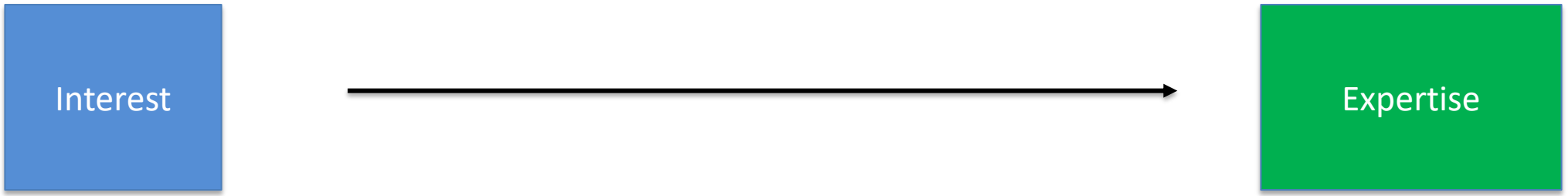
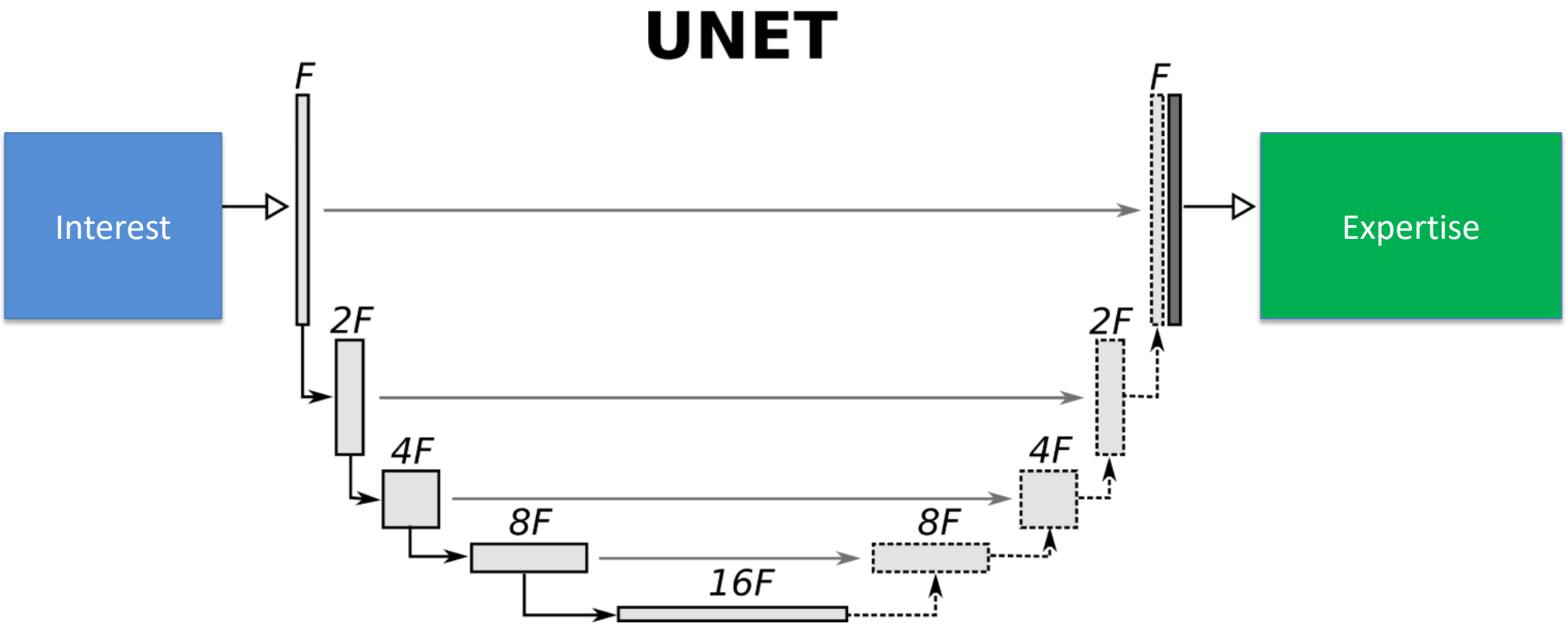
Data-Centric Research

12.S992 AI for Climate Action
Spring 2026
Speaker: Abigail Bodner

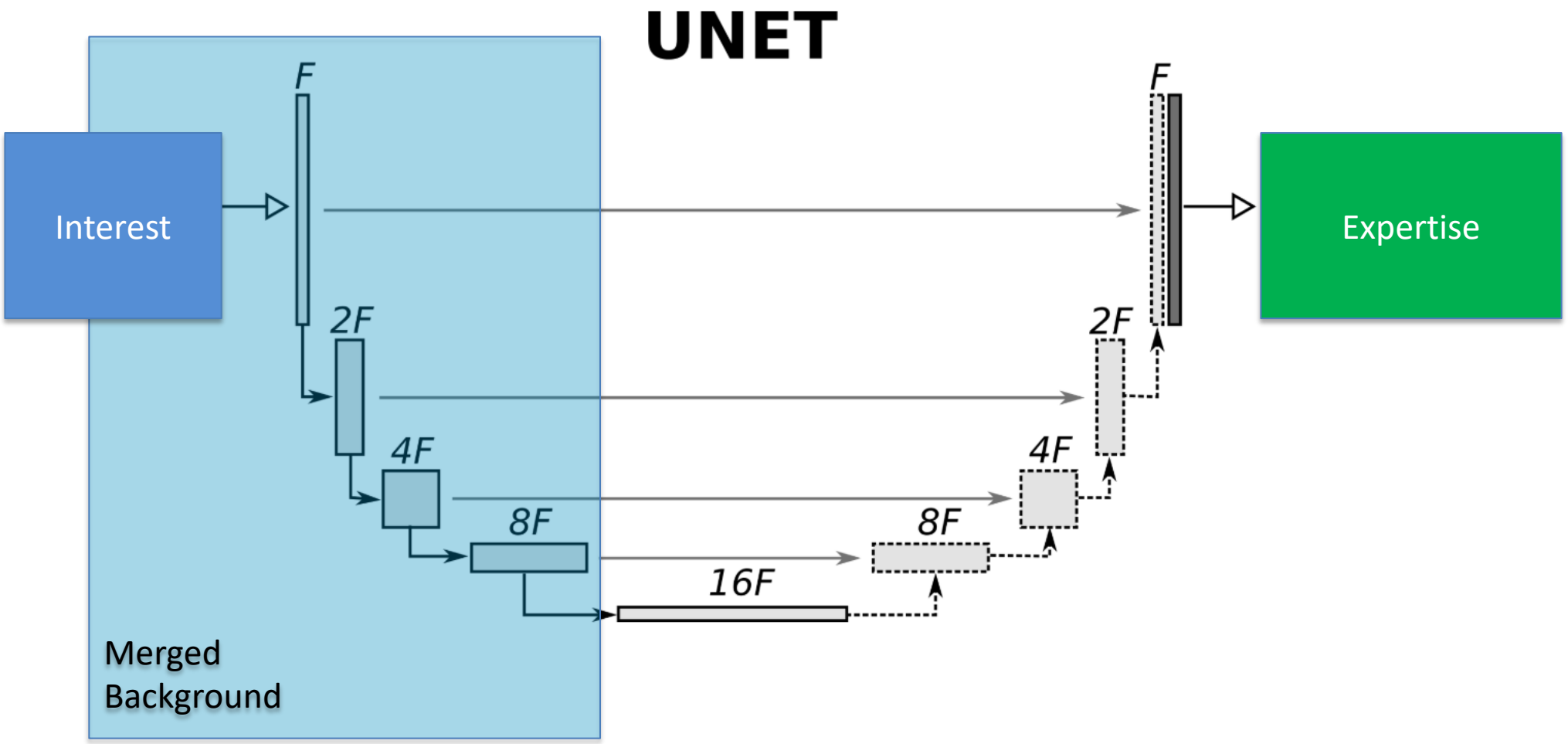
AI for Climate Action



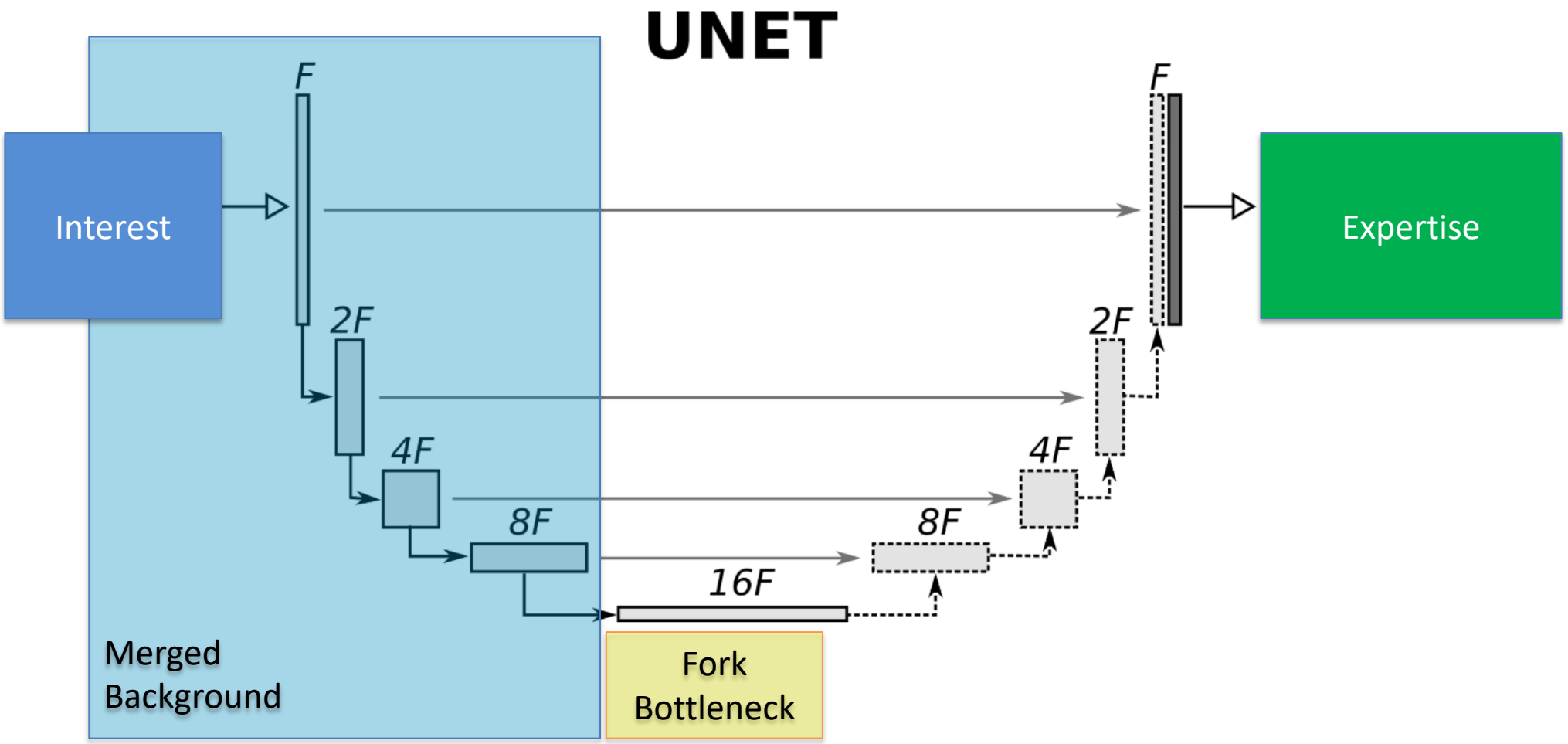
AI for Climate Action class architecture



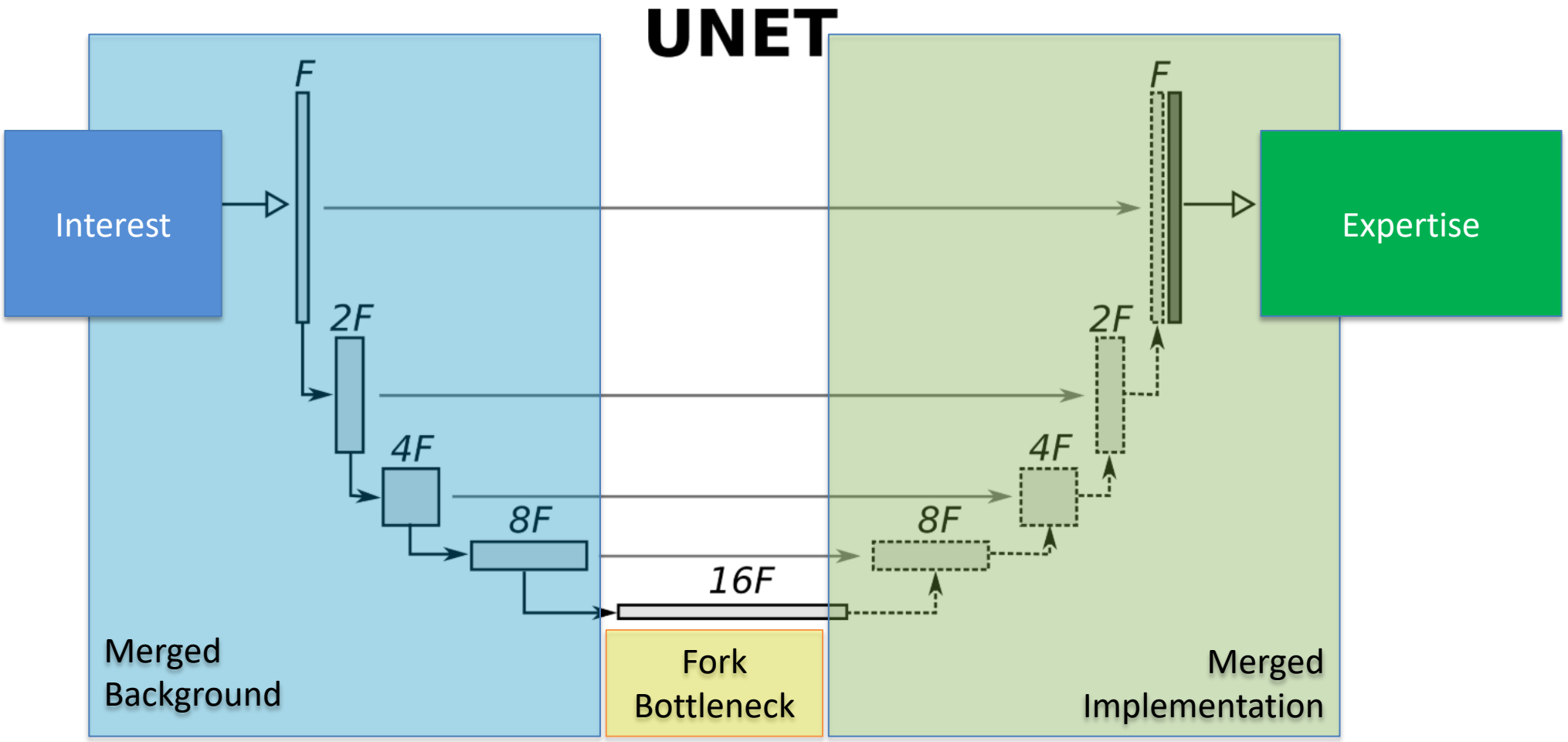
AI for Climate Action class architecture



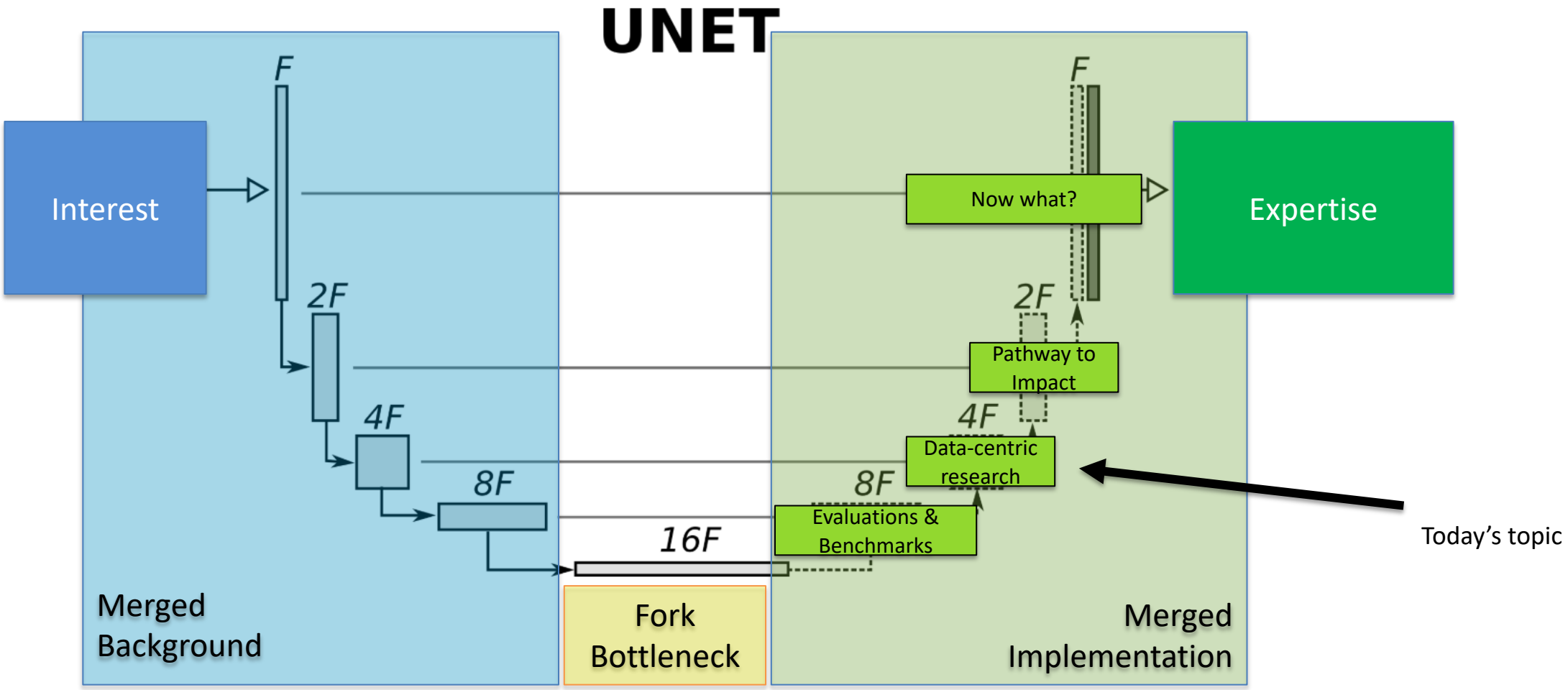
AI for Climate Action class architecture



AI for Climate Action class architecture



AI for Climate Action class architecture



Data-Centric Research

How does data (not just the model) drive performance?

Data-Centric Research

How does data (not just the model) drive performance?

- Selecting meaningful inputs and outputs
- Accounting for changing distributions
- Preprocessing choices
- Handling small amounts of data

Data-Centric Research

How does data (not just the model) drive performance?

- Selecting meaningful inputs and outputs
- Accounting for changing distributions
- Preprocessing choices
- Handling small amounts of data

→ Domain knowledge is required!

Inputs and outputs

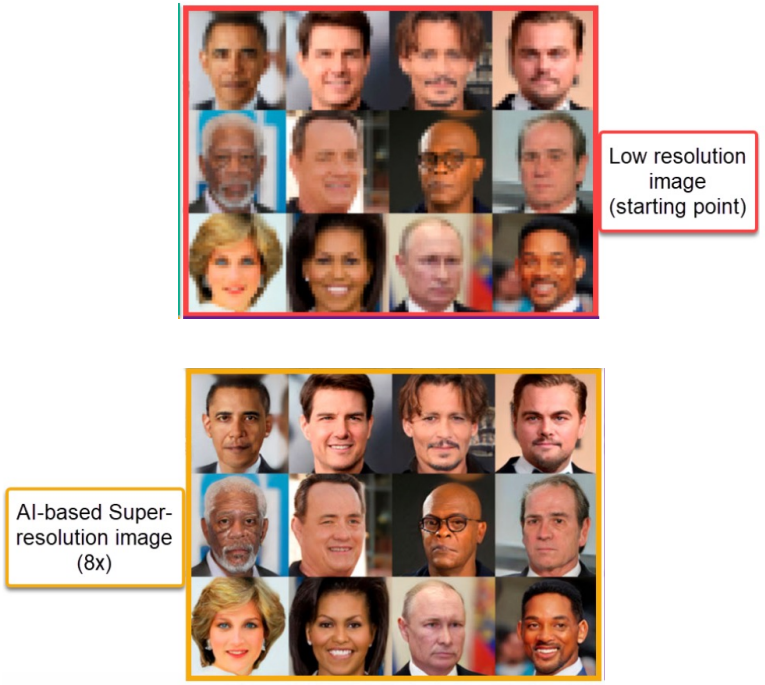
Inputs and outputs

- Meaningful correlations
- Application-driven variables
- Clear targets (corresponding to benchmarks and evaluation metrics)

Inputs and outputs

- Meaningful correlations
- Application-driven variables
- Clear targets (corresponding to benchmarks and evaluation metrics)

Example ✅: upsampling of climate variables

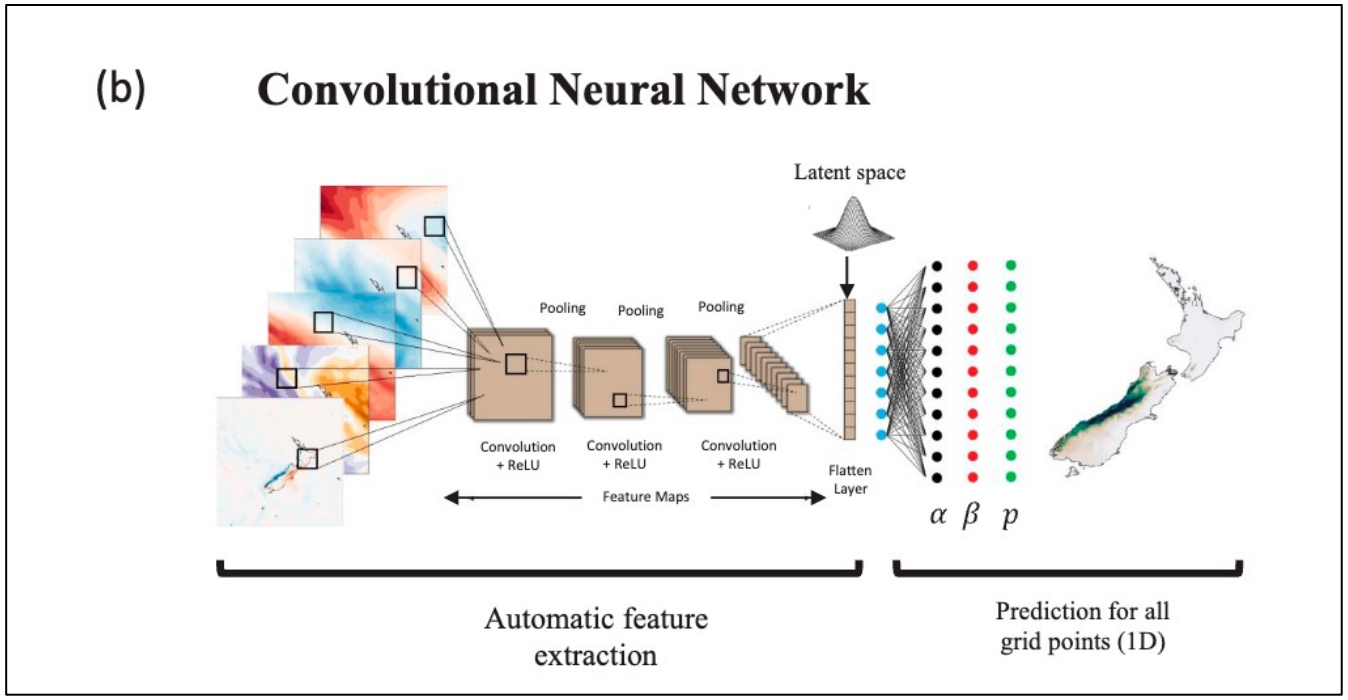


Rampal et al (2024)

Inputs and outputs

- Meaningful correlations
- Application-driven variables
- Clear targets (corresponding to benchmarks and evaluation metrics)

Example ✅: upsampling of climate variables

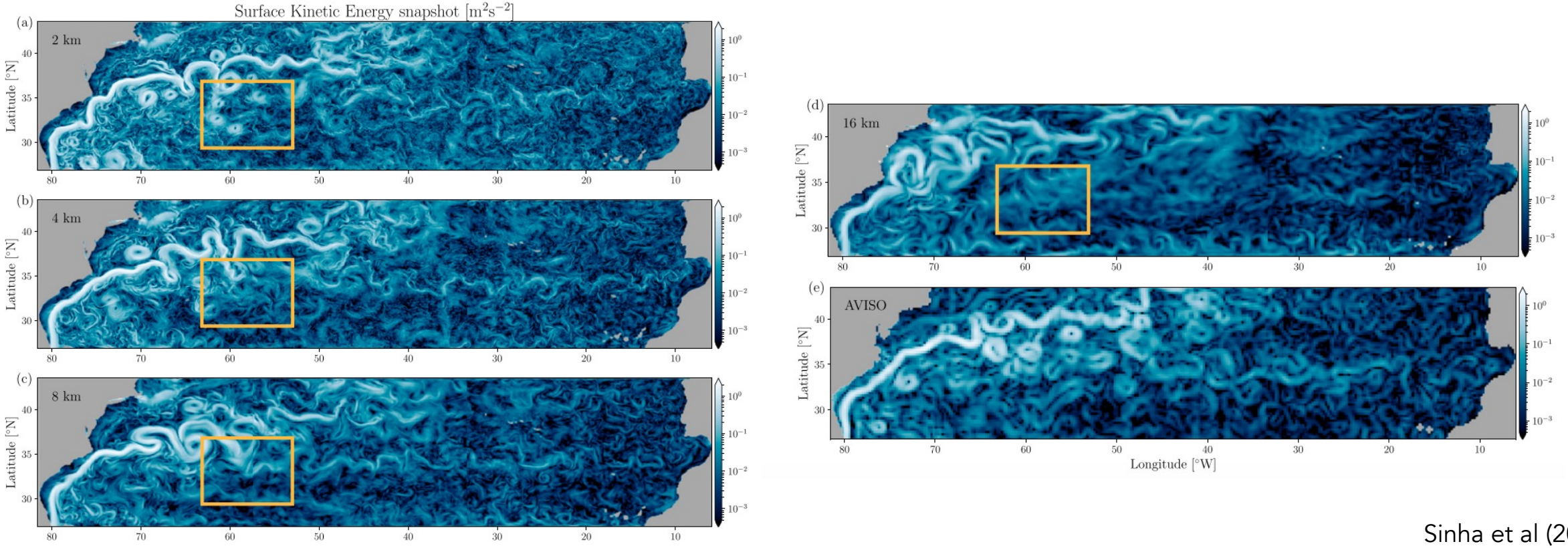


Rampal et al (2024)

Inputs and outputs

- Meaningful correlations
- Application-driven variables
- Clear targets (corresponding to benchmarks and evaluation metrics)

Example ❌: mismatch between simulations & observations



Sinha et al (2023)

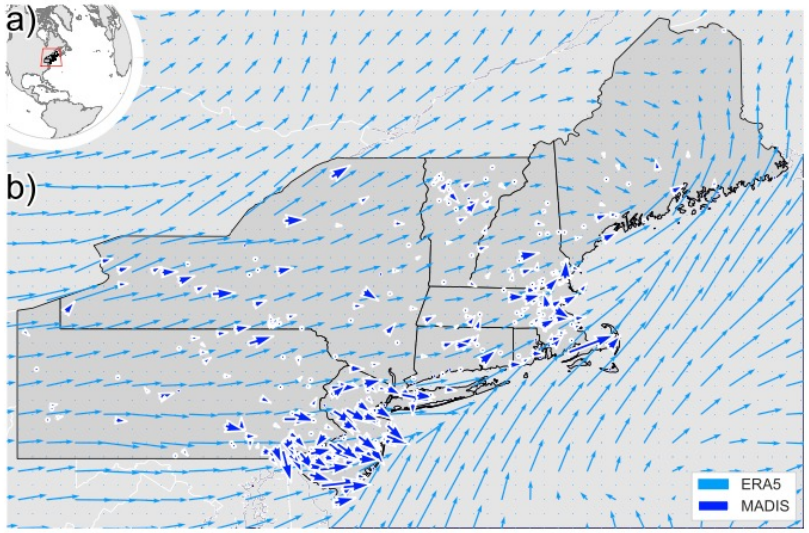
Spatial and temporal distributions

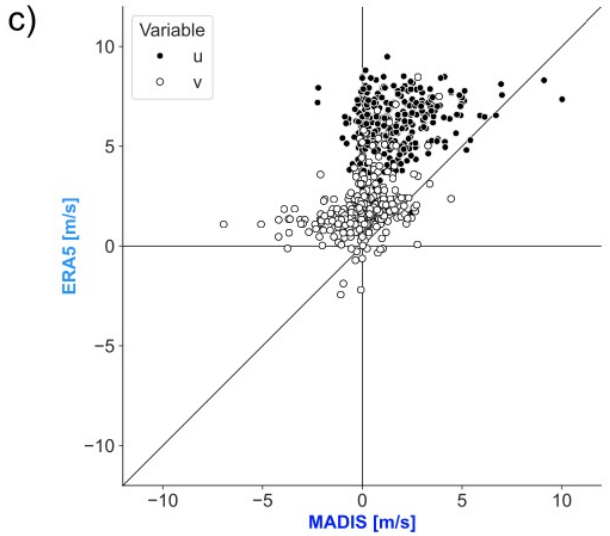
Spatial and temporal distributions

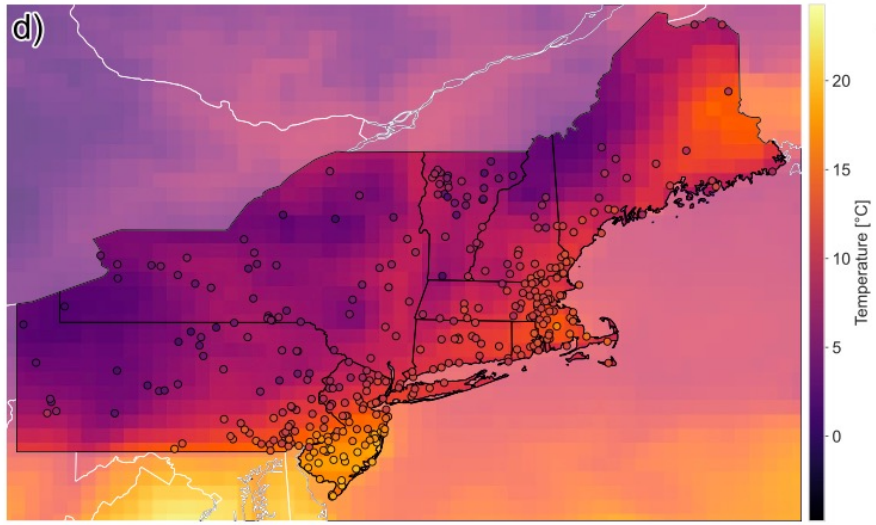
Sparse vs dense

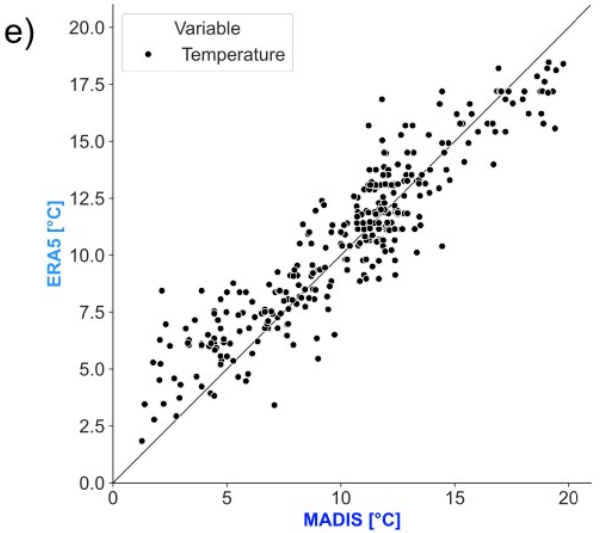
Handling missing data (in space, time, hidden e.g., by clouds)

Co-location of time and space









Yang et al (2025)

Spatial and temporal distributions

Sparse vs dense

Handling missing data (in space, time, hidden e.g., by clouds)

Co-location of time and space

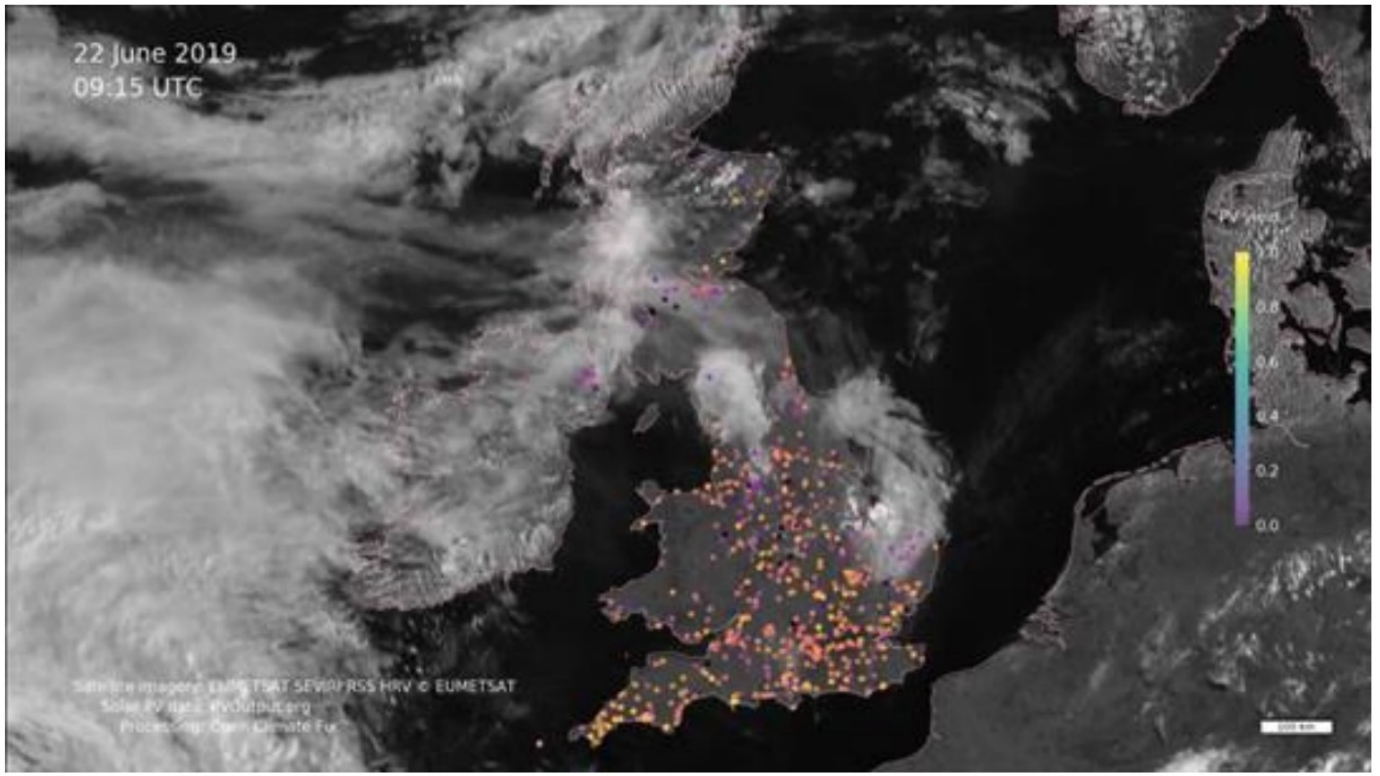
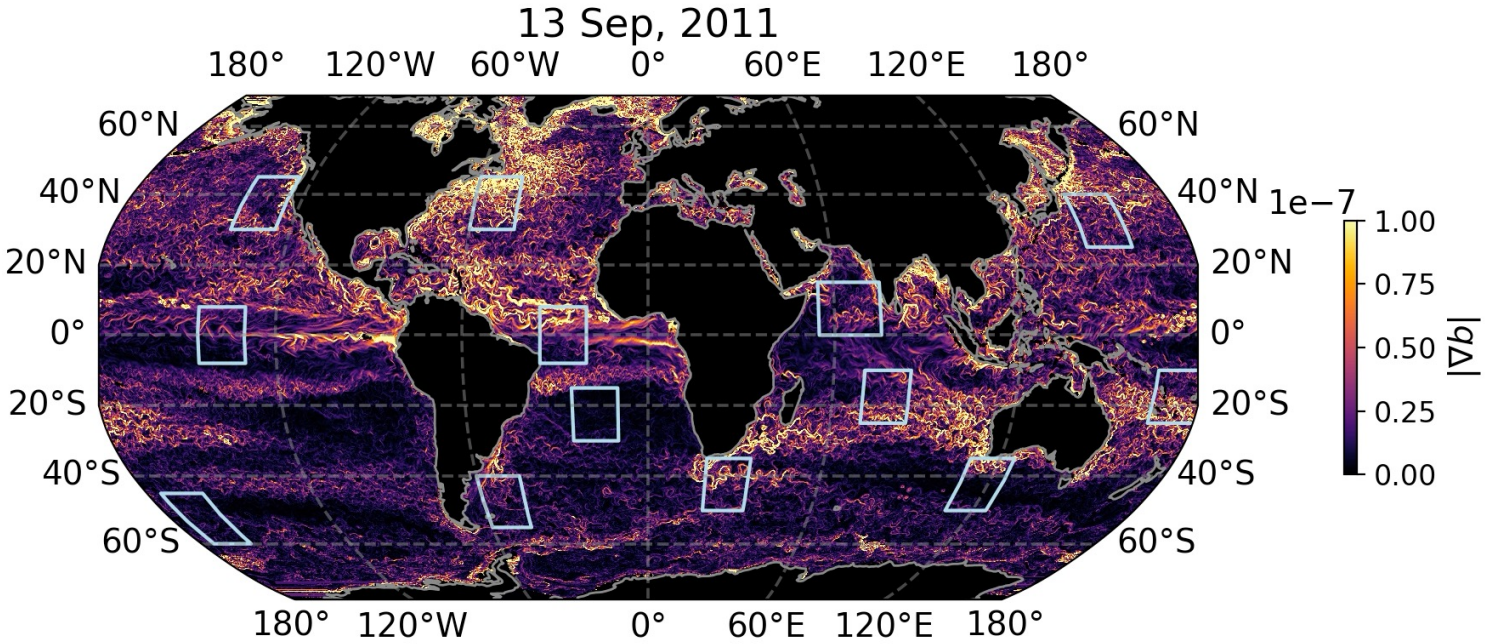


Image source: Open Climate Fix

Spatial and temporal distributions

Sampling bias (e.g., hemispheric, population sizes)



Bodner, Balwada, Zanna (2025)

Spatial and temporal distributions

Sampling bias (e.g., hemispheric, population sizes)

The power of local action, ctd.

ICLEI
Local Governments for Sustainability

2500+
Cities towns & regions

25+
% Global urban population

125+
Countries worldwide

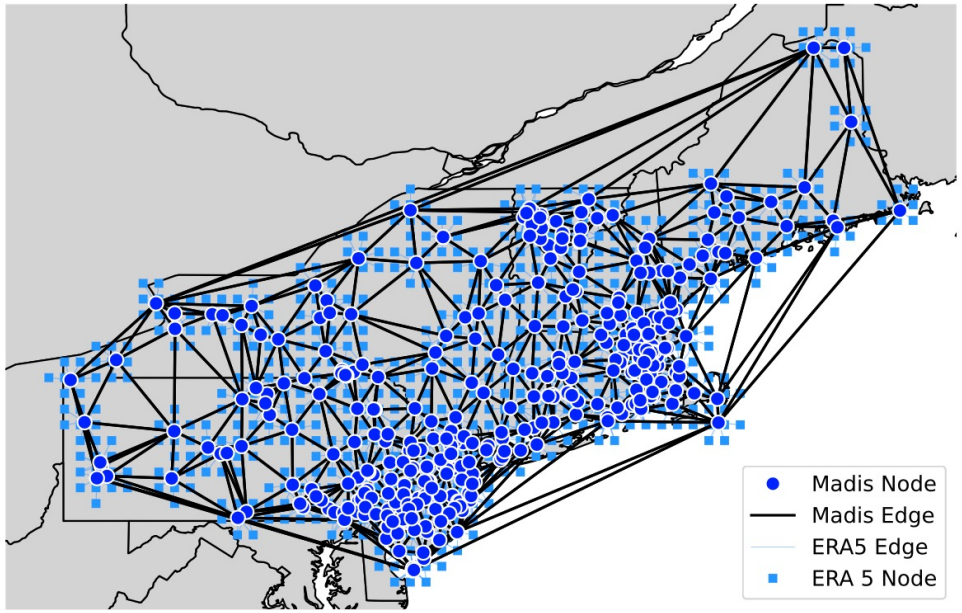
20+
% Global population

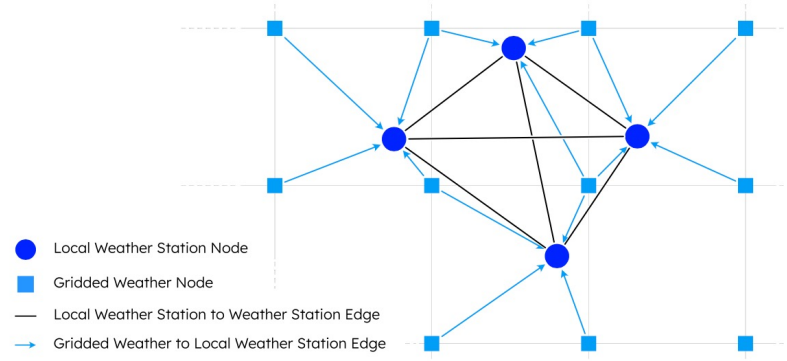
Source: https://iclei.org/our_network/

Spatial and temporal distributions

- How we handle the spatio-temporal variability of the data will guide model choices down the line.

- Examples:
  - Spatial: FNO or GNN
  - Temporal: RNN, LSTM, transformers

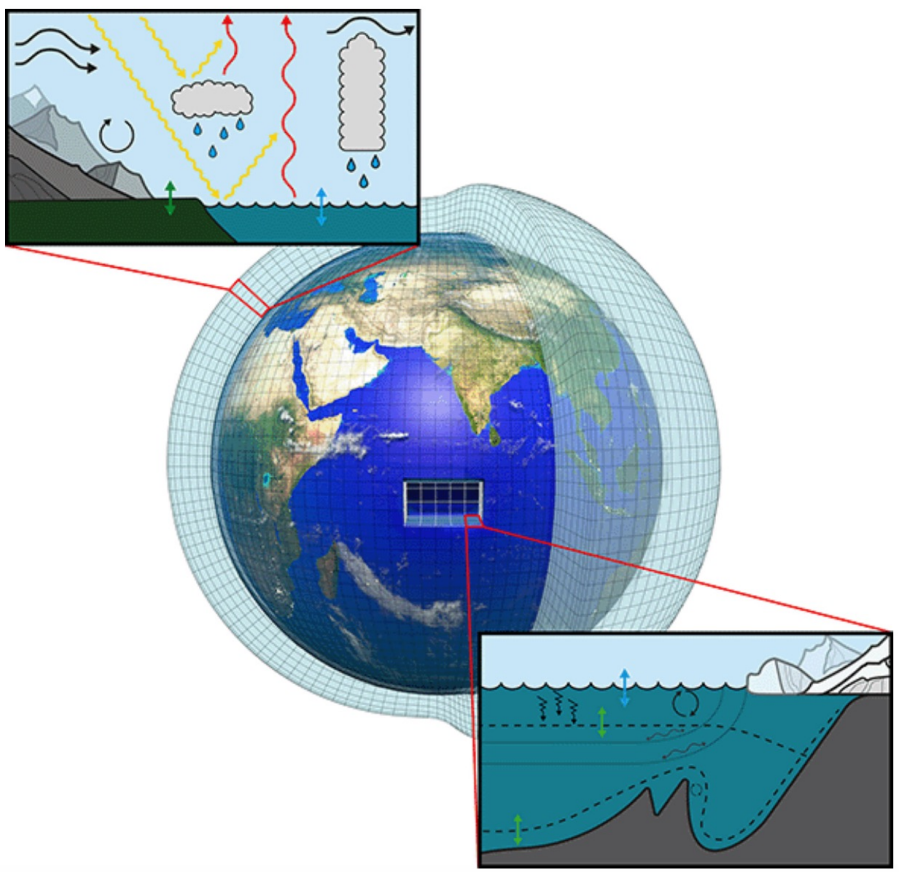


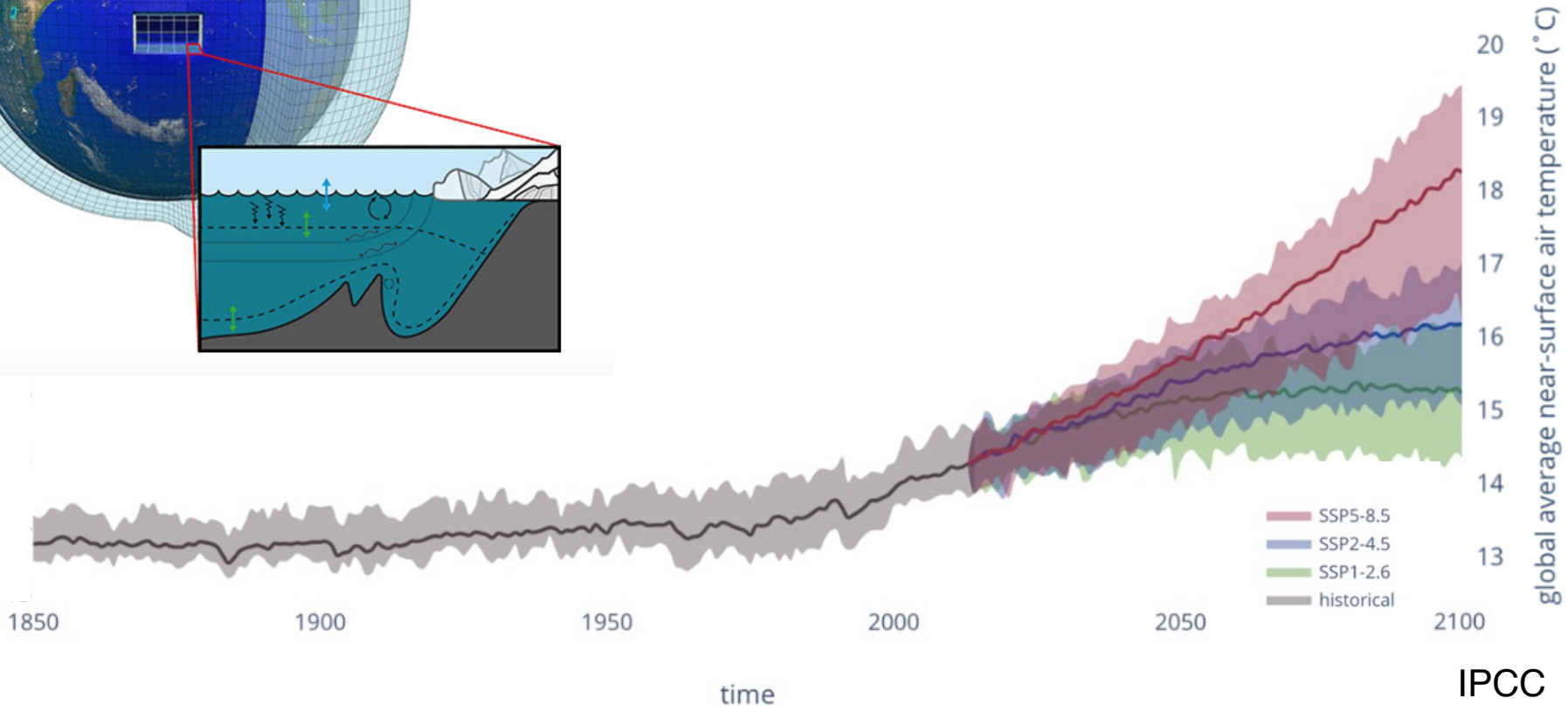


Yang et al (2025)

Data preprocessing

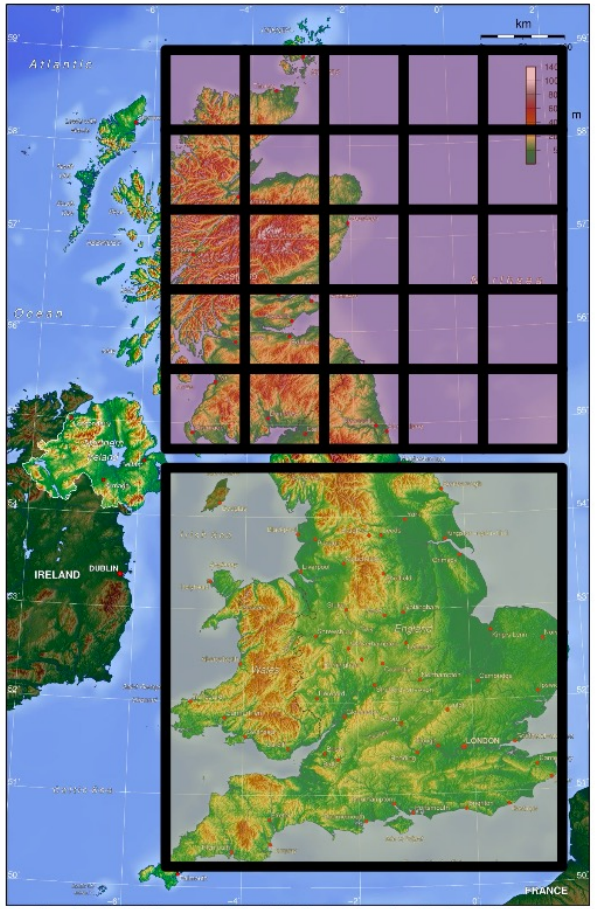
Global Climate Change





IPCC

What's smaller than 100km?



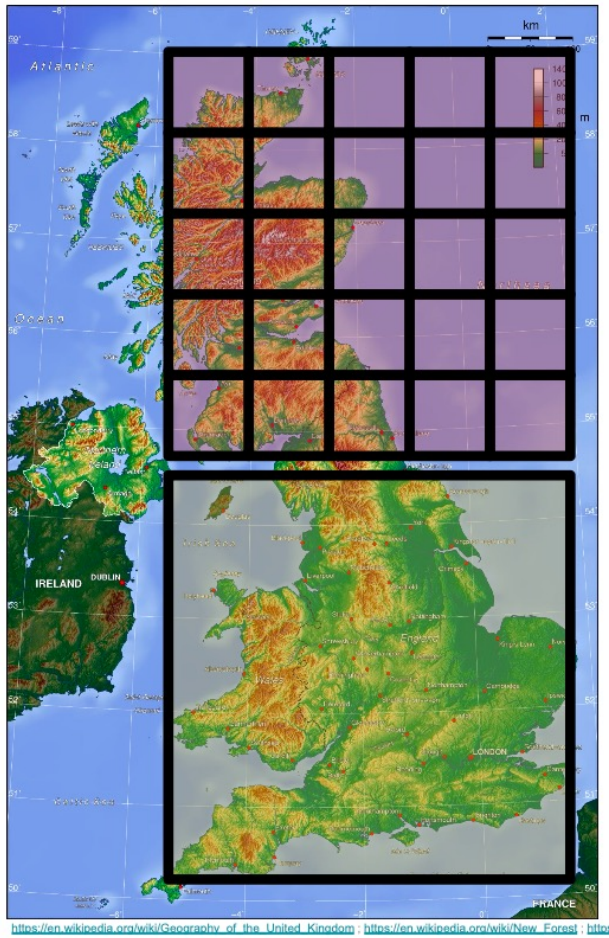
New Models
(~100km)

Old Models
(~500km)

Clouds (atmosphere) & surface conditions (land)

https://en.wikipedia.org/wiki/Geography_of_the_United_Kingdom ; https://en.wikipedia.org/wiki/New_Forest ; https://en.wikipedia.org/wiki/Iceberg ; https://en.wikipedia.org/wiki/Phytoplankton

What's smaller than 100km?



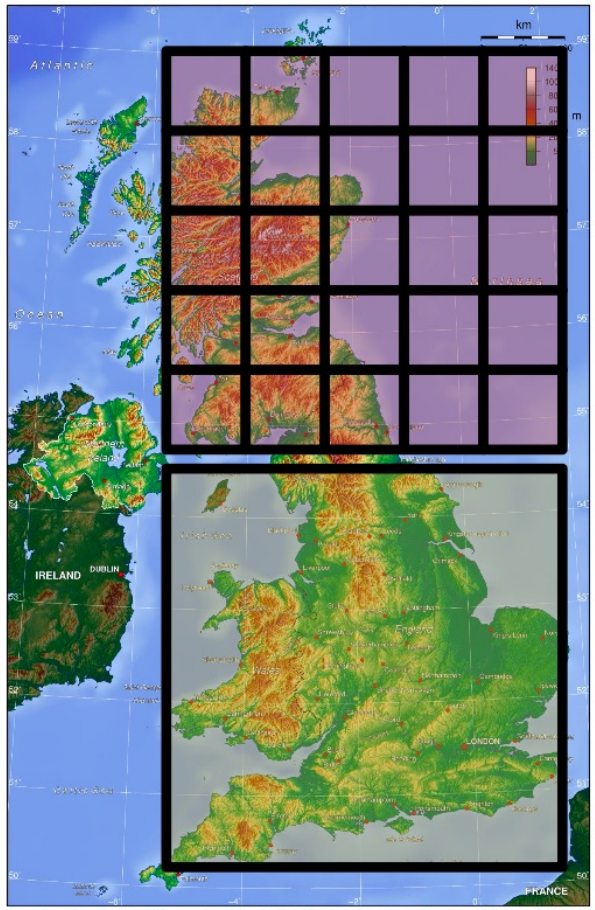
New Models
(~100km)

Old Models
(~500km)

Icebergs (cryosphere)

https://en.wikipedia.org/wiki/Geography_of_the_United_Kingdom : https://en.wikipedia.org/wiki/New_Forest : https://en.wikipedia.org/wiki/Iceberg : https://en.wikipedia.org/wiki/Phytoplankton

What's smaller than 100km?



New Models (~100km)

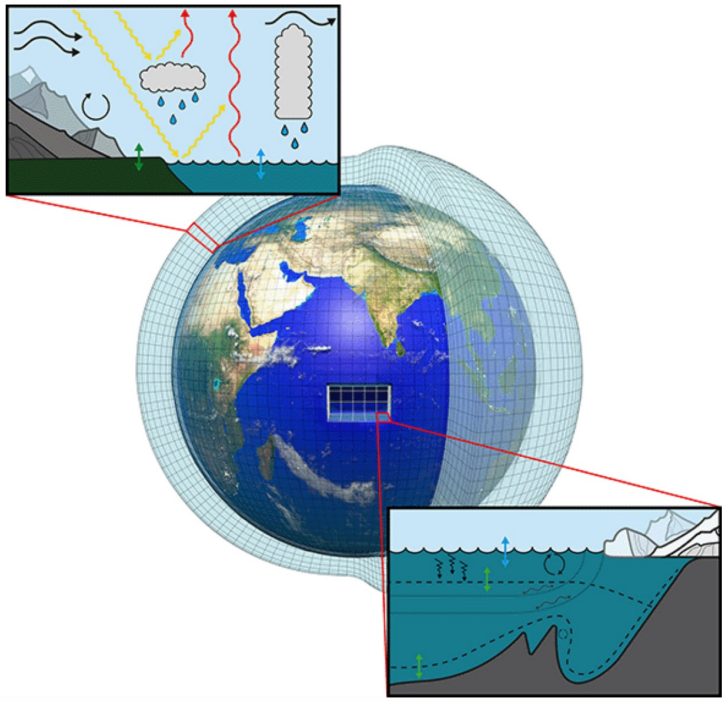
Old Models (~500km)

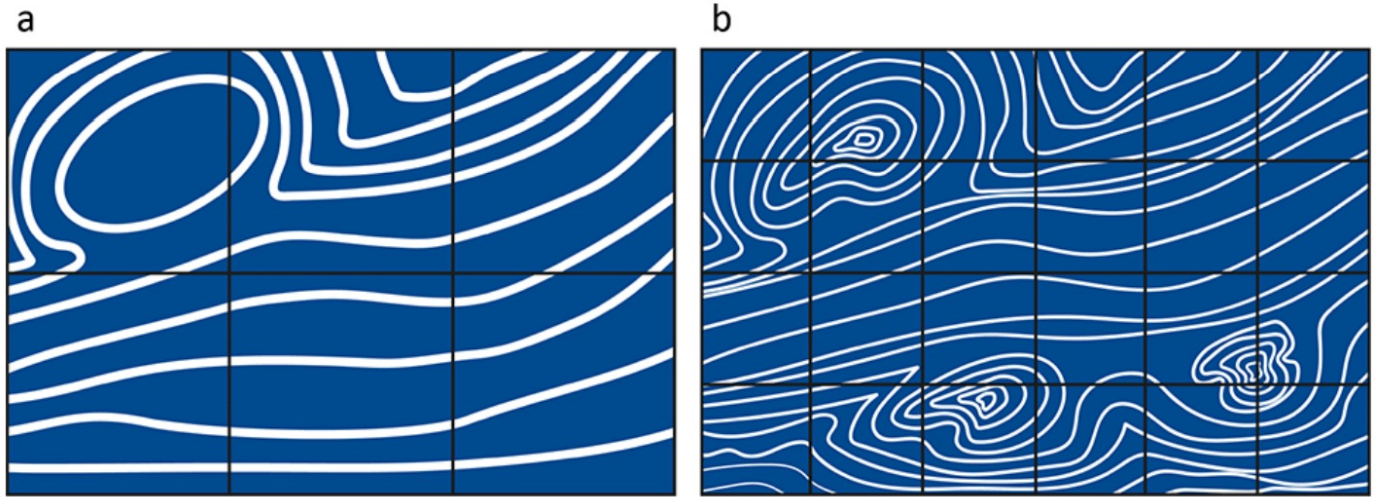
Eddies (ocean) & phytoplankton (biology)

https://en.wikipedia.org/wiki/Geography_of_the_United_Kingdom : https://en.wikipedia.org/wiki/New_Forest : https://en.wikipedia.org/wiki/Iceberg : https://en.wikipedia.org/wiki/Phytoplankton

Data preprocessing

Filtering/smoothing: removing noise or important signal?

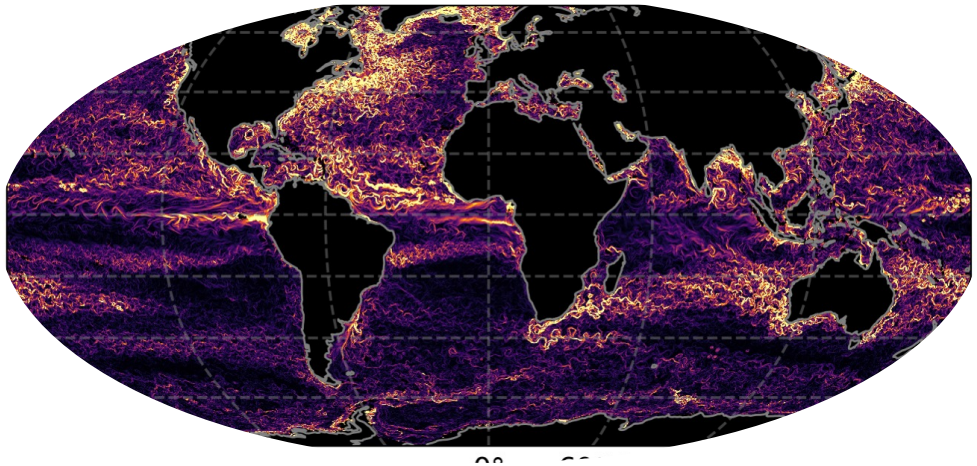


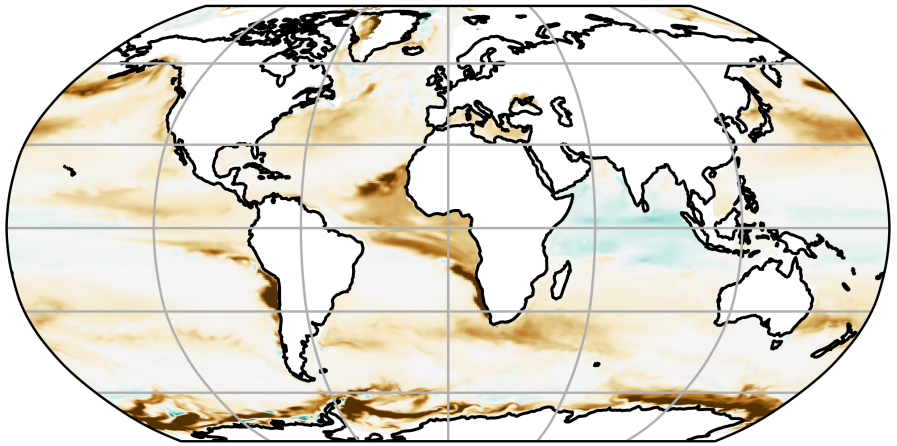


Christensen & Zanna (2022)

Spatial and temporal distributions

Distribution remapping between training and inference

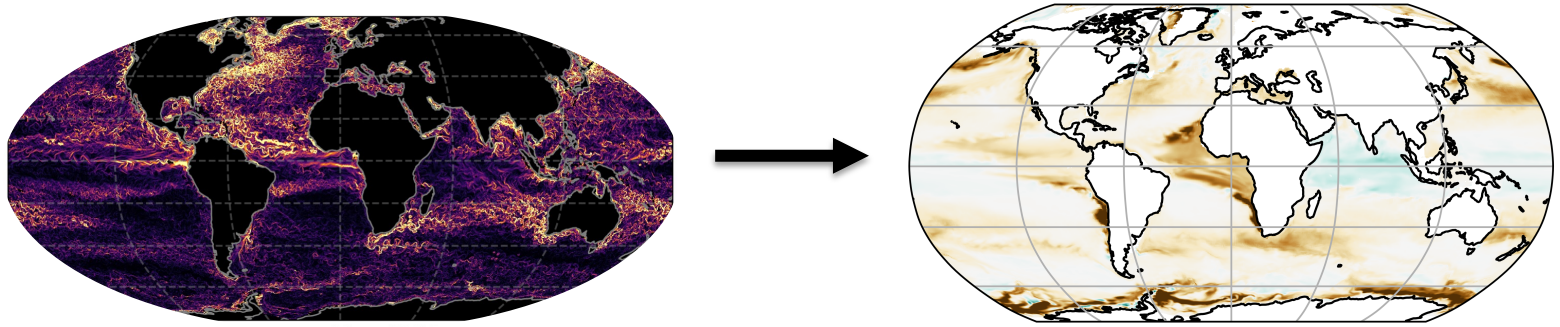




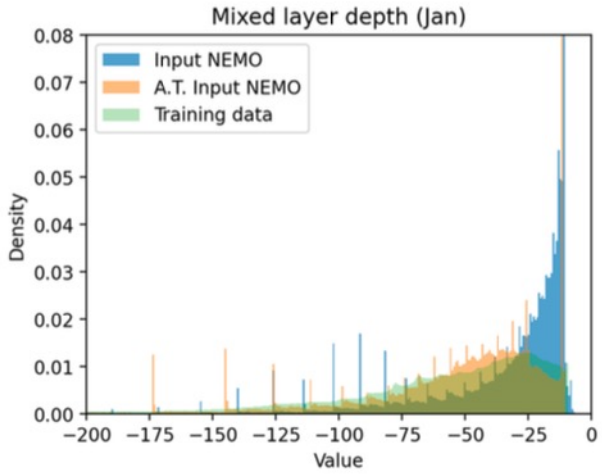
Contreras et al. (in prep.)

Spatial and temporal distributions

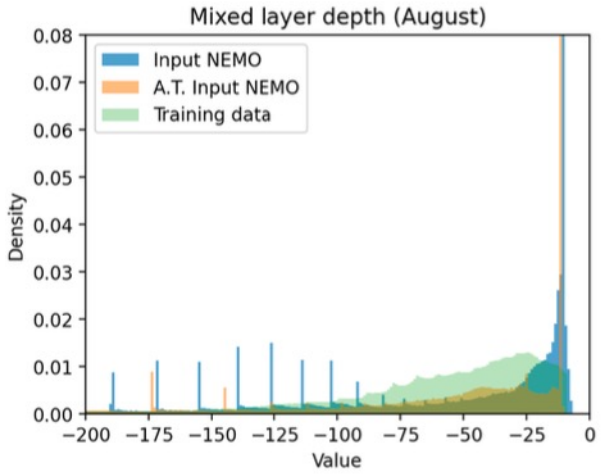
Distribution remapping between training and inference



Winter



Summer



Contreras et al. (in prep.)

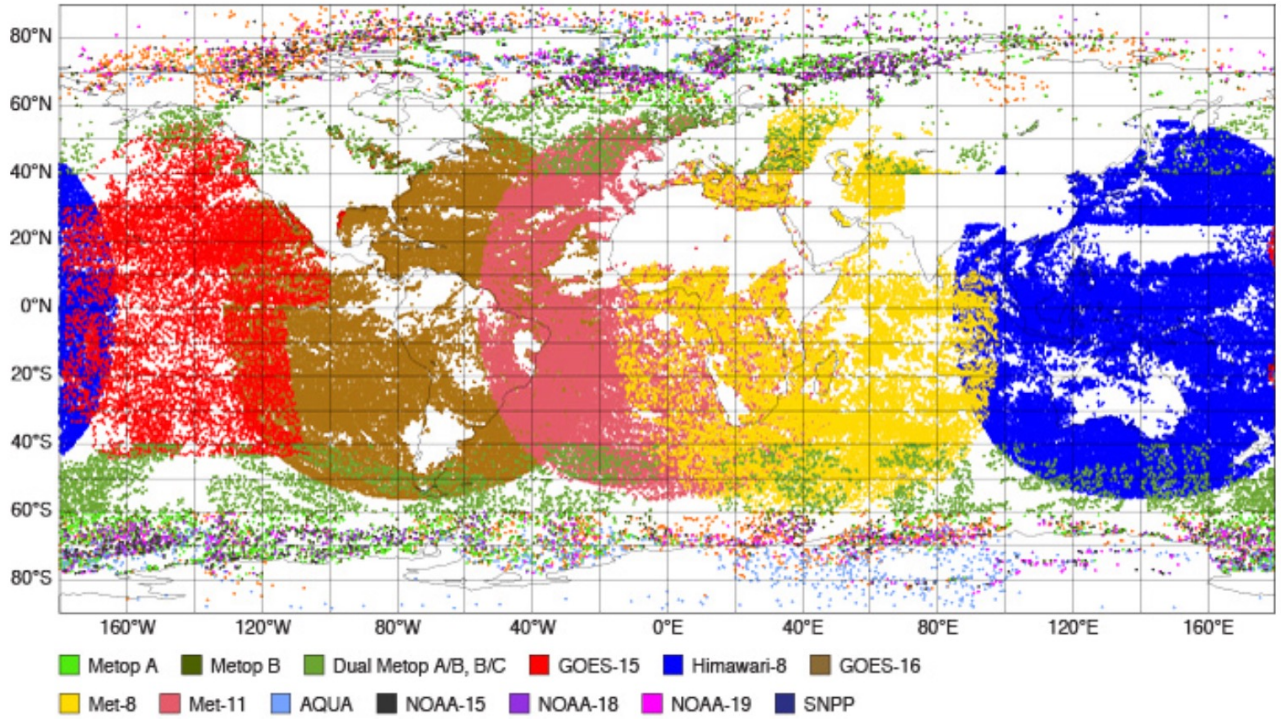
Data preprocessing

Multi-source or multi-process data integration

Data preprocessing

Multi-source or multi-process data integration

Weather observations come from many sources, but they cannot provide a complete picture of the state of the Earth system at a given point in time. (Diagram: WMO)

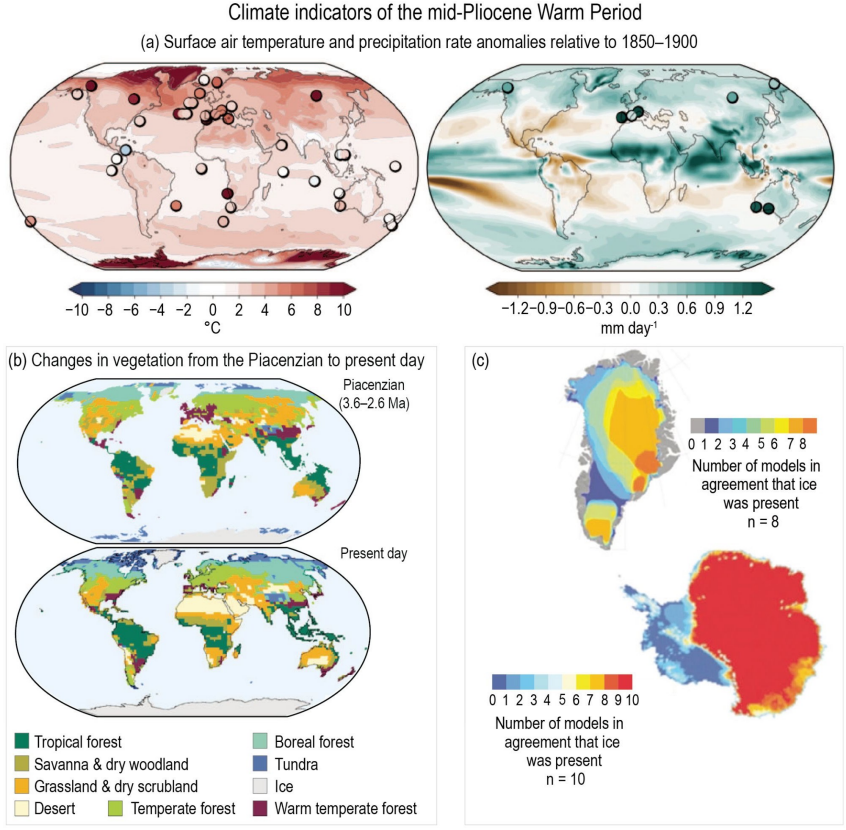


Typical coverage of active atmospheric motion vector (AMV) data for a 12-hour assimilation cycle (00 UTC, 7 March 2019).

Data preprocessing

Multi-source or multi-process data integration

Paleoclimate Models



Proxies and Archives

"Travertine speleothem (Crystal Cave, Main Island, Bermuda) 1" by James St. John is licensed under CC BY 2.0

"Tree rings" by Out of the Fire Blog is licensed under CC BY 2.0

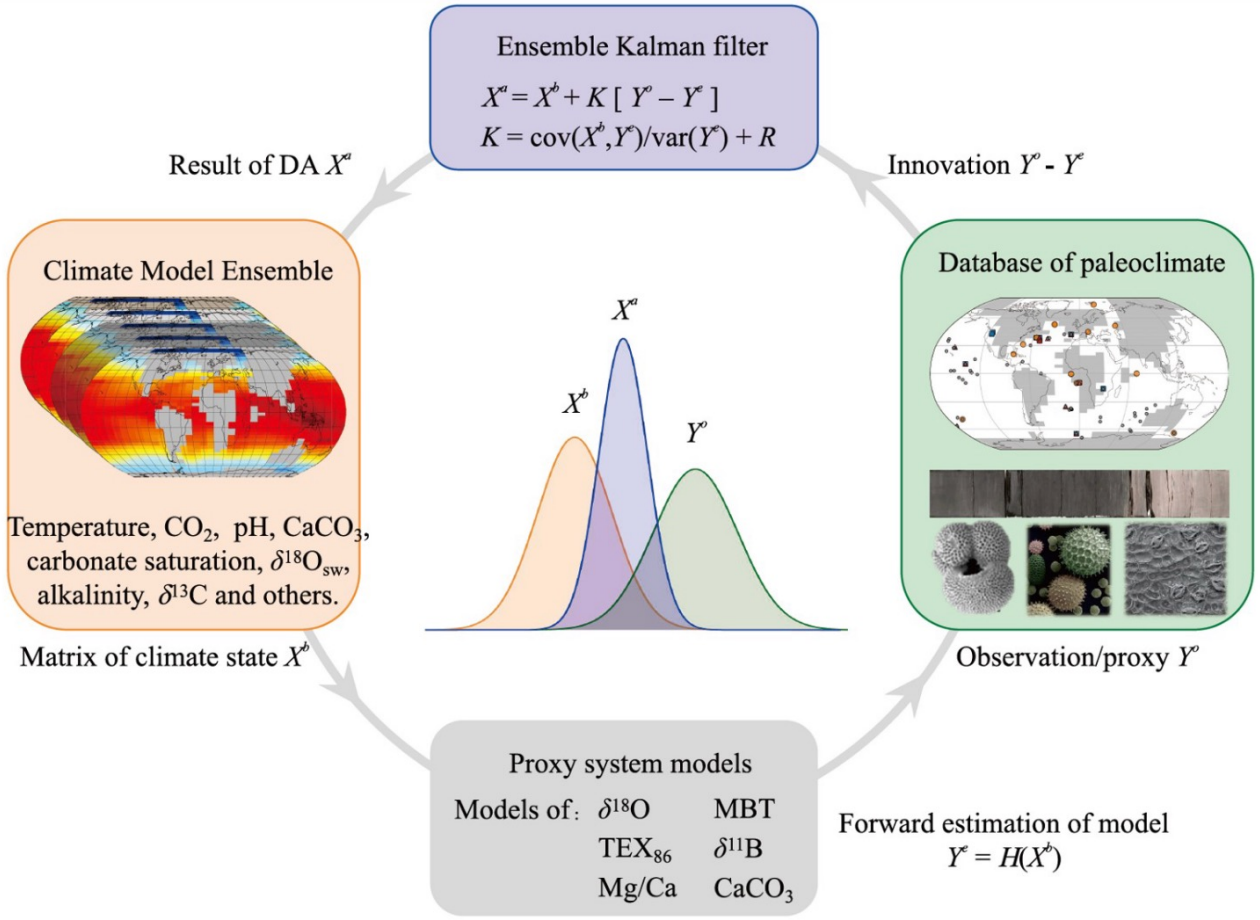
"A volcanic ash layer in the WAIS Divide ice core. Volcanic markers like these were used in the new study to synchronize ice cores from across Antarctica." by Oregon State University is licensed under CC BY-SA 2.0

"2016_Lake_sediment_core_Forlorn_Lakes_Gifford_Pinchot_National_Forest_Washington." by USDA Forest Service is marked with Public Domain Mark 1.0

Diploria fossil brain coral on Devil's Point Hardground (Cockburn Town Member, Grotto Beach Formation, Upper Pleistocene, ~120-123 ka; Cockburn Town Fossil Reef, San Salvador Island, Bahamas) 3" by James St. John is licensed under CC BY 2.0

Data preprocessing

Multi-source or multi-process data integration



Zhang et al (2025)

Data preprocessing

Multi-source or multi-process data integration

Warning! Sometimes skill drops when adding more data sources. Why and how to interpret this can only be done with domain knowledge

BRITISH ECOLOGICAL SOCIETY | Journal of Animal Ecology

RESEARCH METHODS GUIDE | Open Access

Should you use data integration for your distribution model?

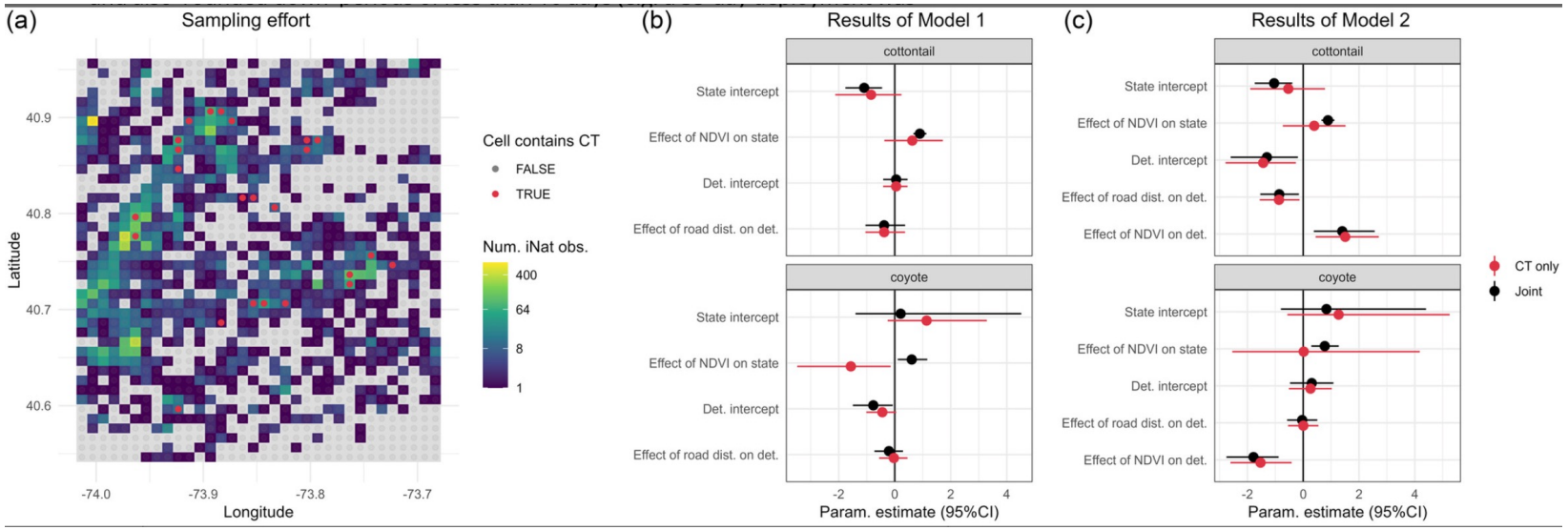
Benjamin R. Goldstein, Jeffrey W. Doser, Brent S. Pease, Krishna Pacifici

First published: 28 January 2026 | https://doi.org/10.1111/1365-2656.70210 | VIEW METRICS

Data preprocessing

Multi-source or multi-process data integration

Warning! Sometimes skill drops when adding more data sources. Why and how to interpret this can only be done with domain knowledge

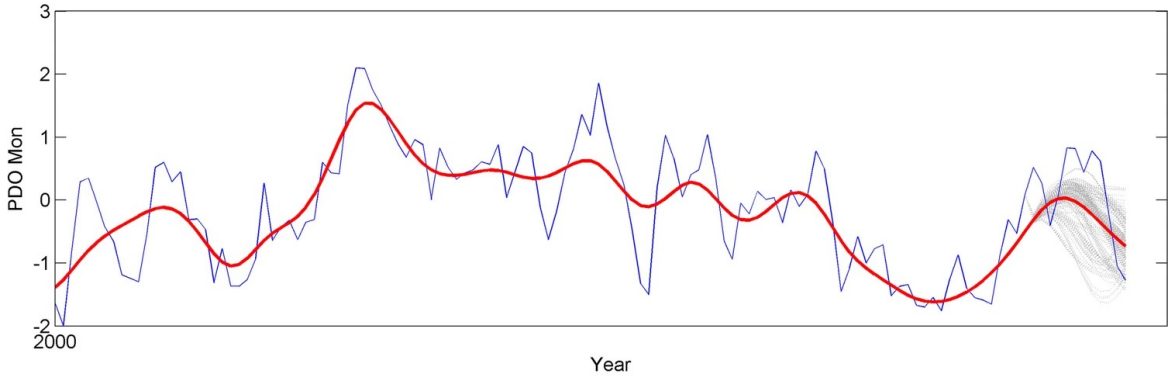


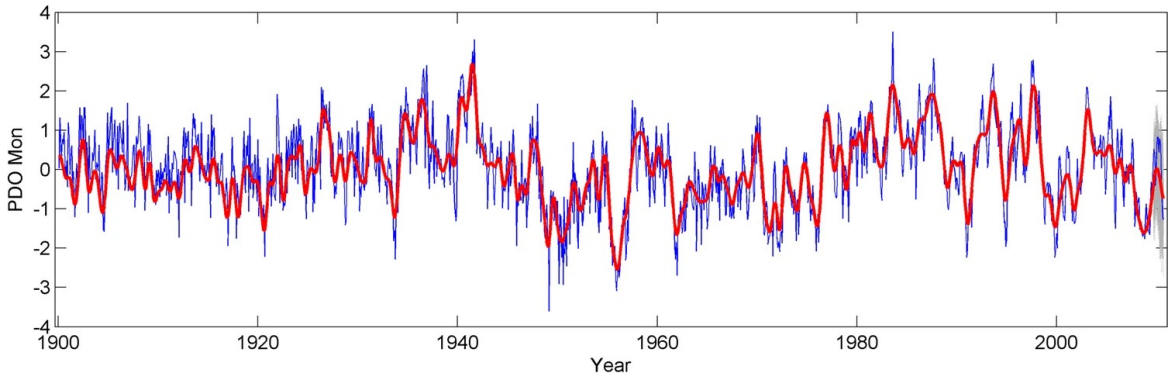
Goldstien et al (2026)

Data preprocessing

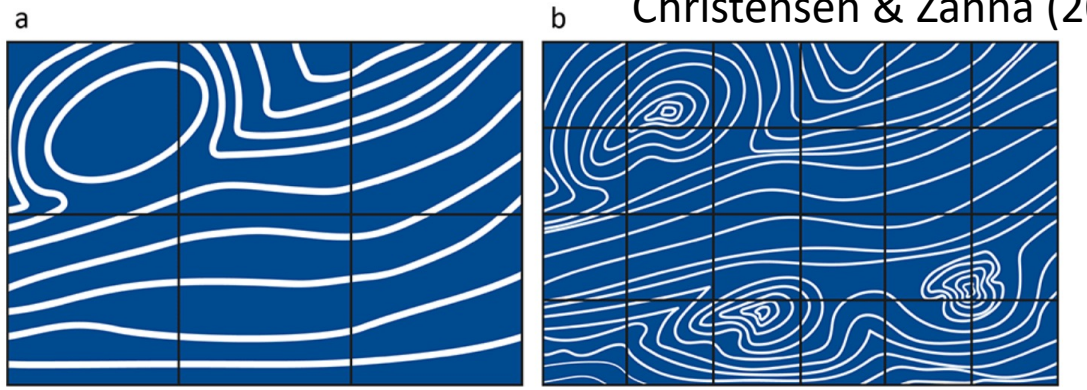
Matching scales/times (local→ global, hours → seasons)

- How to define extreme events and are they relevant?





Christensen & Zanna (2022)

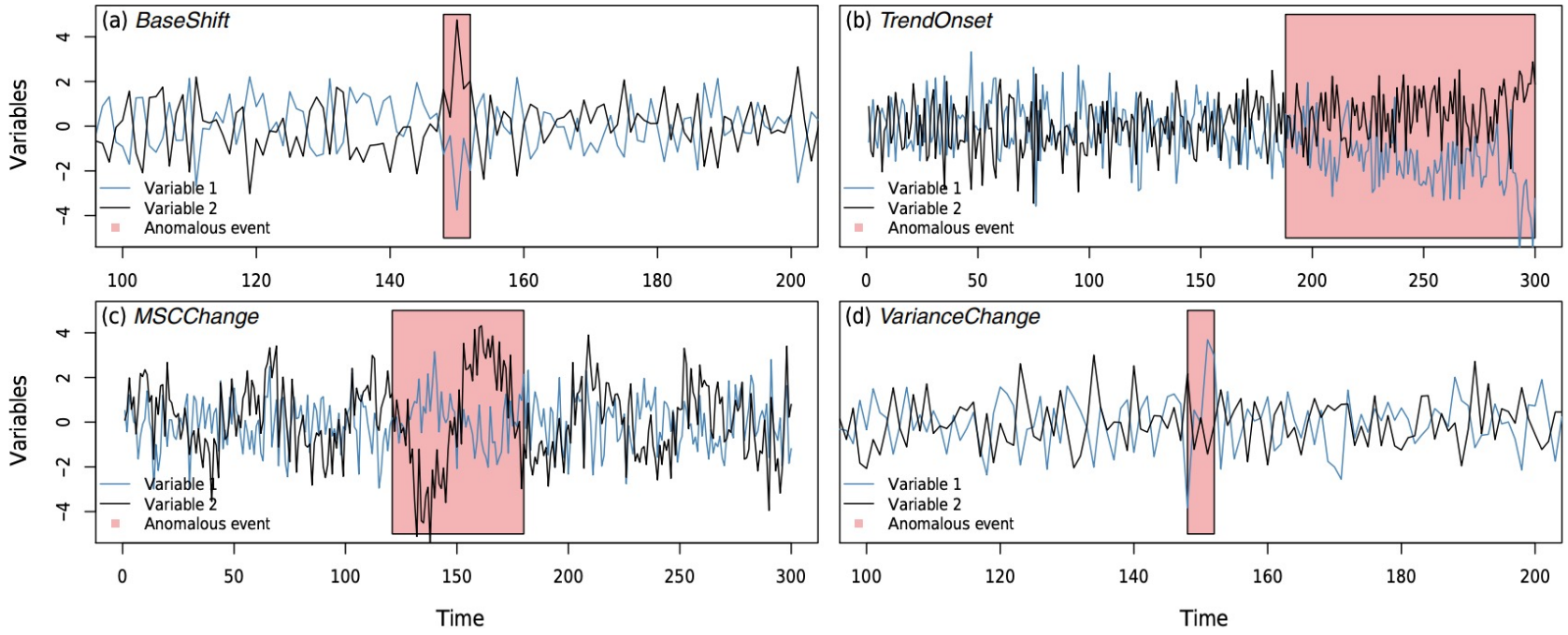


Lee et al (2017)

Data preprocessing

Matching scales/times (local→ global, hours → seasons)

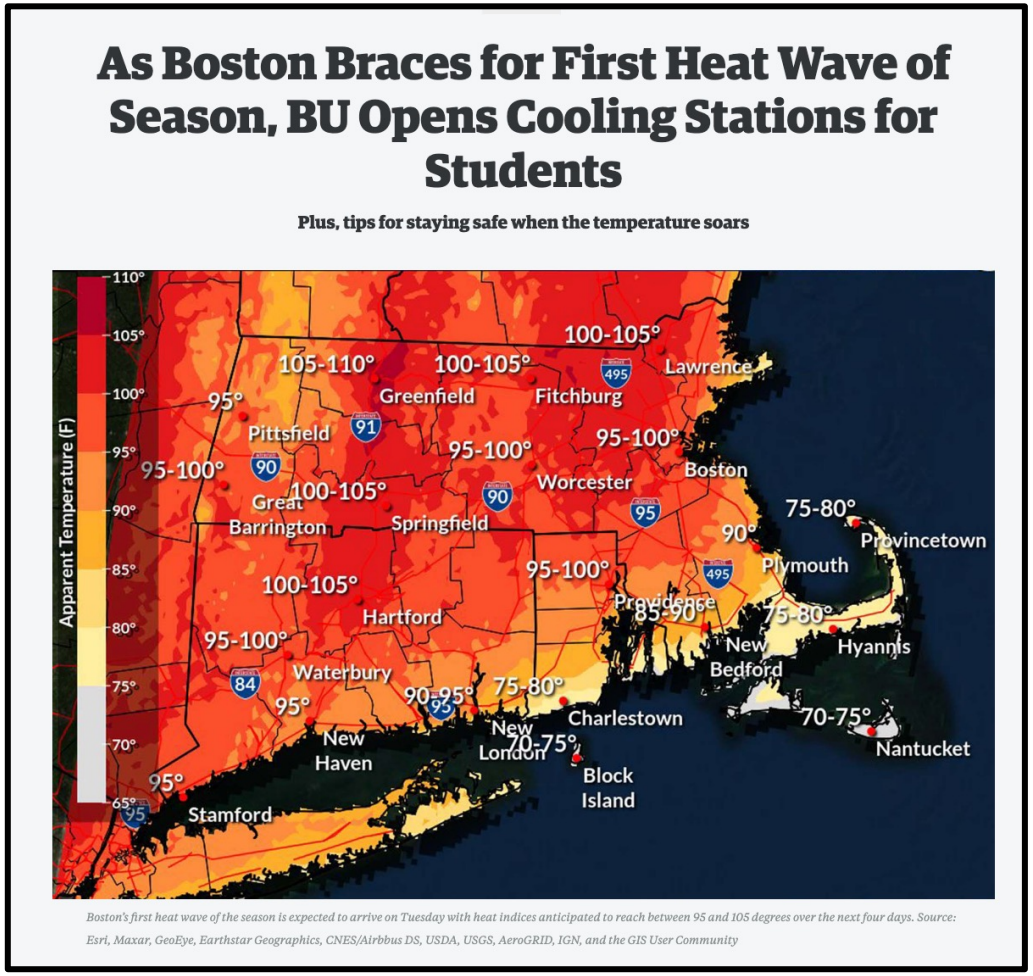
- How to define extreme events and are they relevant?

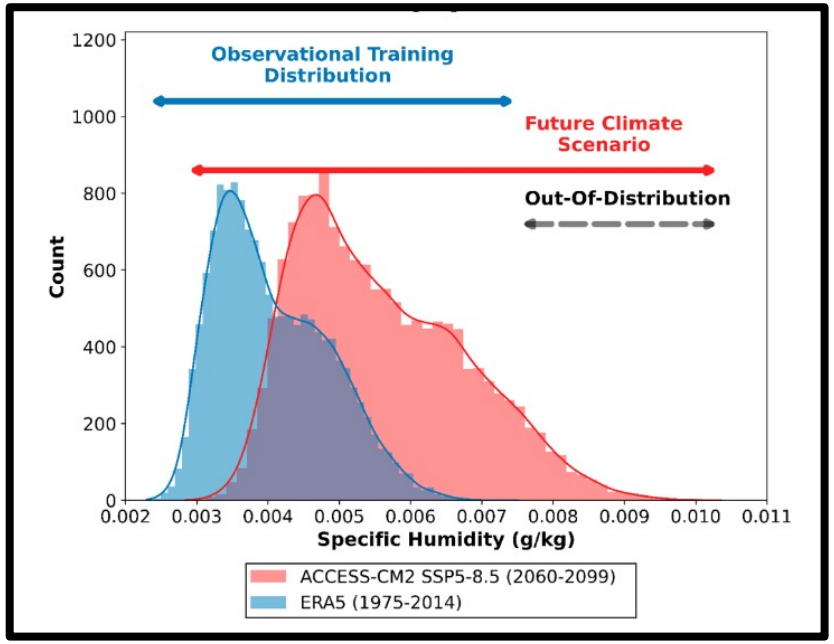


Flach et al (2017)

Data preprocessing

Extreme events, distribution tails

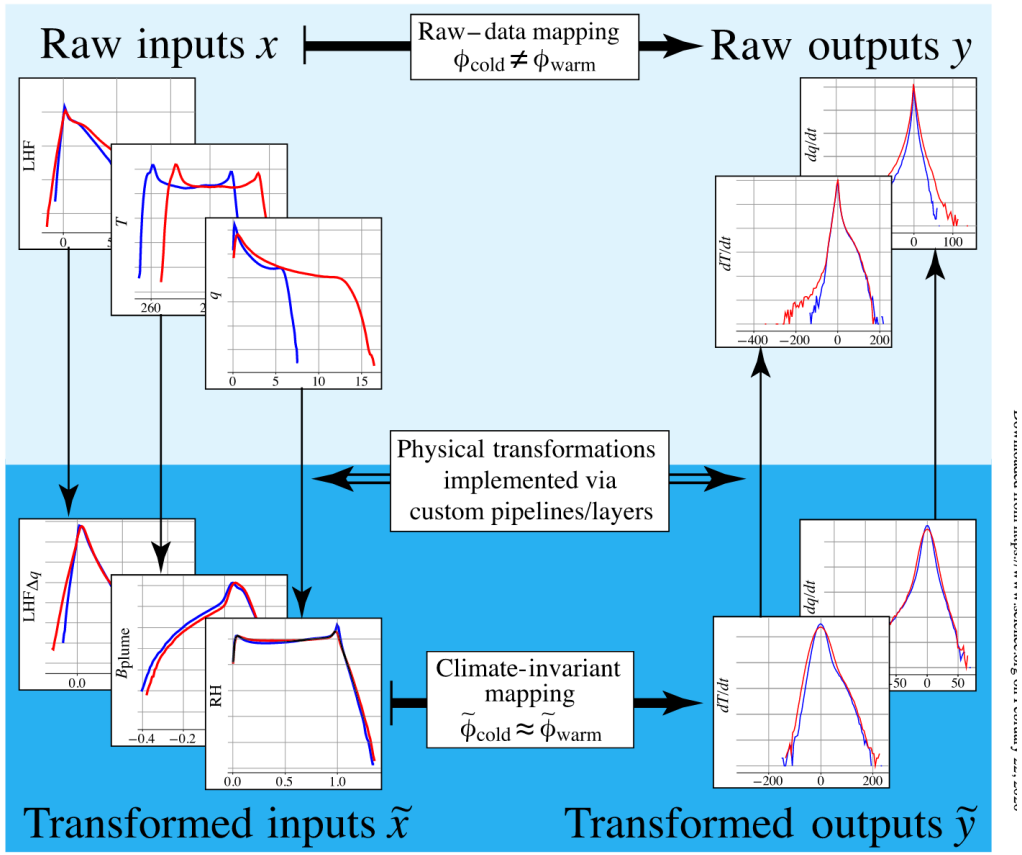


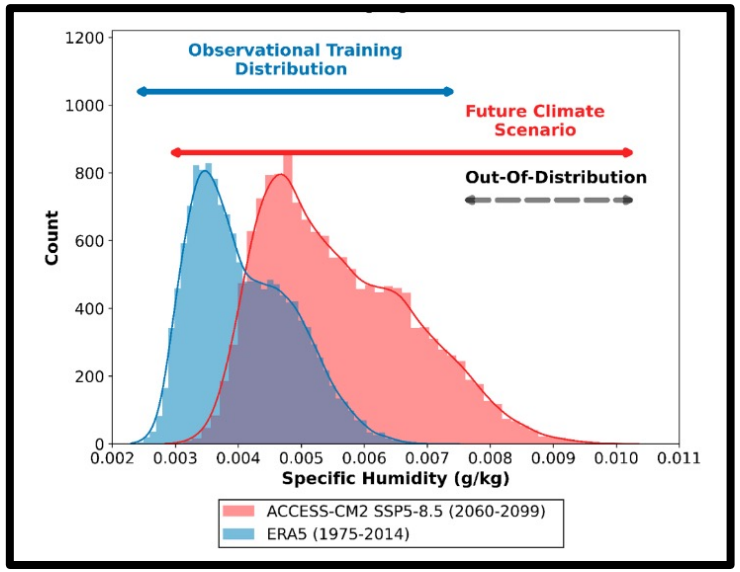


Data preprocessing

Extreme events, distribution tails

Transforming features using physical laws can improve predictions

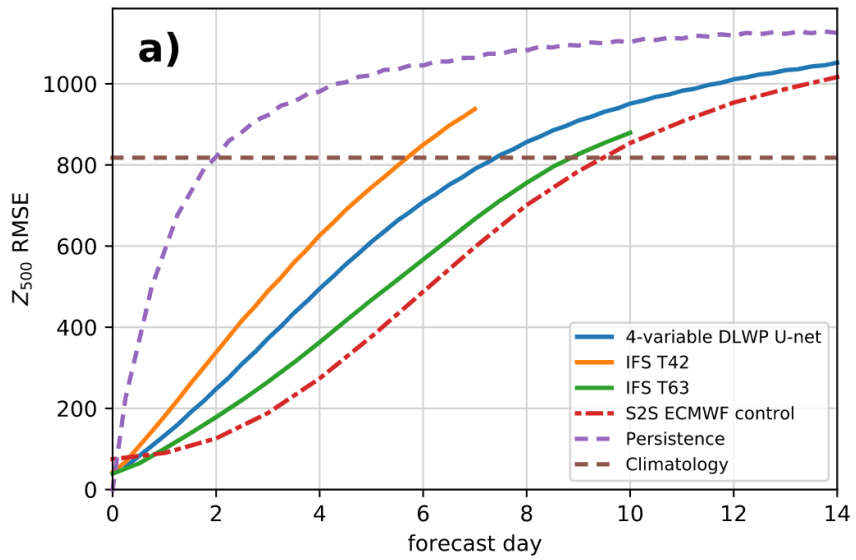


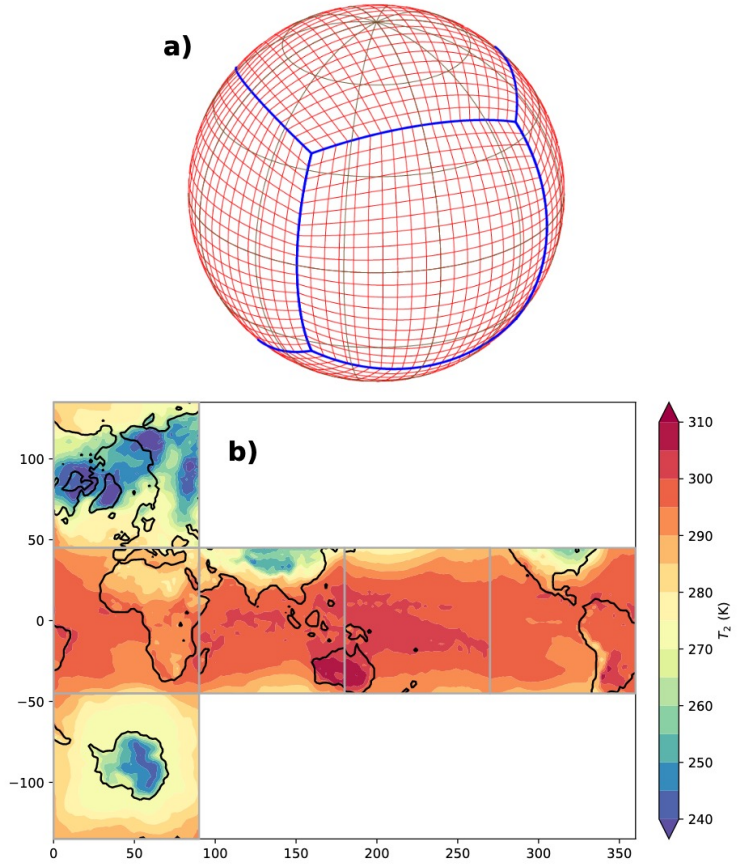


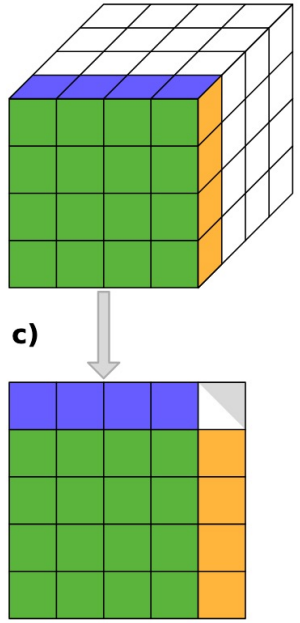
Yuval et al (2024)

Data preprocessing

Sometimes a physical remapping can improve prediction skill







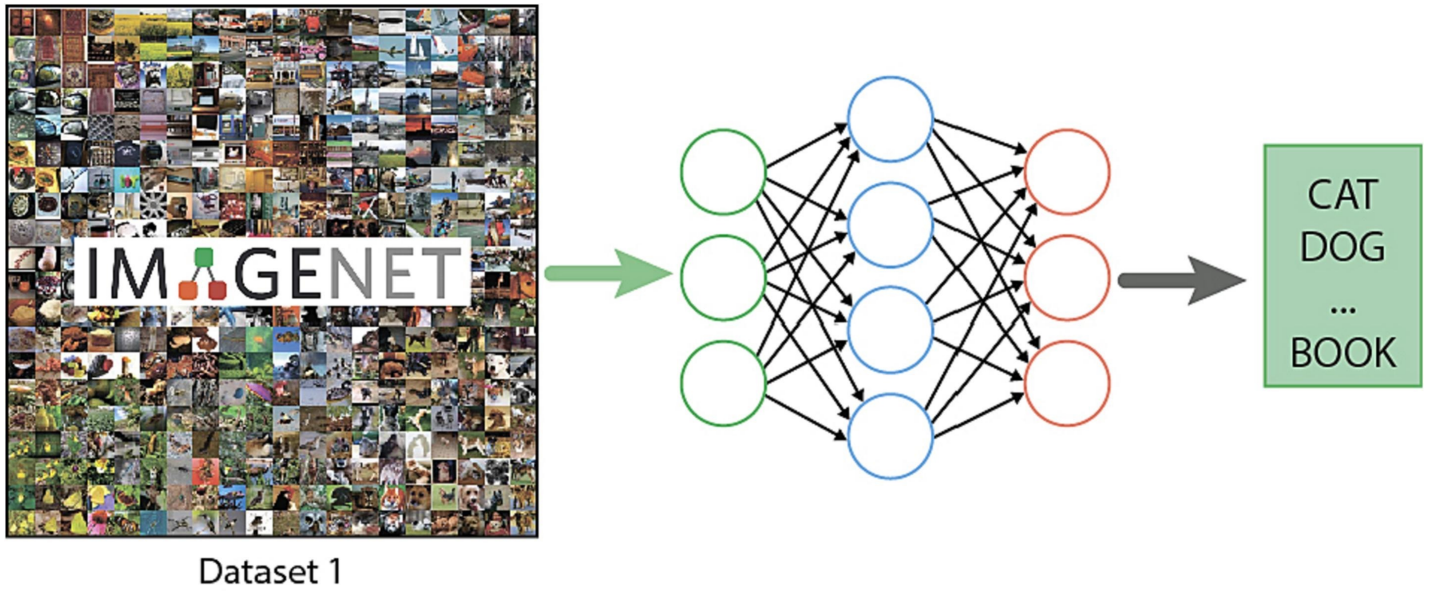
Wayne et al (2020)

Making the most of small datasets

Making the most of small datasets

Pretrained models: transfer learning

- Model is pretrained on abundance of related data, e.g., ImageNet
- Fine-tune model on smaller dataset, unfreezing select layers



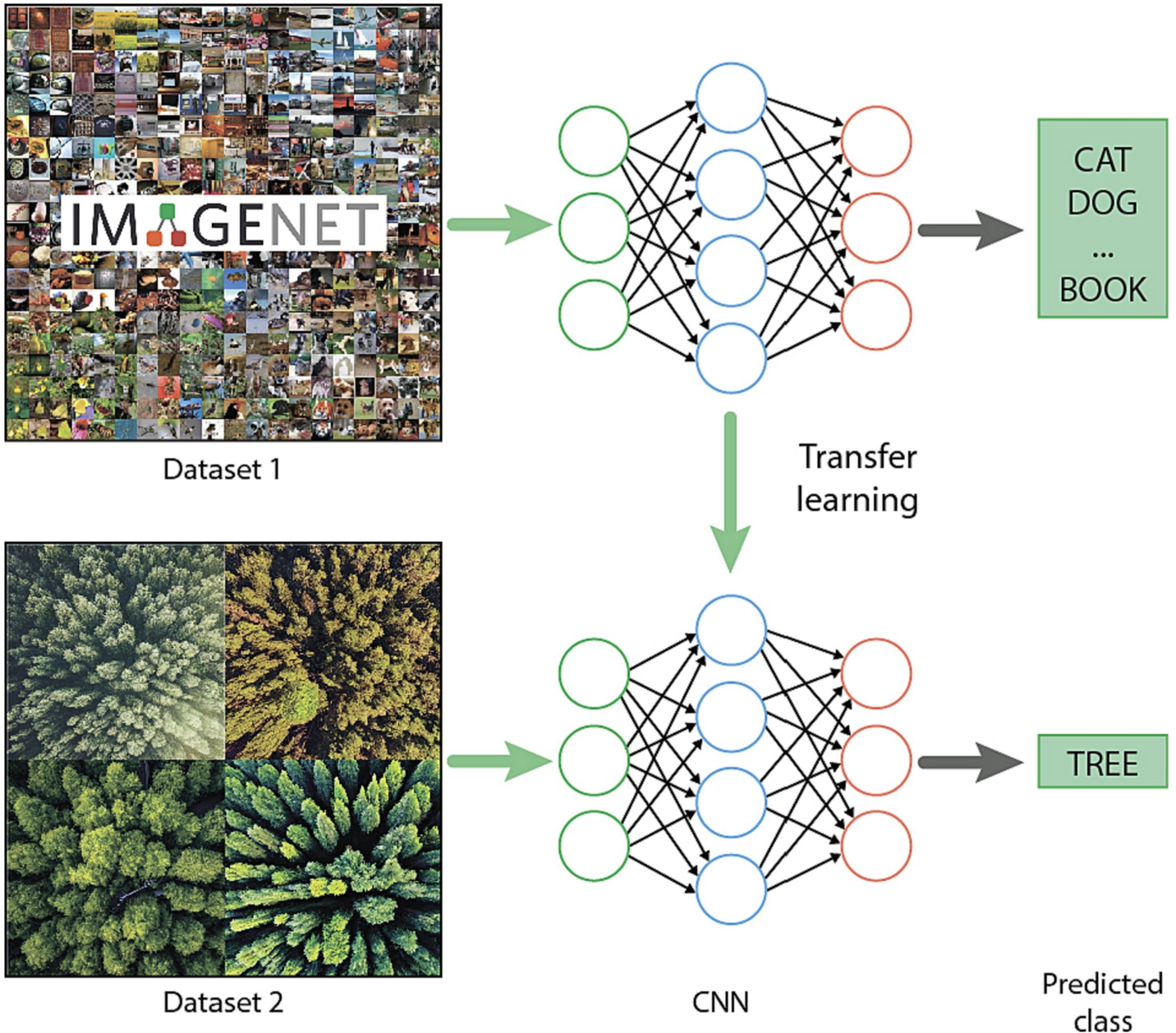
Dataset 1

Safonova et al. (2023)

Making the most of small datasets

Pretrained models: transfer learning

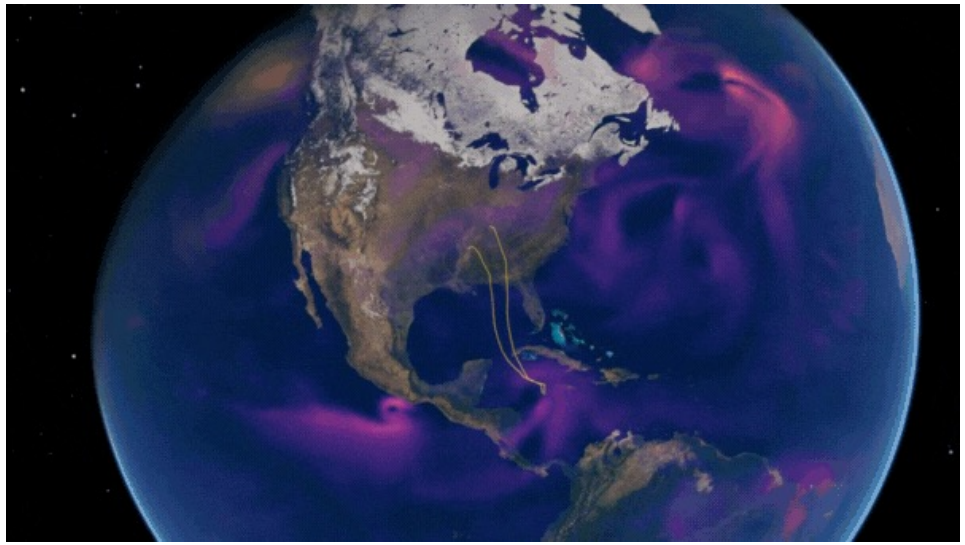
- Model is pretrained on abundance of related data, e.g., ImageNet
- Fine-tune model on smaller dataset, unfreezing select layers



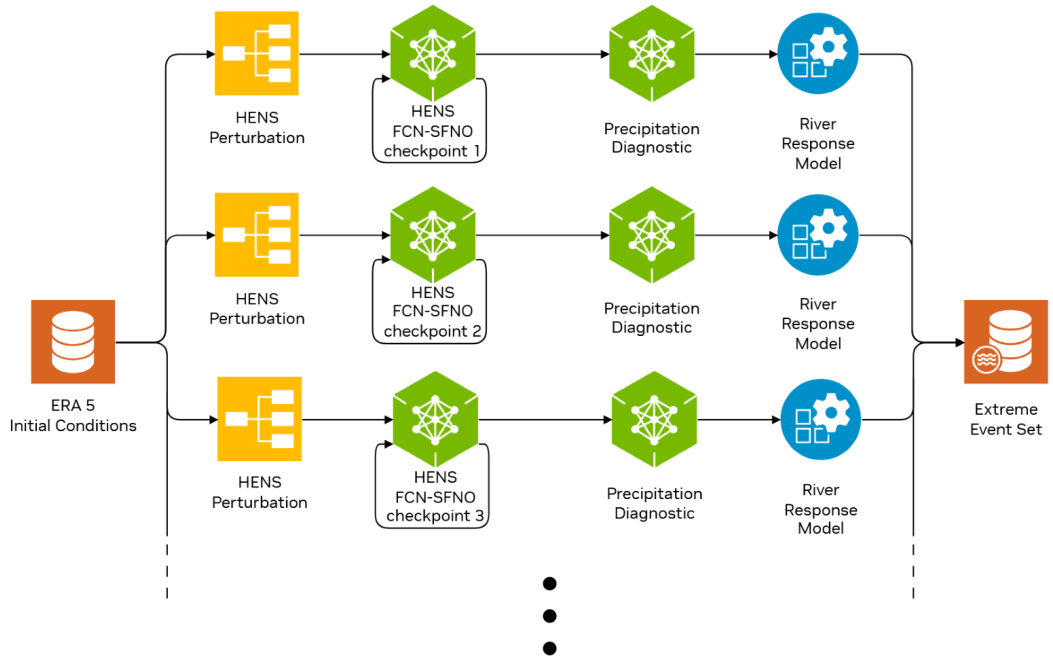
Safonova et al. (2023)

Making the most of small datasets

Pretrained models: perturb initial conditions to incorporate more extremes

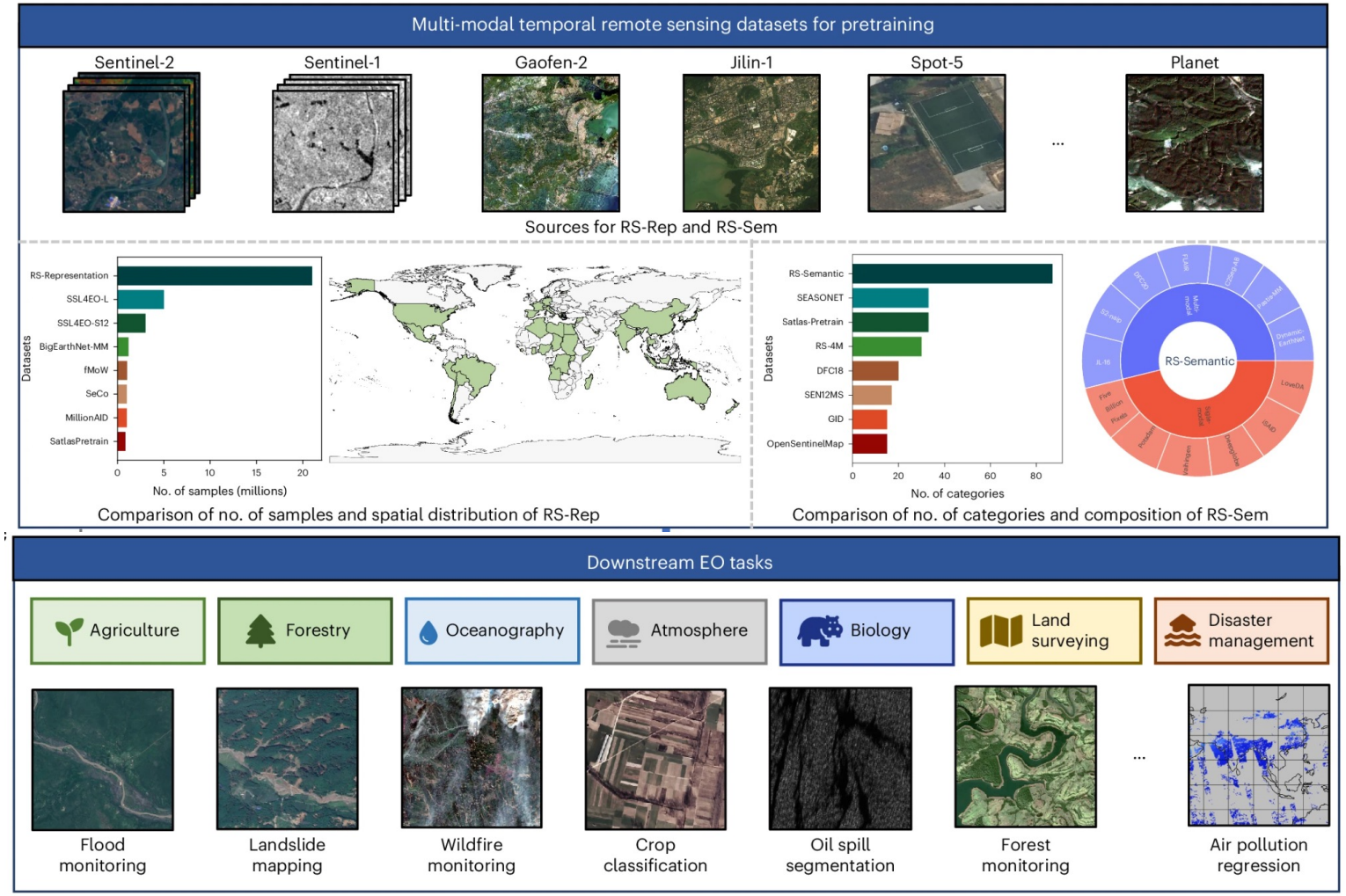


NVIDIA HENS
Huge Ensembles (HENS) for extreme-weather prediction



Making the most of small datasets

Pretrained models: foundation models



(SkySense++)

Wu et al. (2025)

Data-Centric Research

How can we ensure data (not just the model) drives performance?

- Selecting meaningful inputs and outputs
  - Ensuring spatiotemporal correlations with learning goals in mind: apples-to-apples
- Accounting for changing distributions
  - Implementing physical laws or statistical shifts when needed
  - Preserving learning goals: dense to sparse, high to low res, multiple sources, changing climate etc.
- Preprocessing choices
  - Trickiest part!
  - Staying mindful of smoothing/filtering choices and physical data structures
  - Are we removing the signal or the noise? Are we introducing artificial artifacts
- Handling small amounts of data
  - Leveraging pretrained models while not compromising learning goals

→ Domain knowledge is required!

→ helps inform model choice, methods of evaluation and relevant benchmarks

Key Takeaway

Bad data usage cannot be fixed with better models

Use domain knowledge to improve data usage in models