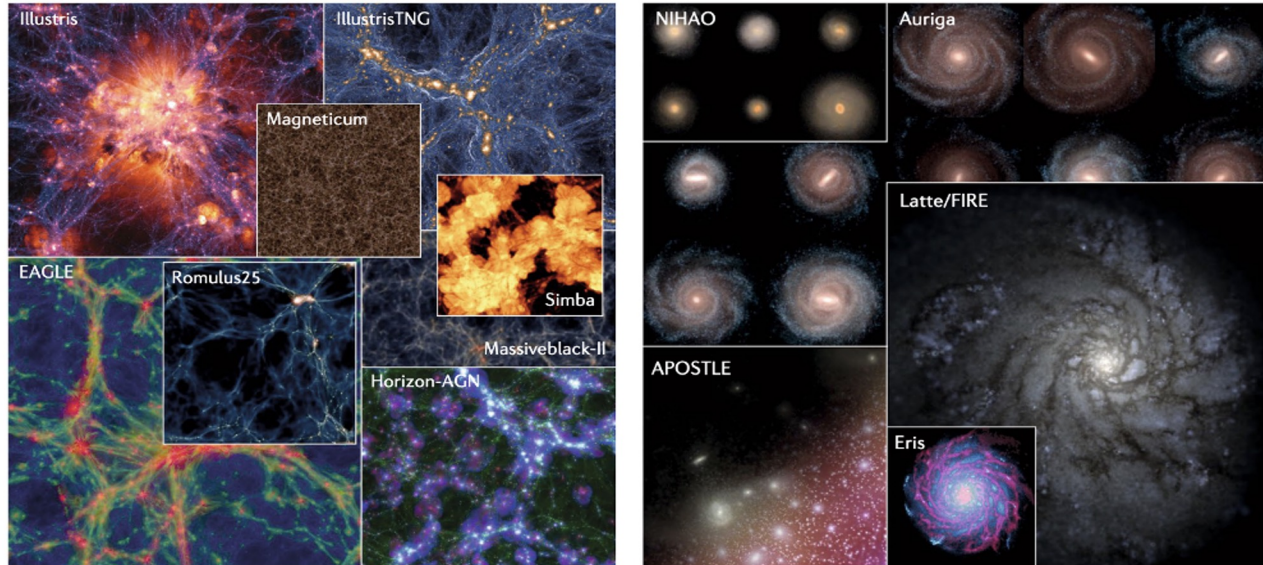
PC 2: Learning the ordinary features of the Earth without Cosmology



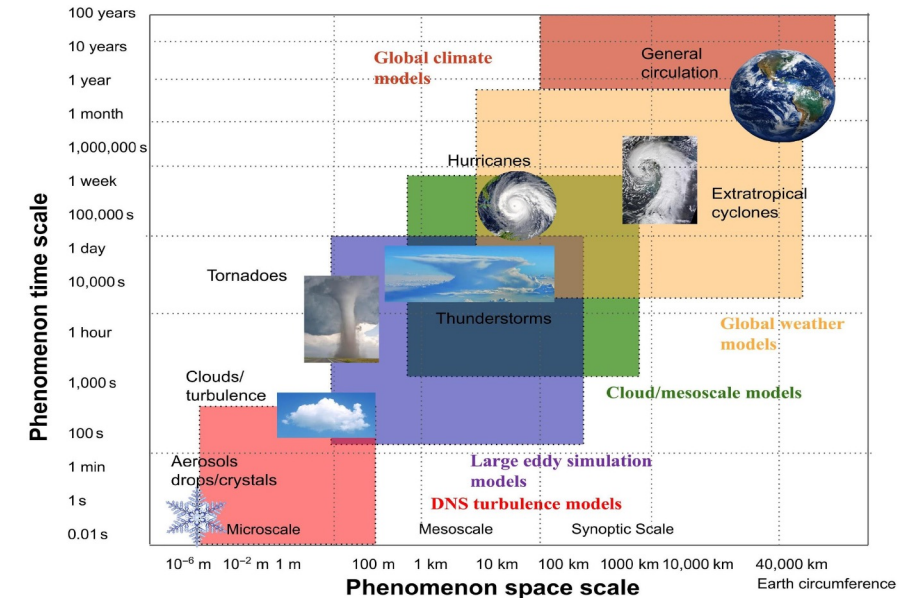
PC 1: Learning the deep mysteries of the Universe with Cosmology

Also realized that if people made a conference for my retirement, it would be a mix of people who never met before!

Nonetheless, I am a bottle half-full person



Vogelsberger et al 2020



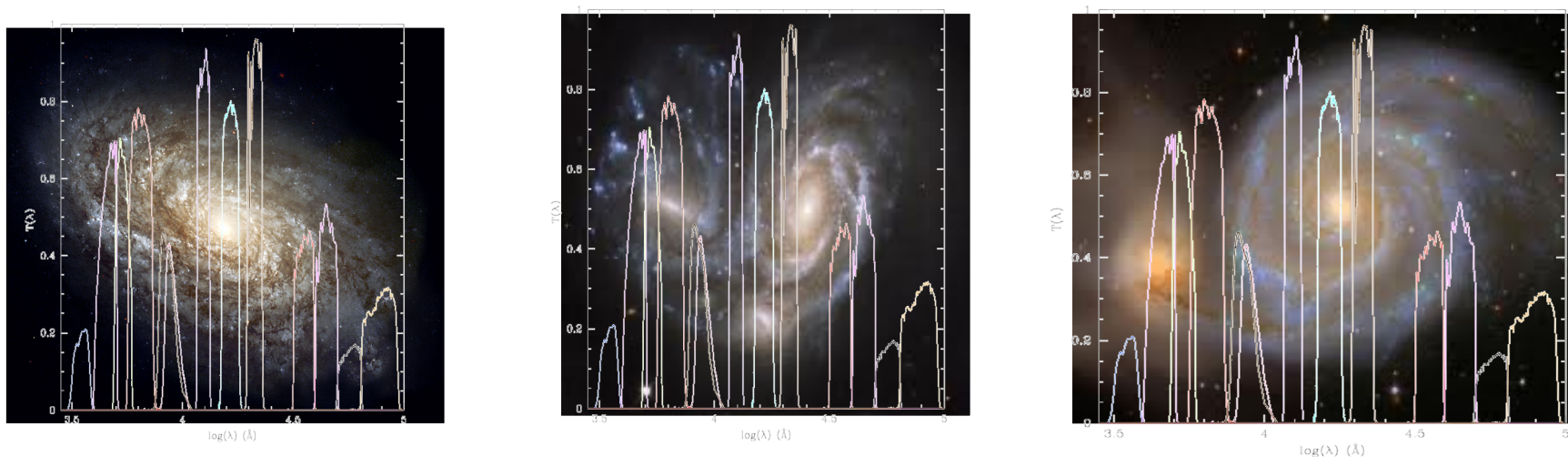
Gettelman et al 2022

- Heavily relying on simulations
- Multi-scale, interactive system
- Resolution is a challenge; subgrid processes + merging
- Need to generate parameterizations/physical models

Goals: Tell you a bit about my new work and hope you have good ideas

Take advantage of the location to make controversial statements such as “climate change is real” or “we should apply the scientific method”

Studying galaxies (up to circa 2017)



HAVE: emission chart at different wavelengths

WANT: PHYSICAL PROPERTIES

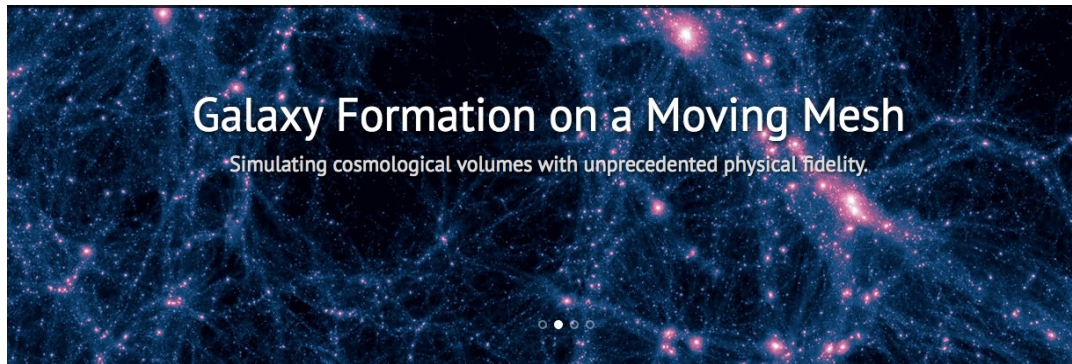
Stellar Mass, Star Formation History, Dust content, Chemical

Spectral Energy Distribution

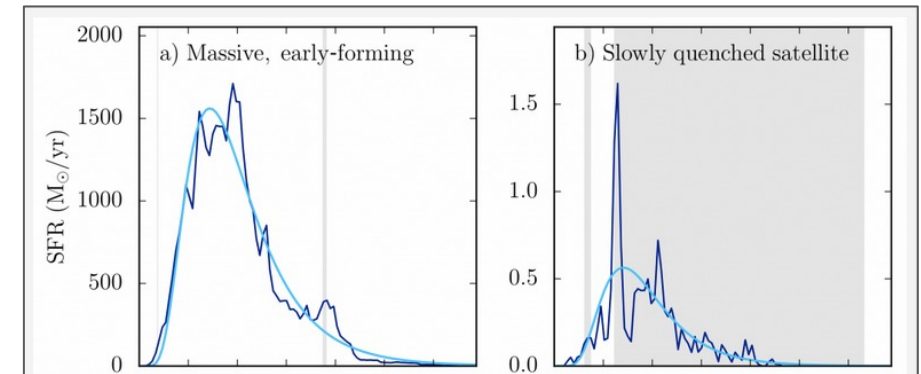


Use machine learning to measure SFHs

Work led by **Chris Lovell**
MNRAS 490 (2019); arXiv:1903.10457



plot by Diemer+ 17



- Take star formation histories of galaxies from two state-of-the-art simulations: Illustris and EAGLE
- Generate realistic spectra using flexible stellar population synthesis (Conroy, FSPS) + self consistent dust attenuation models (Trayford et al 2015)
- Teach a Convolutional Neural Network the connection between spectra (observed) and star formation history (inferred)
- Test behavior across different simulations

The big question (aka the next decade of work in the making)

These models are trained on simulations.

We want to apply them to data.

On real galaxies, for basically any parameters other than redshift,
we don't have labels (ground truth)
to check how we are doing.

What would it take to trust the ML models on data?

Sims = data

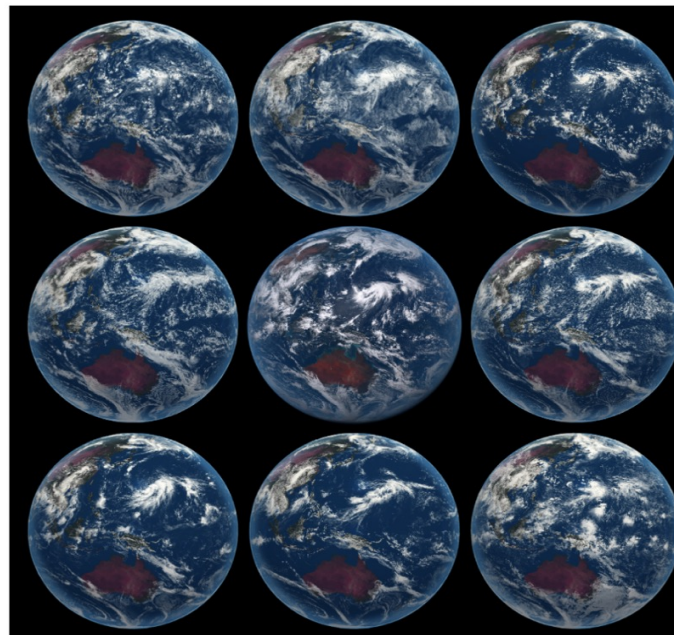


Sims \approx data

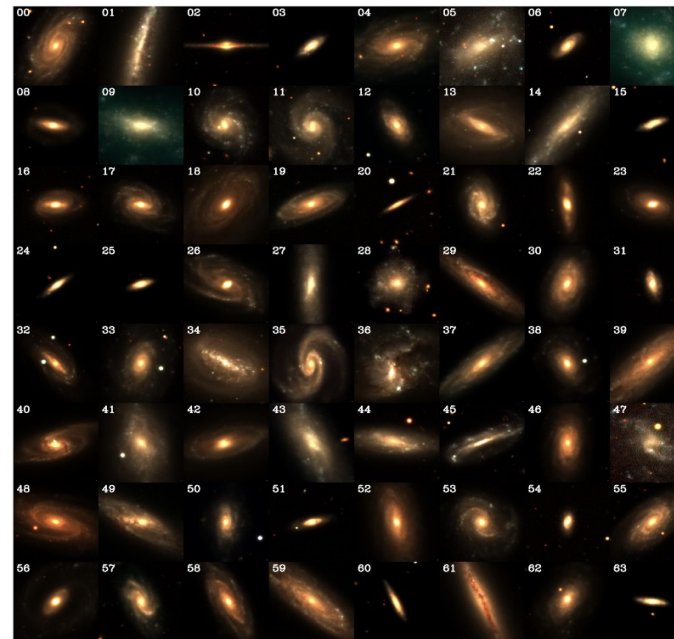


How can we formalize the notion of similarity?

From Cosmology to Climate Science,
I kept thinking of similarity and improved representations.



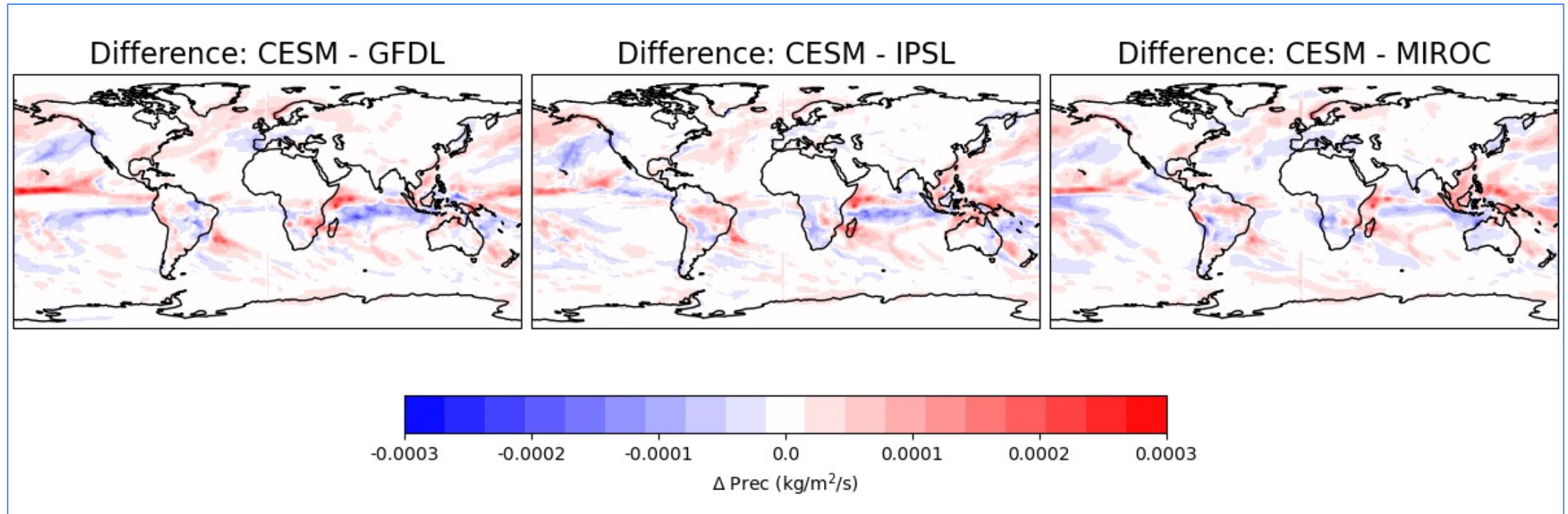
Bjorn et al 2019



Smith et al 2022

When data structures are complex,
what do we measure?

Maps are common outputs of climate models



How do we evaluate differences between models or compare models to data?

New metrics needed!

The “Metrics reloaded” team



Gabriele Accarino
Columbia/LEAP
Post-doc



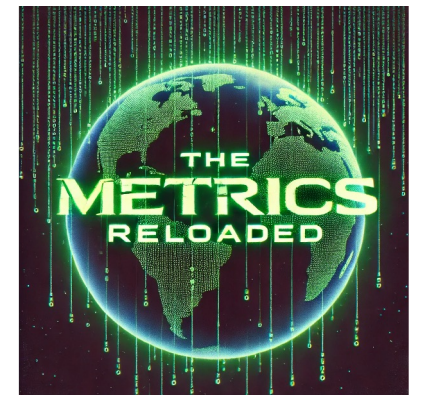
Viviana Acquaviva
CUNY/Columbia



Sara Shamekh
NYU



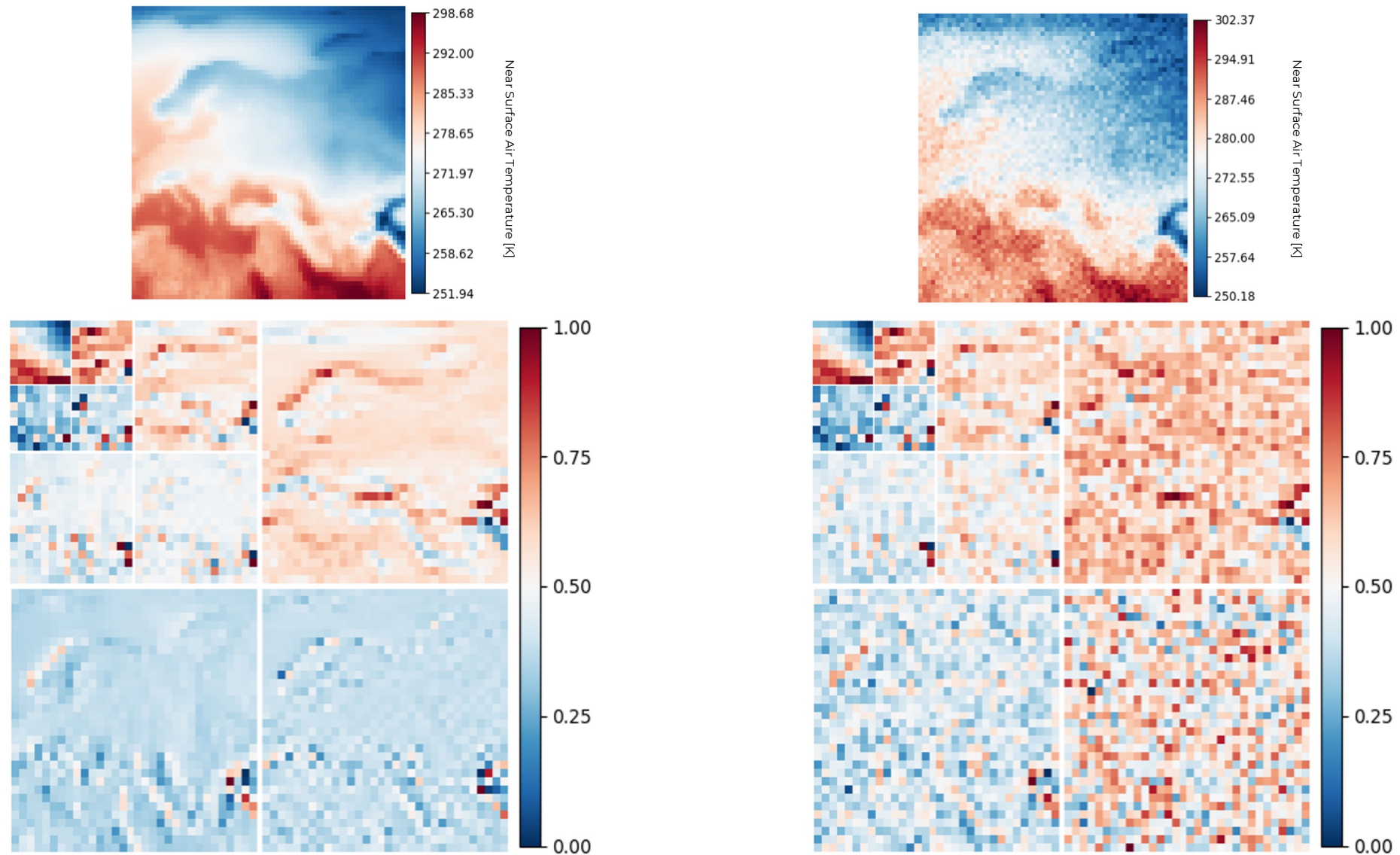
Duncan Watson-Parris
UCSD



Beyond Traditional Metrics

- We need metrics that are
 - Sensitive to perceptual similarities, large-scale bias, spatial structures, and multi-scale variability
 - Tunable, to fit different climate variables
- We found metrics from the Image Processing domain that look promising in assessing perceptual similarity
 - Structural Similarity Index Measure (SSIM):
Brightness + Contrast + Covariance
 - Haar Perceptual Similarity Index (Haar-PSI)
Wavelet-based decomposition of information at different scales

2D Wavelet decomposition to separate scales



Components of Similarity

- We combine three orthogonal axes of similarity at different spatial scales:
 - Magnitude (\mathcal{M}): measures whether the two maps have similar energy magnitude across scales, regardless of where the energy is localized;
 - Displacement (\mathcal{D}): captures whether the spatial distributions of energy are aligned along spatial dimensions (e.g., latitude and longitude), making it sensitive to displacements and invariant to global magnitude differences;
 - Structural (\mathcal{S}): measures whether the structural patterns of wavelet coefficients are preserved, independent of magnitude or exact spatial positioning.



Our Wavelet-based Similarity Metric: WaveSim

decomposition in different levels (physical scales)

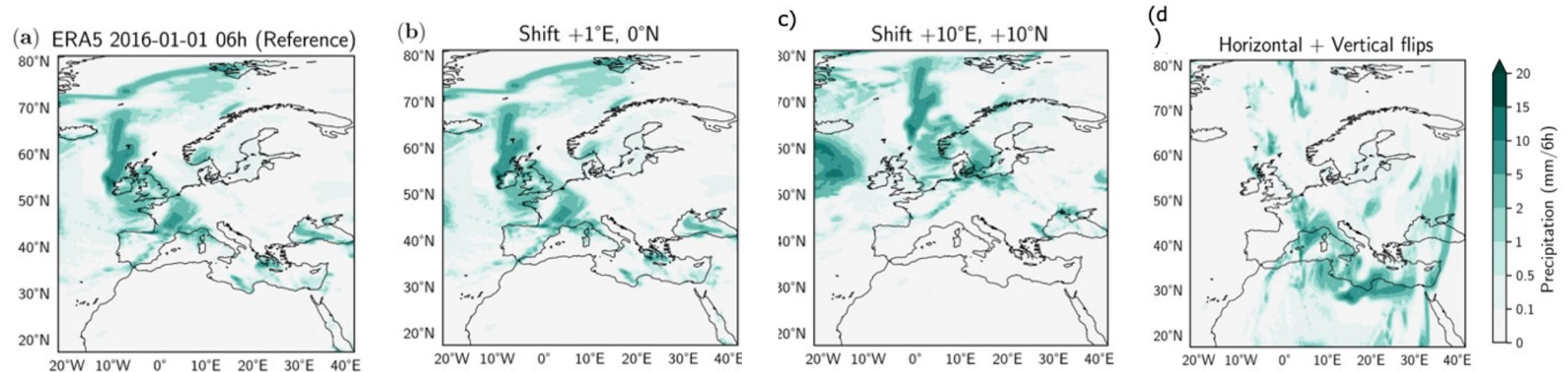
$\alpha, \beta, \text{ and } \gamma$ control the trade-off between the three components

$$WaveSim(X, Y) = \sum_s w_s \cdot \left(\underbrace{\mathcal{M}(X_s, Y_s)^\alpha}_{\text{Magnitude (luminance) component}} \cdot \underbrace{\mathcal{D}(X_s, Y_s)^\beta}_{\text{Displacement component}} \cdot \underbrace{\mathcal{S}(X_s, Y_s)^\gamma}_{\text{Structural component}} \right)$$

weighting scheme can be adjusted to privilege similarity @ desired scale

Accarino, VA et al 2025, in prep

WaveSim on Synthetic Test Cases



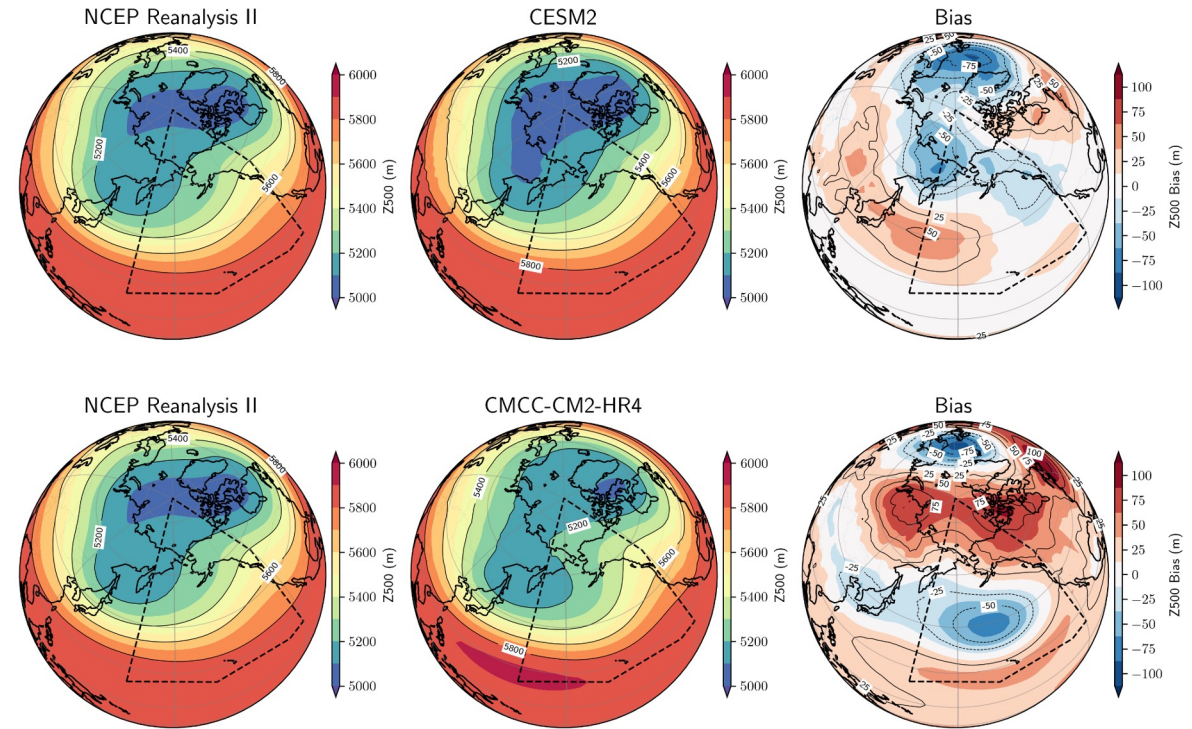
All maps are decomposed across 4 scales; we report Normalized RMSE (NRMSE), Data Structural Similarity Index Measure (DSSIM), and WaveSim scores, with scale-components equally weighted across scales.

Test Case	NRMSE	DSSIM	WaveSim (ours)	WaveSim Components				
				Spatial Resolution: 0.25°; Scales (km): ~50, 100, 200, 360				
(b)	0.949	0.455	0.881	Magnitude	[0.989	0.997	0.938	0.995]
				Displacement	[0.886	0.912	0.916	0.937]
				Structure	[0.994	0.996	0.977	0.974]
(c)	0.876	0.188	0.445	Magnitude	[0.935	0.926	0.998	0.946]
				Displacement	[0.306	0.521	0.536	0.607]
				Structure	[0.964	0.944	0.935	0.955]
(d)	0.893	0.073	0.380	Magnitude	[1.000	1.000	1.000	1.000]
				Displacement	[0.378	0.454	0.426	0.263]
				Structure	[1.000	1.000	1.000	1.000]

Shifts (b–c) affect all components, most strongly displacement, while flips (d) impact only displacement. In contrast, NRMSE overestimates similarity, and DSSIM drops sharply even for small shifts, with extremely low scores for flips.

Evaluating Biases in Earth System Models (ESMs)

- We apply WaveSim to compare DJF-averaged Z500 maps from CMIP6 ESMs (1979–2014) to NCEP Reanalysis II over the dashed region in the figure (32 x 32 grid sub-domain)
- We target 3 scales in the wavelet decomposition, corresponding to ~520, 890 and 1,480 km
- Each component (Magnitude, Displacement, Structure) yields 3 values from the detail coefficients at different scales
- Unlike Normalized Root Mean Squared Error and partly Data Structural Similarity Index Measure, WaveSim is sensitive to the larger CMCC bias, attributing it to differences in magnitude (power) and medium-scale structure



Earth System Model against NCEP Reanalysis II Spatial Resolution: 2.5°	NRMSE	DSSIM	WaveSim (ours)	WaveSim Components Scales (km): ~520, 890, 1480	
CESM2	0.969	0.950	0.836	Magnitude	[0.963 0.927 0.904]
				Displacement	[0.928 0.968 0.965]
				Structure	[0.961 0.952 0.938]
CMCC-CM2-HR4	0.943	0.858	0.656	Magnitude	[0.794 0.834 0.709]
				Displacement	[0.961 0.968 0.964]
				Structure	[0.898 0.906 0.837]

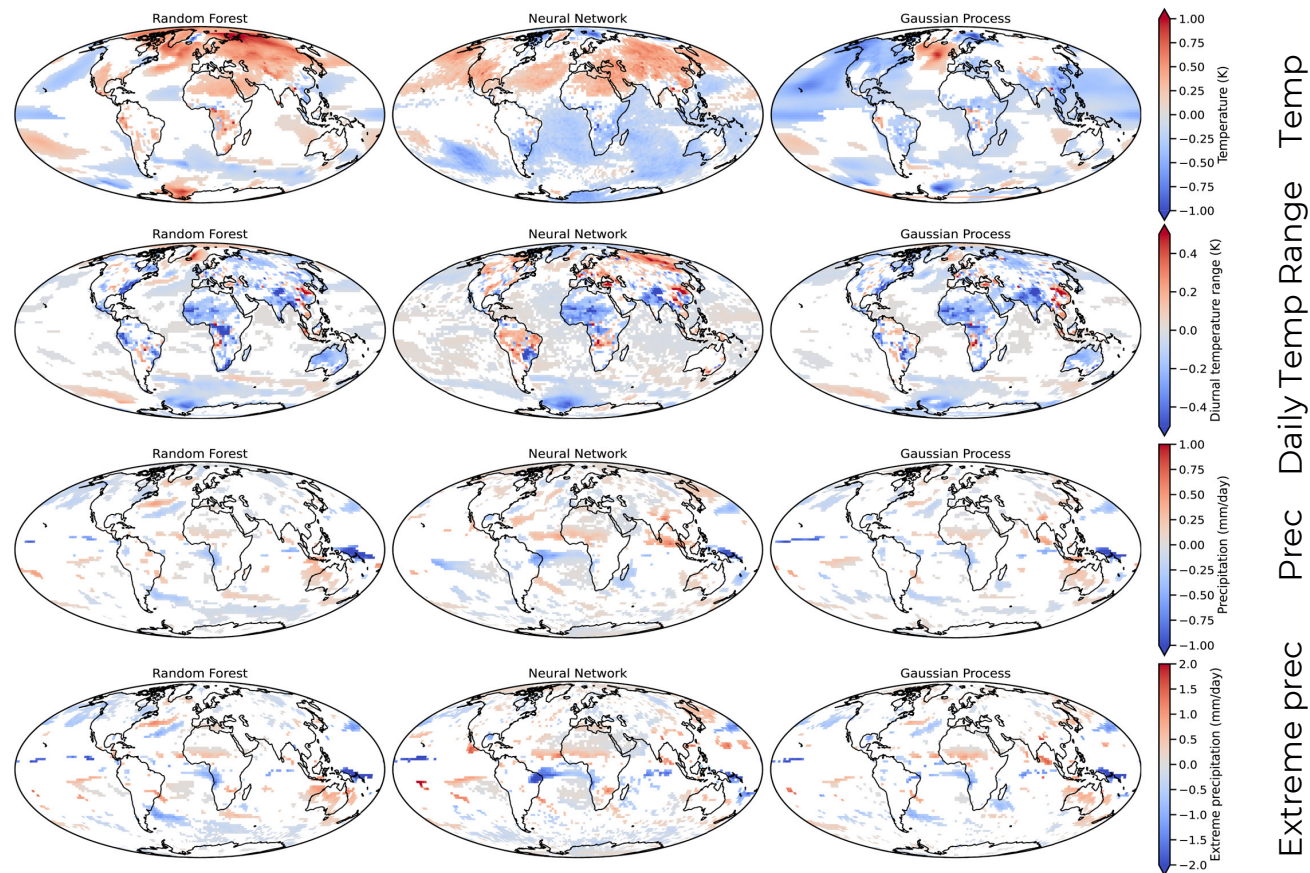
Application 1: Multi-scale Loss Function for Emulators

ClimateBench
(Watson-Parris
et al 2021)

ML emulator
for climate
models

(includes
neural
network)

Random Forest Neural Net Gaussian Proc



Idea:

Use new
metric as
loss function
for NN-based
emulators

Application 2: Open new axes of analysis for AI models



The scorecards below show the skill (measured by the global root mean squared error) of different physical and ML-based methods relative to [ECMWF's IFS HRES](#), one of the [world's best operational weather models](#), on a number of key variables. For a detailed explanation of the different skill metrics and variables, check out the [FAQ](#).

		Geopotential					Temperature					Humidity					Wind Vector				
		500hPa geopotential RMSE [kg ² /m ²]					850hPa temperature RMSE [K]					700hPa specific humidity RMSE [g/kg]					850hPa wind vector RMSE [m/s]				
Physical models	IFS HRES	42	135	304	521	801	0.62	1.16	1.62	2.17	2.63	0.55	0.96	1.27	1.53	1.81	1.69	3.30	5.21	7.13	9.16
	IFS ENS (mean)	42	132	277	439	621	0.65	1.11	1.62	2.17	2.80	0.51	0.84	1.06	1.22	1.38	1.63	2.99	4.45	5.75	6.95
	ERA5-Forecasts	43	142	316	534	811	0.59	1.19	1.87	2.68	3.6	0.53	1.01	1.33	1.59	1.85	1.63	3.41	5.38	7.27	9.25
ML / hybrid models	Pangu-Weather (oper.)	45	136	300	510	785	0.65	1.09	1.74	2.54	3.55	0.53	0.85	1.05	1.45	1.76	1.71	3.03	4.86	6.77	8.84
	GraphCast (oper.)	40	124	276	477	754	0.53	0.93	1.56	2.36	3.40	0.48	0.76	1.03	1.29	1.60	1.48	2.74	4.52	6.42	8.57
	GenCast (oper.) (mean)	41	129	274	440	623	0.55	0.96	1.51	2.11	2.76	0.49	0.78	1.00	1.18	1.35	1.54	2.80	4.30	5.67	6.90
	Keisler (2022)	66	174	345	544	787	0.81	1.22	1.87	2.63	3.55	0.65	0.94	1.19	1.41	1.65	2.27	3.51	5.18	6.87	8.64
	Pangu-Weather	44	133	294	501	778	0.62	1.05	1.71	2.54	3.54	0.53	0.88	1.19	1.47	1.79	1.66	3.01	4.83	6.73	8.81
	GraphCast	39	124	274	467	731	0.51	0.94	1.55	2.33	3.36	0.47	0.79	1.06	1.30	1.59	1.42	2.76	4.45	6.23	8.20
	FuXi	40	125	277	433	631	0.54	0.97	1.59	2.14	2.91						1.47	2.80	4.51	5.66	7.04
	NeuralGCM 0.7	37	115	267	469	751	0.54	0.97	1.58	2.38	3.42	0.48	0.83	1.12	1.40	1.71	1.49	2.81	4.59	6.51	8.66
	NeuralGCM ENS (mean)	43	126	266	424	606	0.65	1.02	1.53	2.10	2.75	0.54	0.81	1.02	1.19	1.37	1.76	2.89	4.29	5.60	6.84
	GenCast (mean)	39	123	262	420	600	0.51	0.94	1.48	2.07	2.73	0.49	0.80	1.02	1.19	1.37	1.49	2.78	4.23	5.55	6.83

Screenshot



LEAP

From sparse data to full spatio-temporal fields: surface ocean carbon and beyond (PIVOT Research Award from Simons Foundation)



Galen McKinley



Amanda Fay



Thea H Heimdal



Abby Shaum



Tian Zheng



Romina Wild

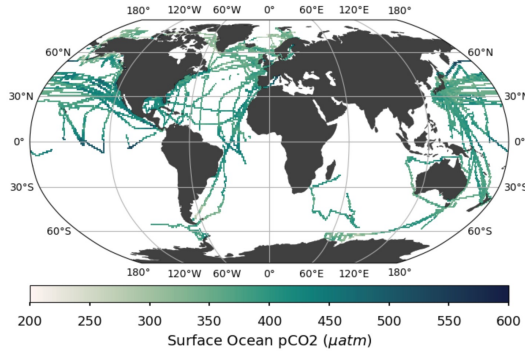


Alessandro Laio

I'll be hiring!

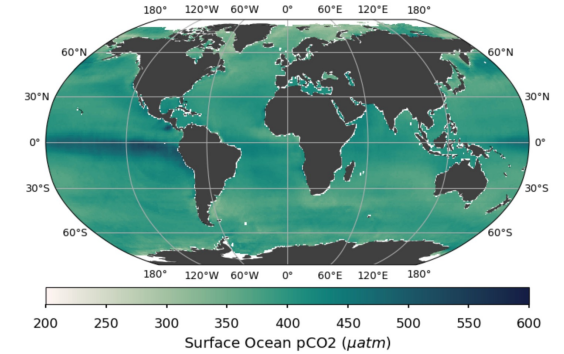


The problem and why we care

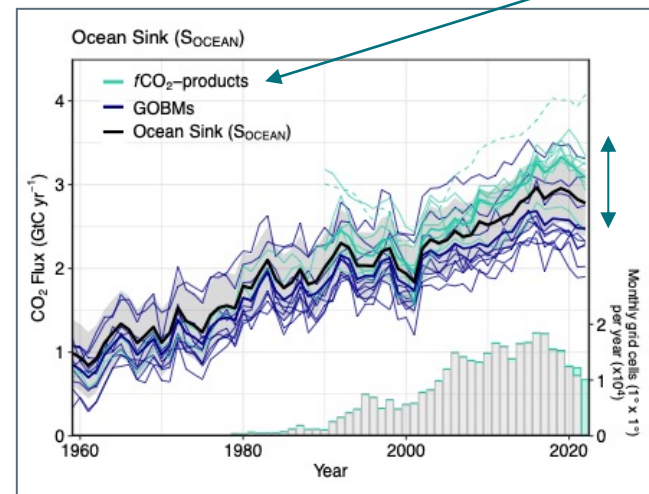
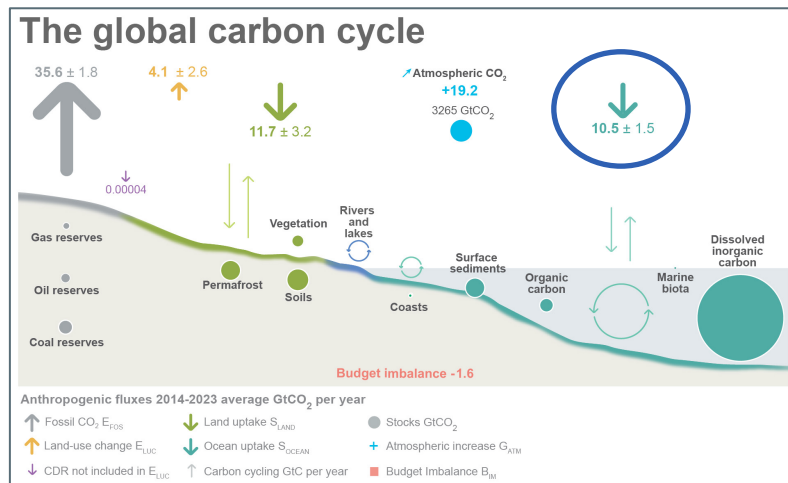


existing $p\text{CO}_2$ observations:
sparse + biased
in space and time

RECONSTRUCTION



dense $p\text{CO}_2$ field in space and time




large uncertainties
hinder plans
for mitigation and
adaptation to climate change

improved reconstruction =
smaller uncertainties

Friedlingstein et al 2024, Global Carbon Budget

improved representation with Tiny Ocean



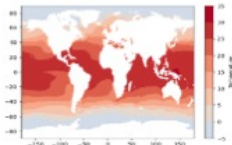
feature engineering



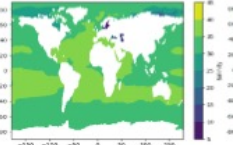
train ML model

features

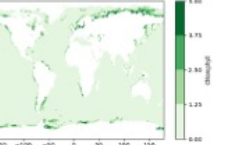
temperature



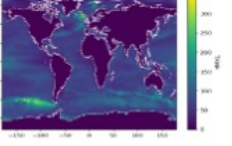
salinity



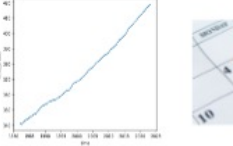
chlorophyll-a




mix layer depth




atmos CO2



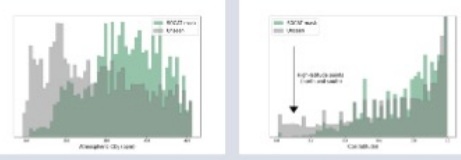
t of year



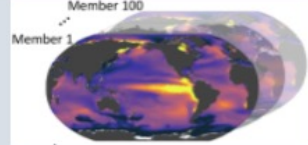
lat/lon



domain adaptation

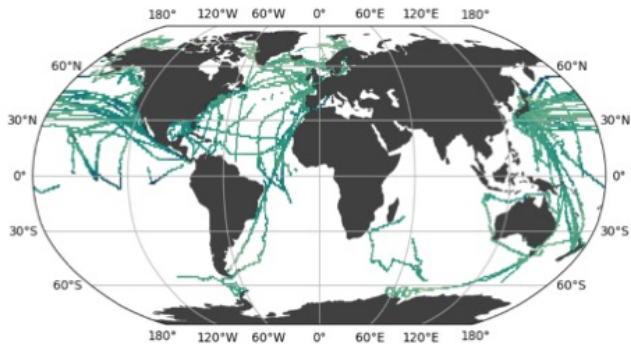


validation through ESMs

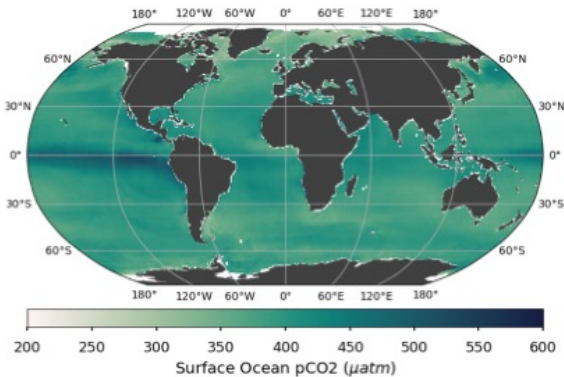



use ML model to fill gaps

sparse pCO2 observations in space / time



dense pCO2 field



What is a good representation?

Representation = Metric or distance or combination of feature variables

PHYSICAL

We seek a representation that lives in the space of **physical variables**

To keep the results maximally **interpretable** (need to communicate across communities)

PROBLEM AWARE

We seek a representation that **correlates well with the target** (pCO₂)

To improve **skill** and **generalization** properties of ML models

SMALL, POSSIBLY TINY 😊

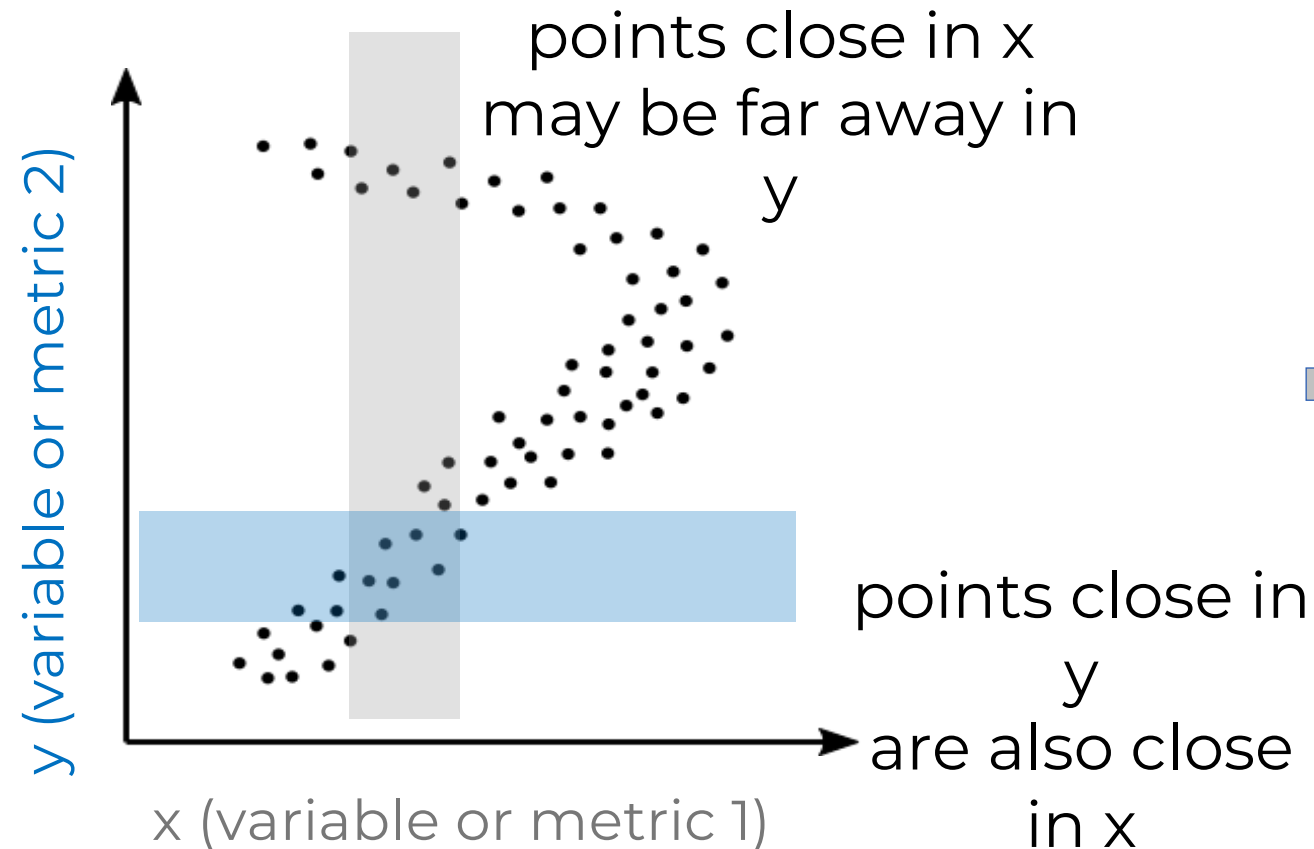
We seek a representation that is as **small** as possible (fewer features and fewer data points)

For computational **agility**

To aid **visualization** and **physical interpretation** of data and relationships

How to find informative metrics

How do we define an informative metric?

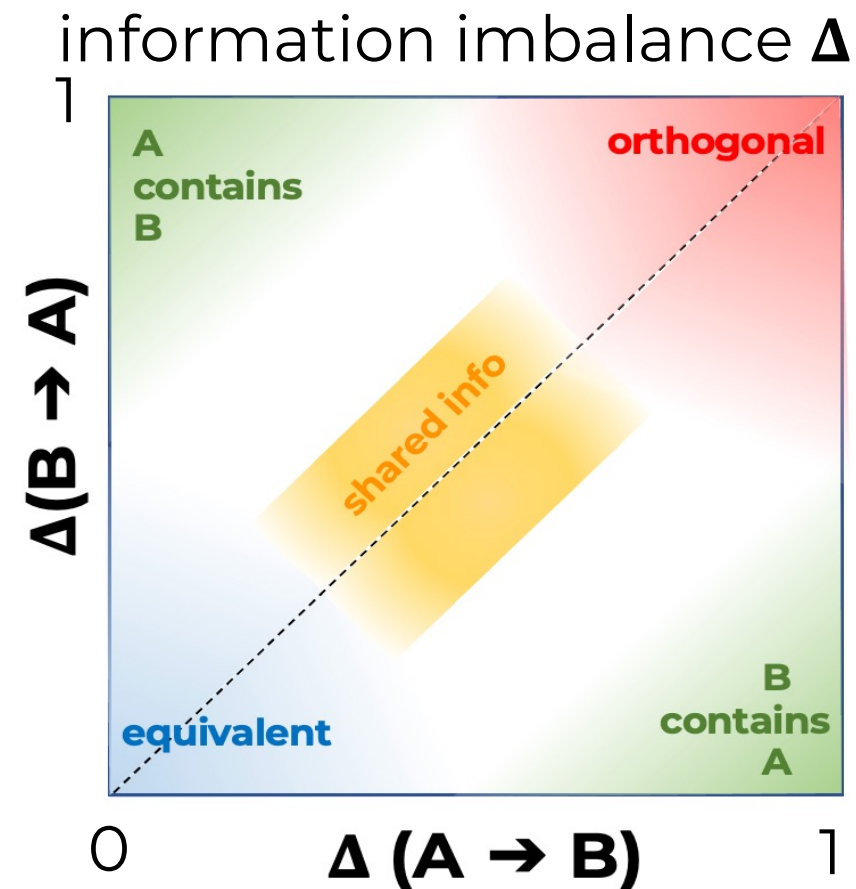


y is more informative than x:
vicinity in y
is predictive of
vicinity in x,
but not viceversa

Information imbalance

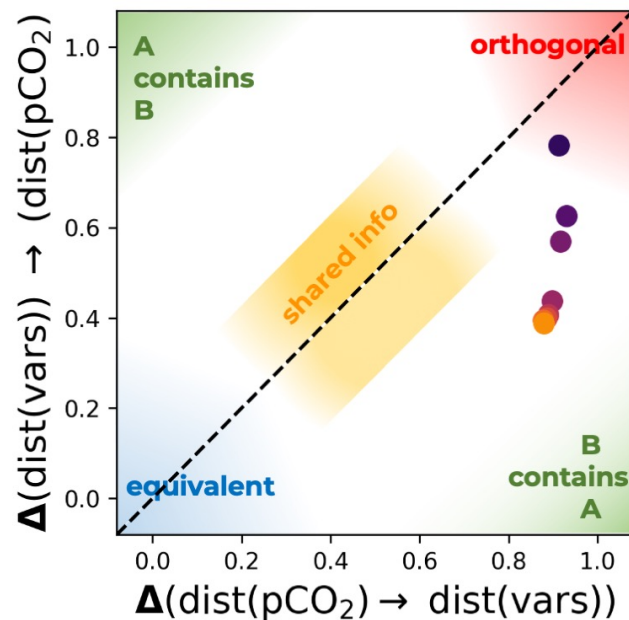
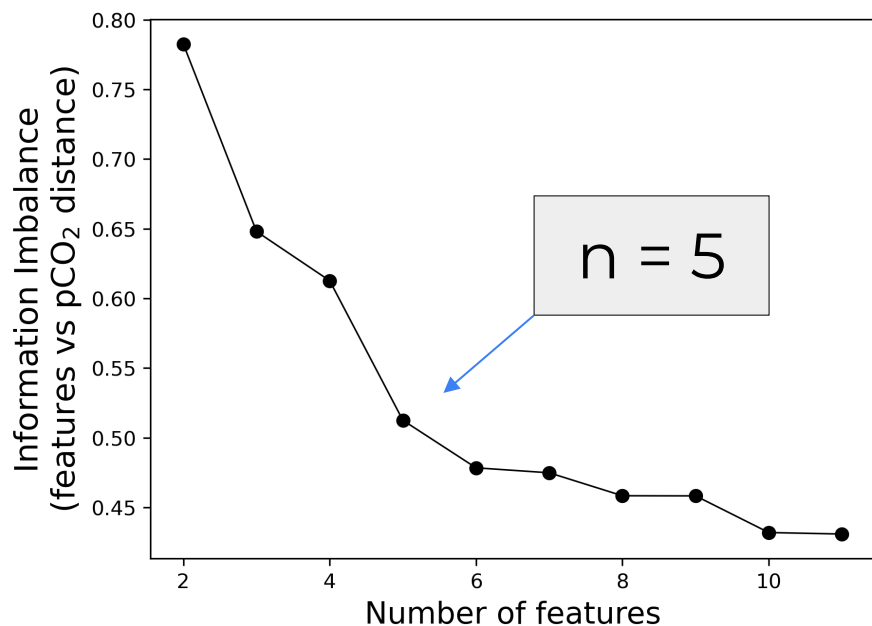
The information imbalance between A and B, $\Delta(A \rightarrow B)$ is the average of the ranks according to distance B of the first neighbors according to distance A

$$\Delta(A \rightarrow B) = \frac{2}{N} \langle r_B | r_A = 1 \rangle = \frac{2}{N} \sum_{i,j: r_{ij}^A=1} \frac{r_{ij}^B}{N}$$



Best metric search as f(dimension)

We explore subspaces of 11 features (A, B, C, T_0 , T_1 , xCO₂, sst, sss, mld, chl, sst anom) to find the one **with the smallest information imbalance vs pCO₂** for any dimension



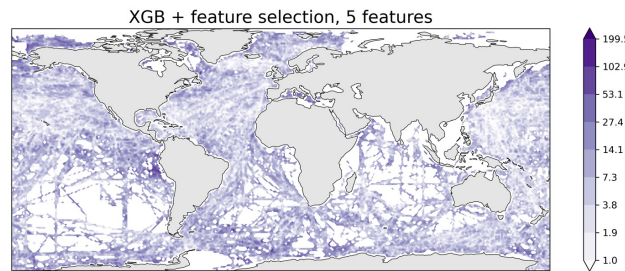
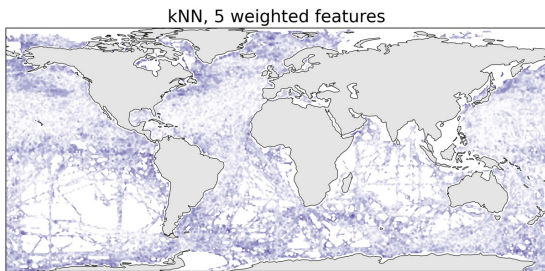
- pCO₂ → weighted ['A', 'sst']
- pCO₂ → weighted ['A', 'C', 'sst']
- pCO₂ → weighted ['T1', 'A', 'C', 'sst']
- pCO₂ → weighted ['T1', 'A', 'C', 'xco2', 'sst']
- pCO₂ → weighted ['T0', 'T1', 'A', 'C', 'xco2', 'sst']

Using the differentiable information imbalance
(Wild et al 2025, Nature Communication),
we find an optimal metric
in a $D=5$ space of physical variables

What can we do with the improved representation?

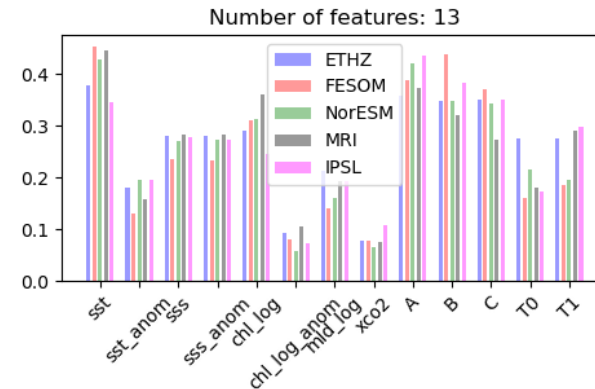
Define **feature importance** that converges quickly and is not algorithm dependent

Improve predictive power of ML models that are based on distances, such as kNN (this goes in the direction of powering up interpretable models!)



darker = higher error

Define custom metrics to compare representations of variables in data and model spaces (**model validation**)



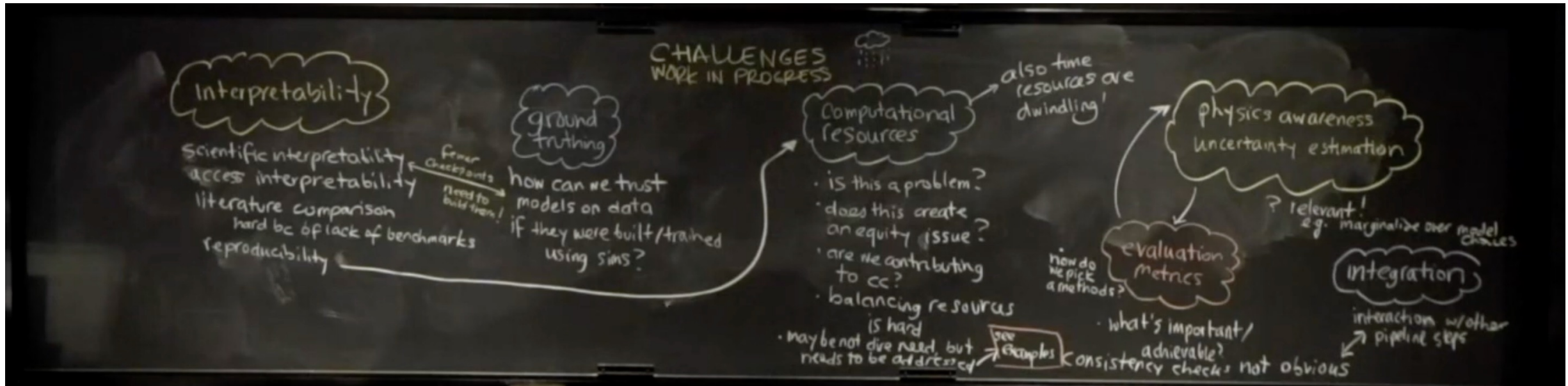
ETHZ	1	0.9	0.95	0.9	0.91
FESOM	0.9	1	0.95	0.89	0.9
NorESM	0.95	0.95	1	0.92	0.92
MRI	0.9	0.89	0.92	1	0.86
IPSL	0.91	0.9	0.92	0.86	1
	ETHZ	FESOM	NorESM	MRI	IPSL

Climate AI: ethics and explainability



LEAP

Challenges posed by AI

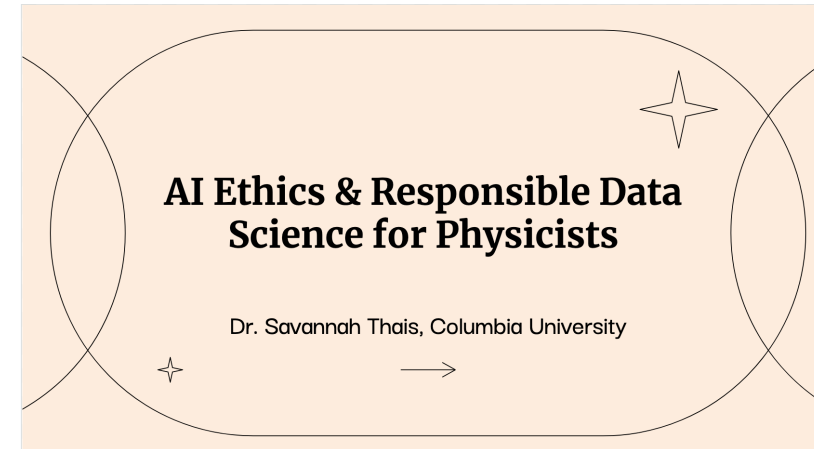


[Learn the Universe, Discussion: ML for Science, Promises and Problems](#) (August 2021)

- Only students/researchers at some institutions have access to ML tools
- Only a part of the academic community is following the AI discourse; lack of benchmarks makes it difficult to read literature
- Only **a part of society has a basic understanding of AI language and principles**

A useful framework: AI Functionality

What my graduate students (May 2023)
wanted to hear about the most:



The fallacy of AI functionality (Raji et al 2022)

Think of AI as an **infrastructure** or **system**

Thinking about reporting or regulating **performance**
is less overwhelming, more actionable

Table 1. Failure Taxonomy

Impossible Tasks	Conceptually Impossible Practically Impossible
Engineering Failures	Design Failures Implementation Failures Missing Safety Features
Post-Deployment Failures	Robustness Issues Failure under Adversarial Attacks Unanticipated Interactions
Communication Failures	Falsified or Overstated Capabilities Misrepresented Capabilities

Ethics and Explainability in Climate AI

Panel Discussion @ Climate Informatics 2024

VA et al 2024, PLOS Climate,
Ethics in climate AI: From theory to practice



1. Science is not apolitical; choice of priors, data, metrics, all carry biases; goal is not to resolve them but to be transparent and open
2. Functionality (failure modes) can be a lens to analyze AI systems (e.g. robustness); can protect us from unexpected failure modes
3. Data has many use cases; we hold no power on how they will be used but can say how we would like them to be used
4. Culture shift is needed to incentivize and reward **slower but well documented, robust, interpretable work**
5. Role of academia vs industry in “race” to climate models: **crucial for us to be deliberate/slow and ask the hard questions.**

New direction: LLMs for increased accessibility

- With codes becoming more complex to understand and use, making them available is not sufficient to ensure access
- The burden of accessibility falls onto early career researchers (grad students and postdocs), whose work cycle is not suited for slower work and not rewarded through those metrics
- Can we fine tune LLMs on pieces of code, github pull requests, etc to generate first drafts of documentation and tutorials?

The AI revolution and the role of scientists

climate
change

is here, whether we like it or not: better adapt than become extinct

Another way to think about it: ***Mitigation (agency)***

Editorial | Published: 20 March 2025

Using large language models wisely

[Nature Astronomy](#) 9, 315 (2025) | [Cite this article](#)

Some questions (for the under 40!)

What parts of our job do we want to keep?

What are the key skills and how do we practice them?

What is the role of creativity in science? How do we make room for diverse skills and ideas?

How do we train scientists for a job that may look very different in 5-10 years?