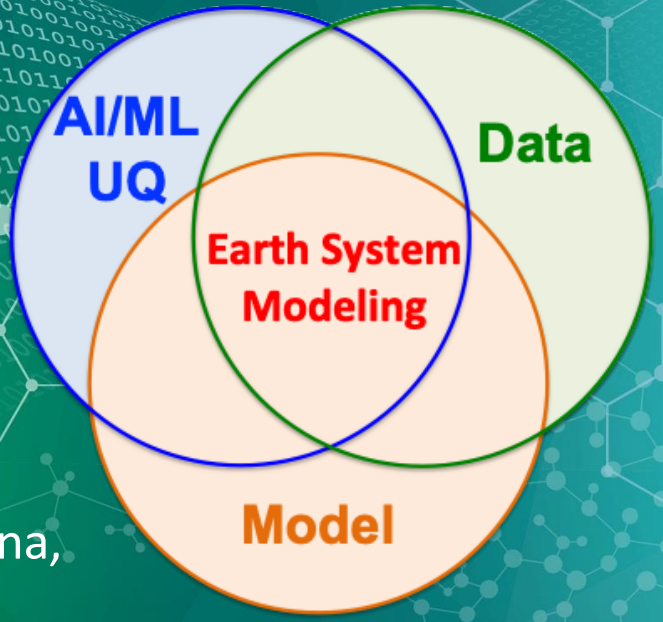
ORBIT: AI Foundation Model for Earth System Modeling



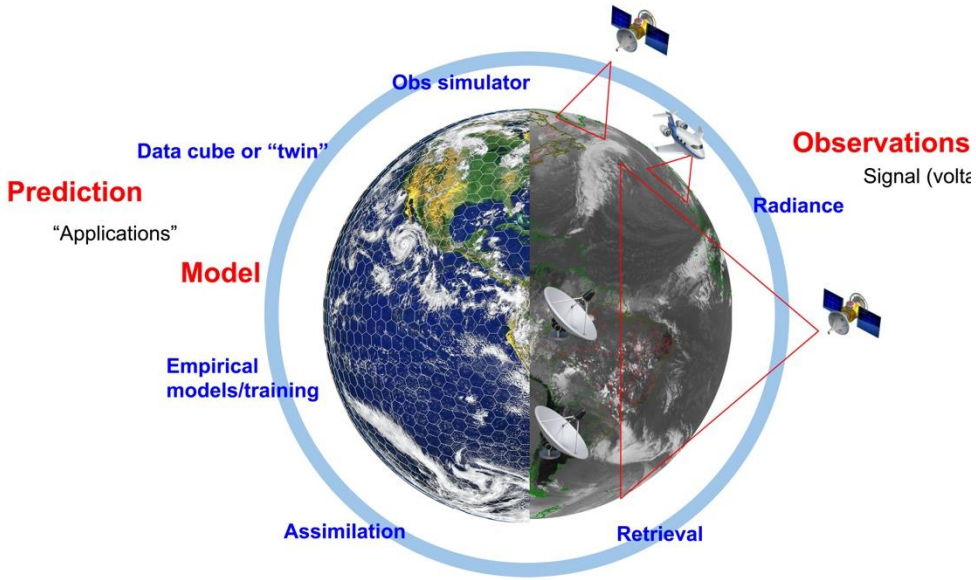
Dan Lu
Senior Computational Earth Scientist
Oak Ridge National Laboratory

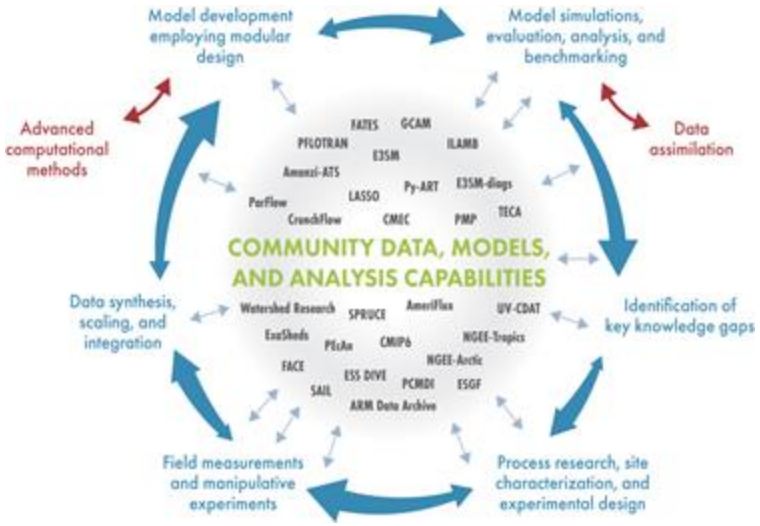
Team: Xiao Wang, Jong Youl Choi, Siyan Liu, Aristeidis Tsaris, Takuya Kurihana, Ming Fan, Moetasim Ashfaq, Wei Zhang, Junqi Yin, Prasanna Balaprakash

March 18th, 2025

Advancing Earth system modeling through data-model integration

- Understanding and predicting Earth systems are crucial for society and environment.
- Current observational systems capture only a fraction of Earth's complexity, making Earth system models (ESMs) essential to improve process understanding, reconstruct past conditions, and predict future changes.





❖ Our research advances Earth system modeling by combining diverse data and models, leveraging advanced computational methods, and integrating AI/ML techniques to enhance predictions and support informed decision-making.

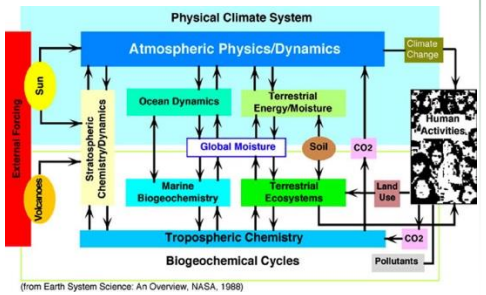
Five paradigms of Earth system modeling

1st Paradigm: Empirical Model

Experiments to explain or empirically describe natural phenomena

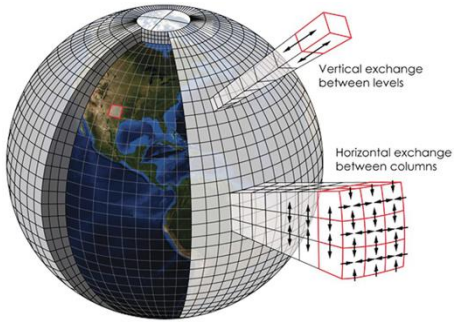
2nd Paradigm: Theoretical Model

Develop physical laws, theoretical models



3rd Paradigm: Computational Model

Computational models, simulating complex, coupled Earth system



4th Paradigm: Data-driven ML Model

A deluge of Earth system data have become available; Derive ML models from observation and simulation data.

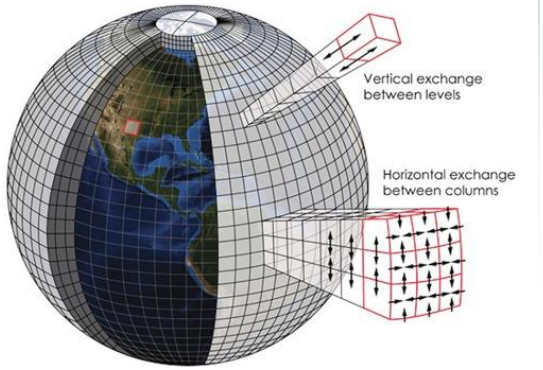
5th Paradigm: AI Foundation Model

1600 — 1950 — 2000 — 2020

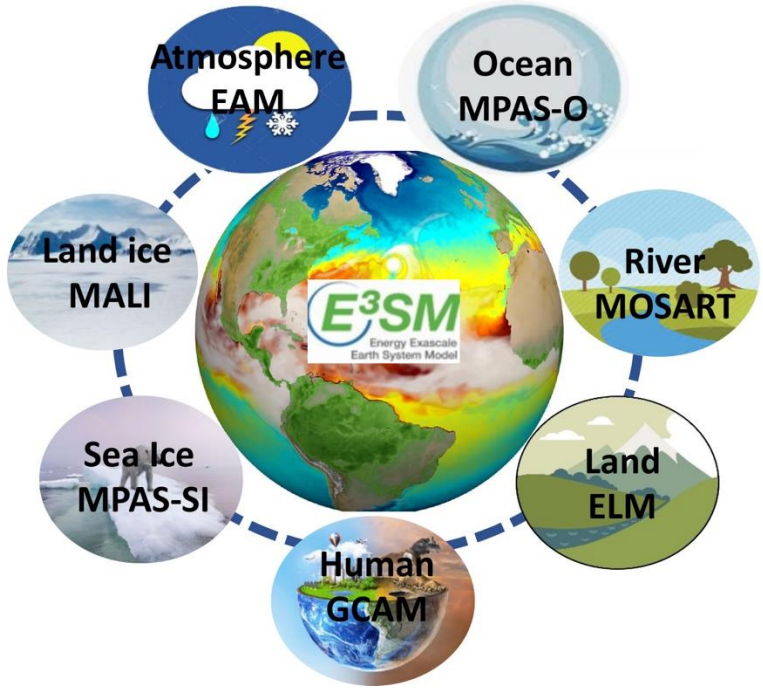
Physics-based Earth system prediction is filled with uncertainty

3rd Paradigm: Computational Model

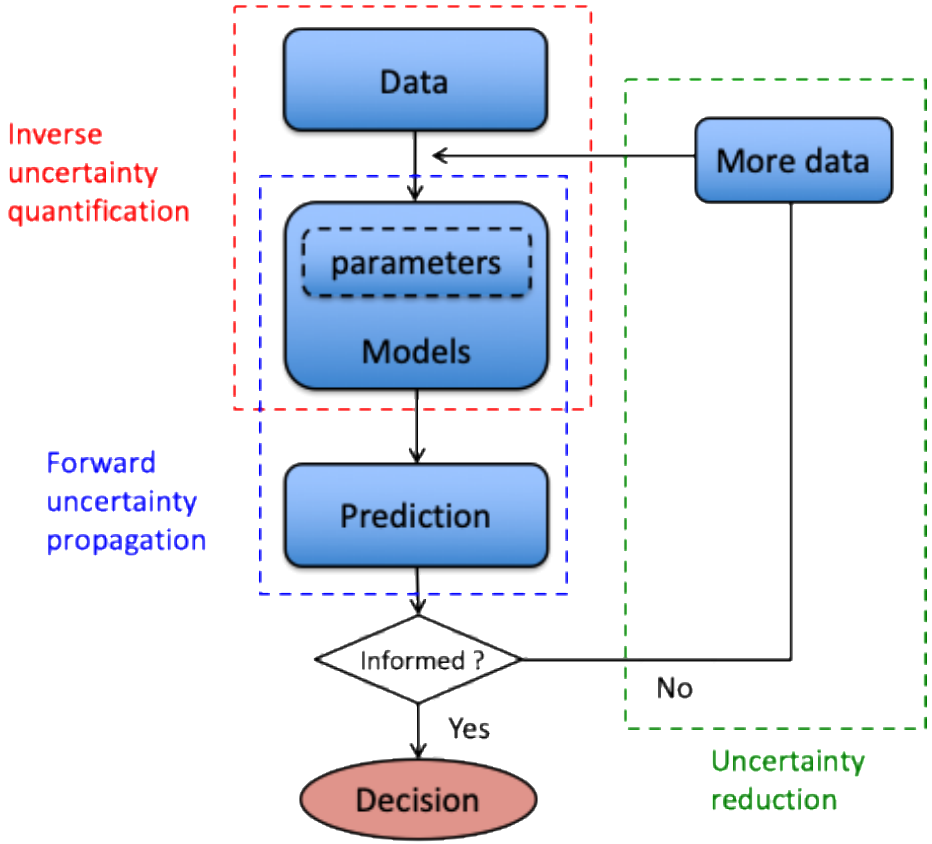
Computational models, simulating complex Earth system



Department of Energy (DOE)'s Energy Exascale Earth System Model (E3SM)



Physical-based ESM prediction

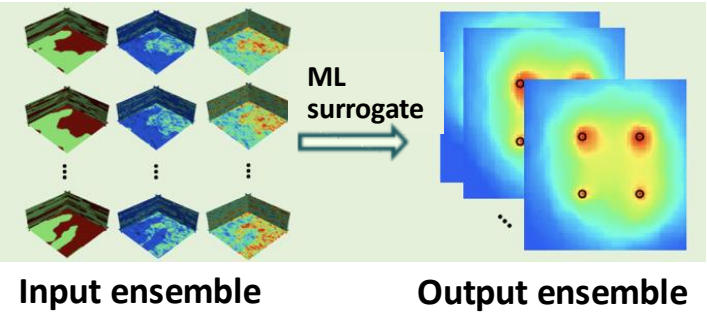


❖ Uncertainty quantification (UQ) is crucial for physics-based Earth system modeling to enhance prediction accuracy and support informed decision-making.

ML techniques for fast physical model prediction and UQ

Surrogate Modeling

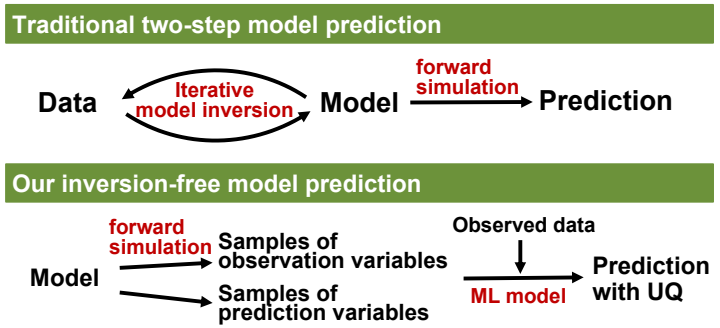
Build a fast surrogate of expensive numerical model based on ensemble model simulations



- Surrogate modeling reduces time of a single model run.
- Evaluation of the surrogate in UQ reduces total costs.
- Use NNs to build a surrogate.

Inversion-Free Prediction

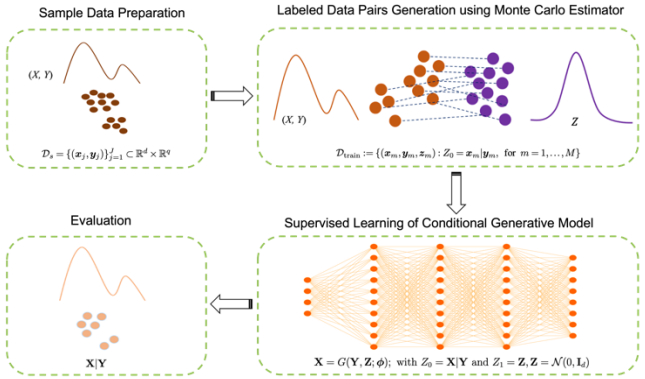
Learn obs-pred relationship and then make direct prediction from observed data



- Avoids expensive, iterative inverse modeling.
- Computationally efficient, fully parallel, fast data assimilation.
- Consider various uncertainties.

Generative AI

Diffusion model generates samples for both forward and inverse UQ by evaluating NN



- Our conditional diffusion model uses NN to estimate a generator and evaluates the NN to generate samples for UQ.
- Computationally and storage efficient.

❖ HPC and ML are critical tools to advance physics-based Earth system modeling and UQ.

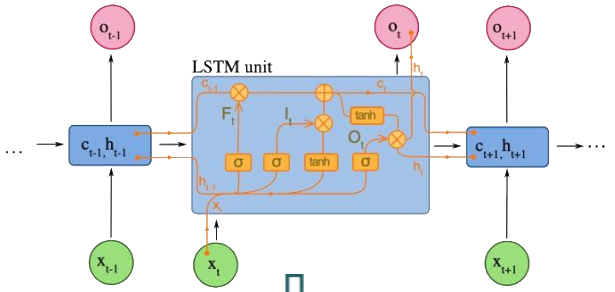
Ensuring trustworthiness in data-driven Earth system modeling

4th Paradigm: Data-driven ML model

A deluge of Earth system data have become available; Derive ML models from observation and simulation data.

LSTM network learns system dynamics from observations of environmental drivers and carbon/water fluxes to predict future carbon/water fluxes

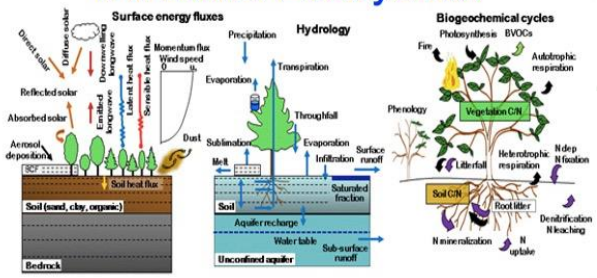
Long Short-Term Memory (LSTM)



LSTM simulates a mapping for the inputs over time to an output to consider the memory effect of drivers.

Input: Observation of environmental drivers

Terrestrial Ecosystem



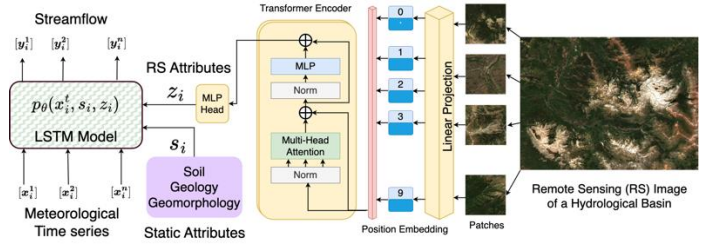
Output: Observation of carbon/water flux

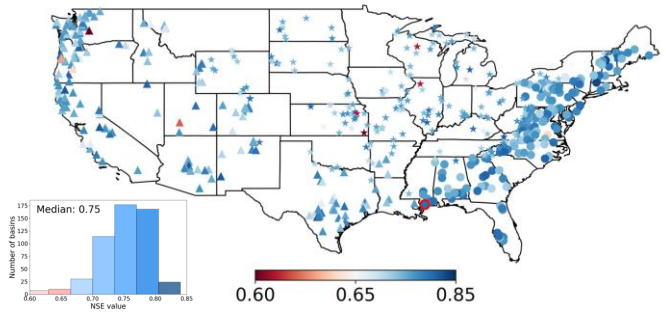
- ML models have shown success in Earth system prediction, but they have challenges for trustworthy prediction:
  - How can we ensure that ML solutions generalize across space and time?
  - How do we verify that models are making good predictions for the right reasons?
  - How can we guarantee prediction reliability under changing environmental conditions?

Advanced, explainable, reliable ML for Earth system prediction

Advanced ML

- Integrate diverse data from satellite and sensor networks
- Develop advanced model architectures



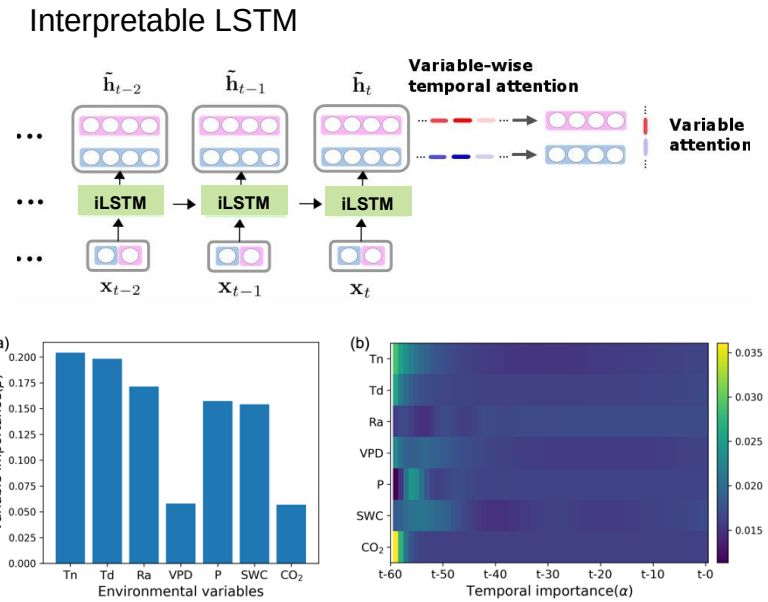


NSE value: = 1.0 means perfect fit; > 0.65 suggests good prediction

❖ Leverage diverse data and advanced ML models to improve accuracy and generalizability.

Explainable ML

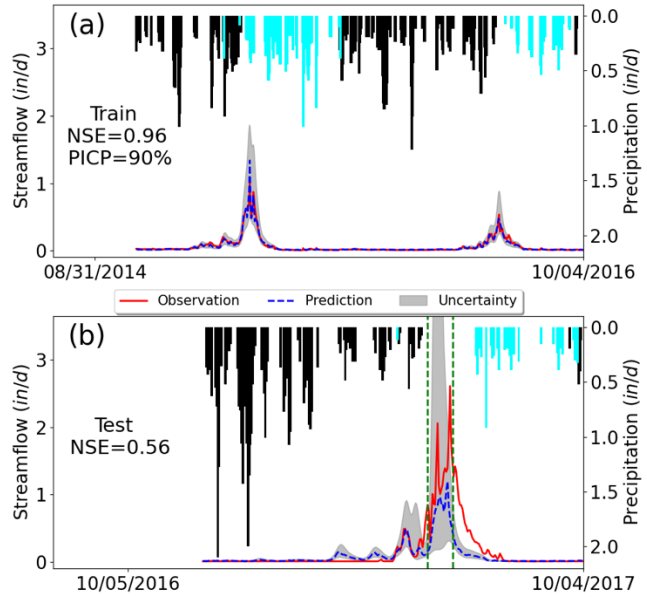
- Permutation analysis: SHAP
- Gradient-based method: IG
- Interpretable LSTM network
- Attention maps of transformer model



❖ Validate model decisions ensuring physical consistency; identify key drivers for prediction.

Reliable ML

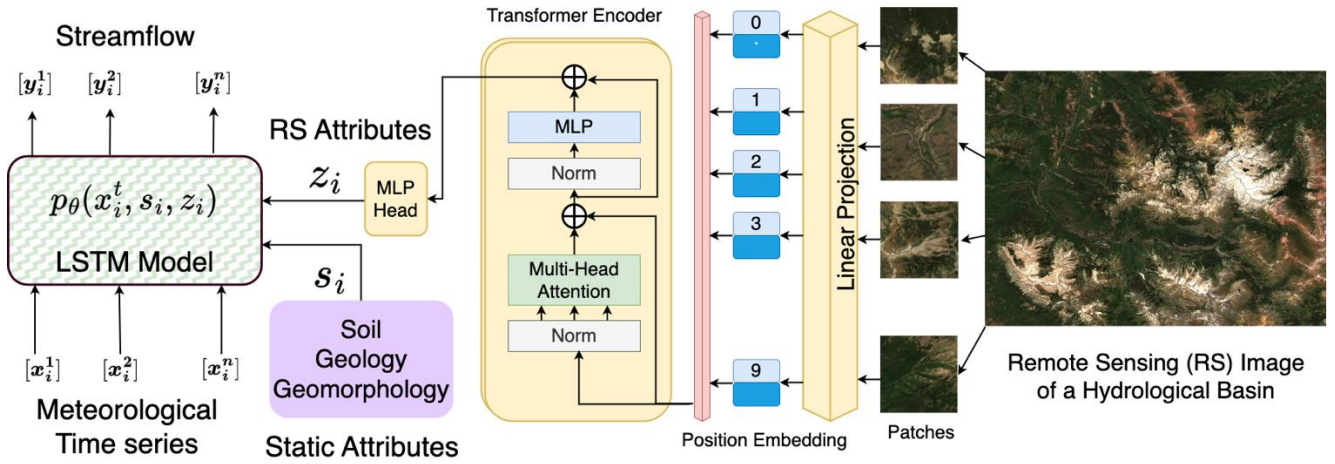
- Bayesian neural networks
- Gaussian processes
- Ensemble-based methods
- Prediction interval methods



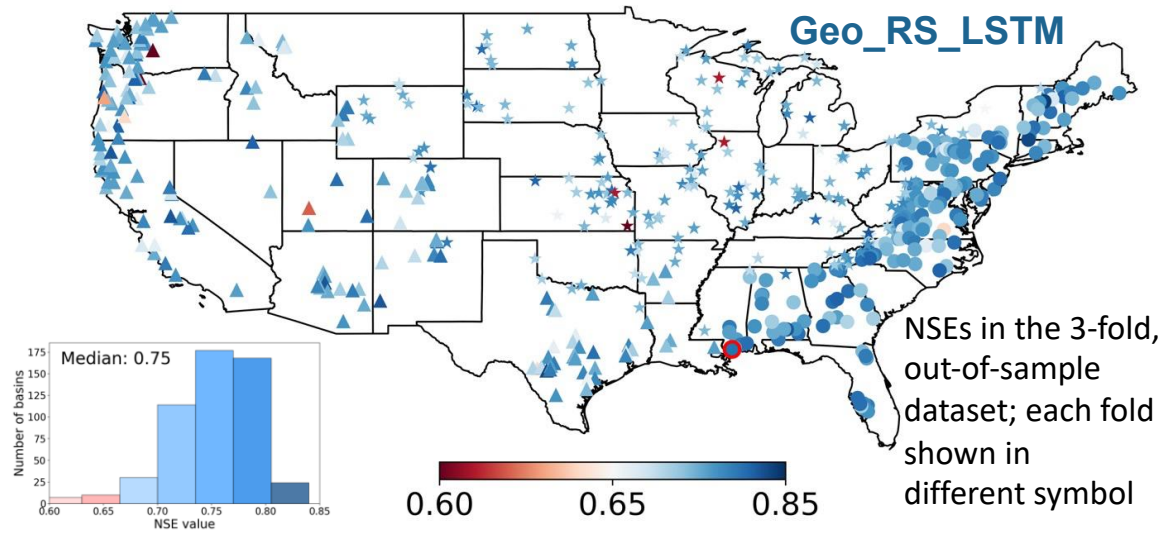
❖ Quantify prediction uncertainty to evaluate & ensure reliability under changing conditions.

Advanced ML models with diverse data enhance generalization

- Problem: Predict streamflow across the US;
- Data: 35 years of CAMELS dataset of 531 basins and Sentinel-2 satellite images;
- Model: Transformer + LSTM integrating multiple data sources;
- Evaluate: Performance in spatiotemporal out-of-sample prediction using the NSE metric (value of 1 is the best).
- Perform 3-fold cross-validation.



	1980-2007	2008-2014
354 basins	Training	
177 basins		Evaluation



NSE value: = 1.0 means perfect fit; > 0.65 suggests good prediction

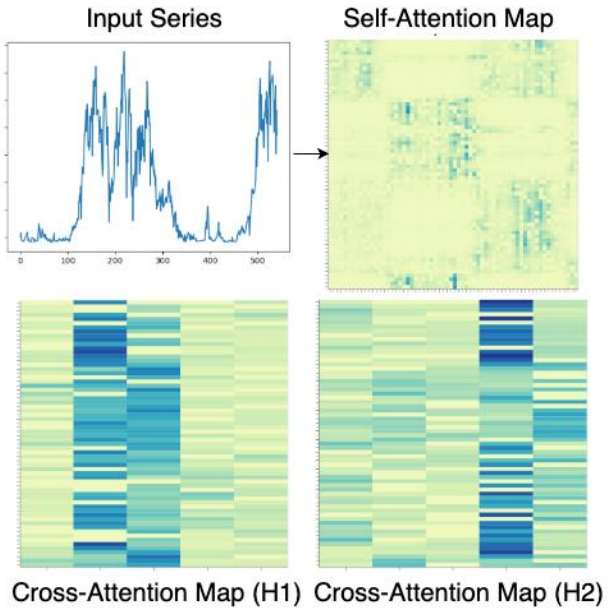
❖ Advanced ML model, integrating diverse data, improves prediction performance at out-of-sample regime.

- Singh, Lu, and Tayal, *NeurIPS*, 2024. Tayal, Renganathan, and Lu, *ERL*, 2024.

Explainable AI can guide ML and process-based model development

Transformer-based model

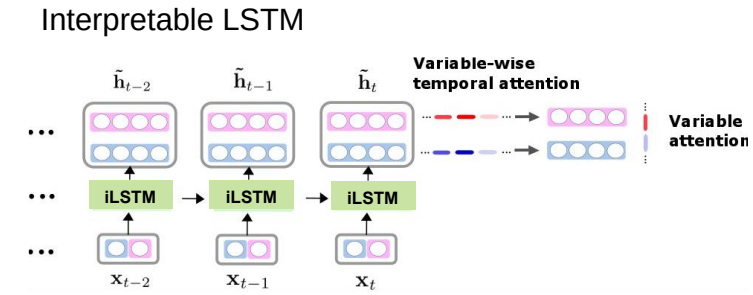
- Visualize Transformer model's learning process to improve prediction understanding.



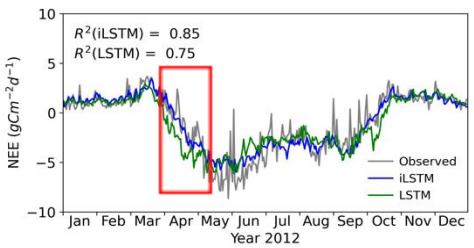
- Self-attention identifies temporal pattern of each driver;
- Cross-attention captures relationships among drivers.

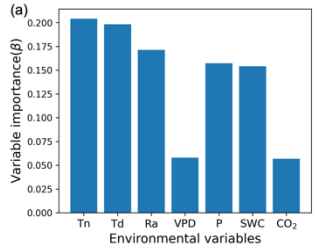
Interpretable LSTM (iLSTM)

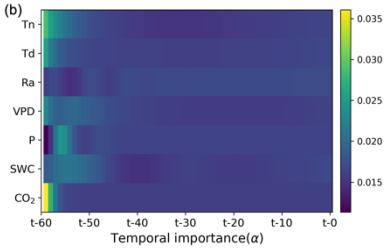
- iLSTM explains variable and temporal importance through its advanced model architecture.



- Uses variable-wise hidden matrix;
- Adds temporal and variable attention;







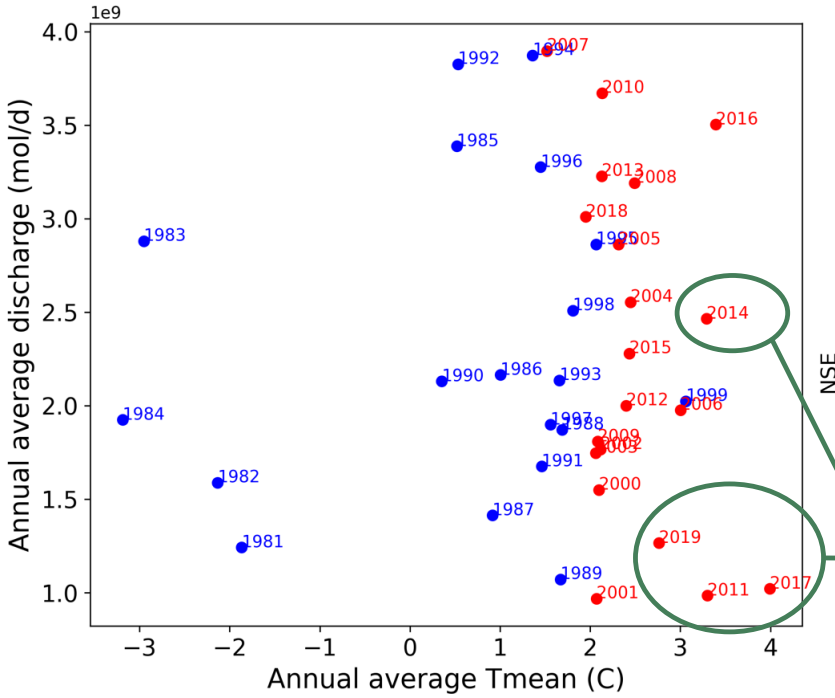
- iLSTM achieved more accurate prediction;
- iLSTM revealed new variable relationships and their temporal importance.

❖ Advanced interpretable ML models enhanced prediction accuracy, revealed learning processes, and provided insights to inform process-based model development.

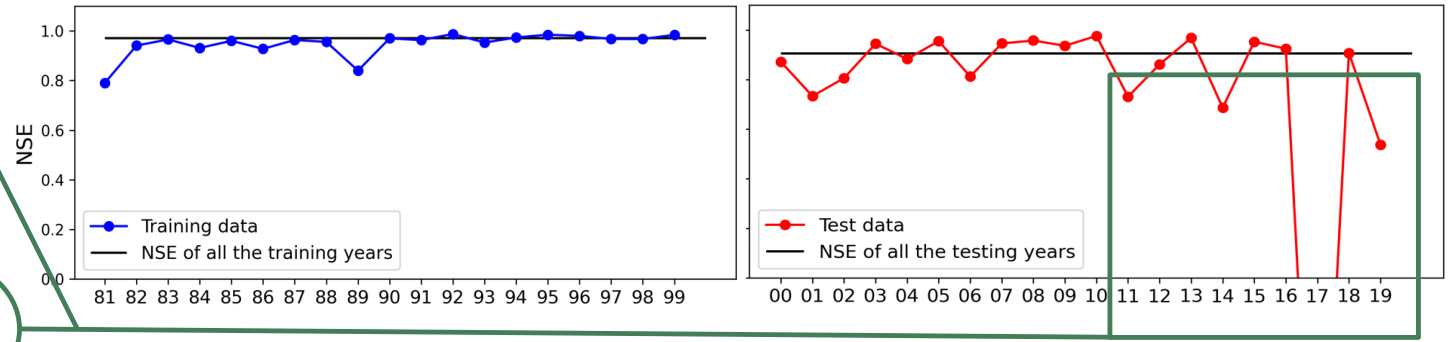
- Lu, Ricciuto, and Liu, *ICLR*, 2022. Tayal, Renganathan, and Lu, *ICML*, 2024.

ML model needs UQ for trustworthy prediction under climate change

- ML model typically perform well under conditions similar to those they have been trained on but struggle with new, unseen conditions.
- Identifying the reliability of ML predictions is crucial for their effective use.
- UQ helps address the challenge of assessing ML model reliability in climate projection.



- Use LSTM to predict streamflow in East River from met. data.
- Train on 20 years of data (blue dots in cool years); and evaluate on subsequent 19 years (red dots in warm years)
- LSTM performance deteriorates when extrapolating the warmer years.



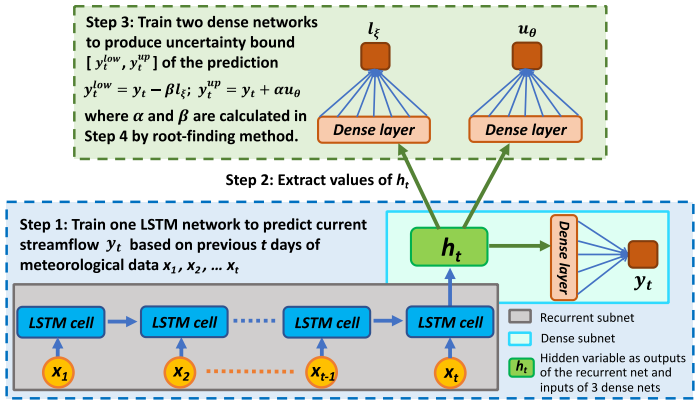
Train: 1981-1999

Test: 2000-2019

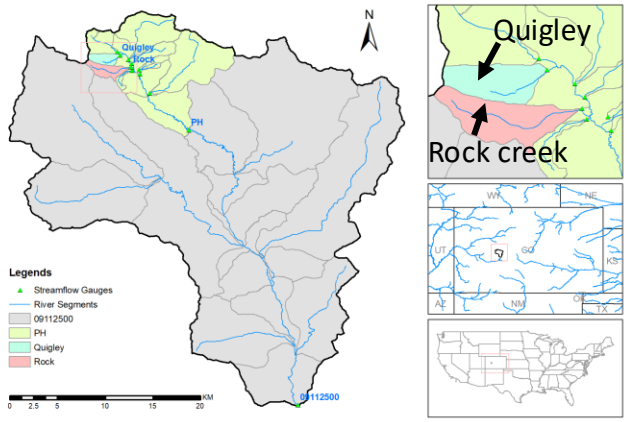
- Topp, S., Barclay, J., Diaz, J., Sun, A., Jia, X., Lu, D., Sadler, J., and Appling A., *WRR*, 2022.

UQ ensures reliable prediction under changing conditions

Our UQ method produces prediction and its uncertainty using three NNs.

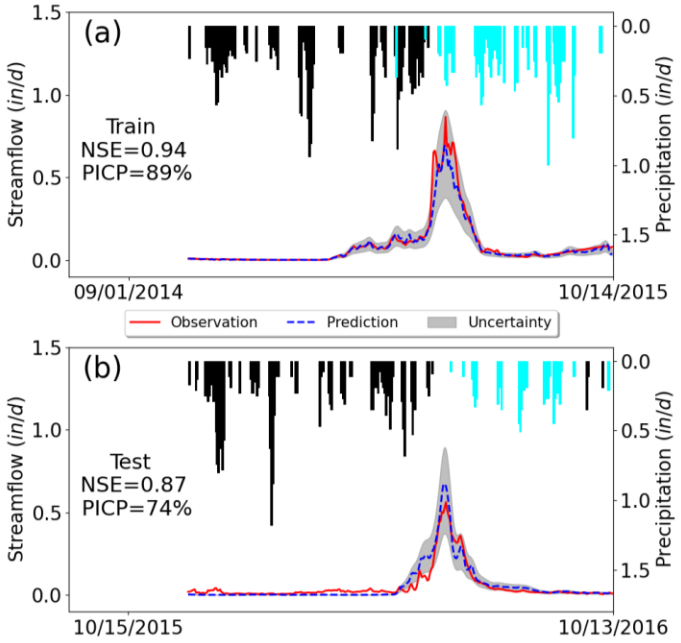


East River Watershed, CO



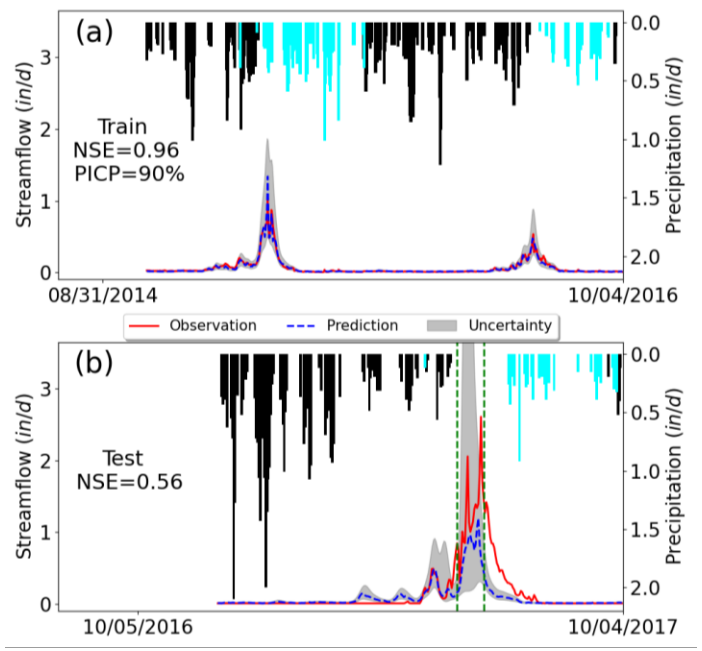
- Input: precip, max and min air T
- Output: daily streamflow
- Model: LSTM network
- UQ: calculate 90% prediction interval

Catchment Quigley



- In Quigley where test and training conditions are similar, LSTM accurately predicts the streamflow.
- Our UQ method accurately quantifies prediction uncertainty consistent with the confidence level.

Catchment Rock creek



- In Rock Creek, LSTM cannot predict the test data well due to data shift and new conditions.
- Our UQ method detects this shift by producing a wider uncertainty consistent with larger errors.

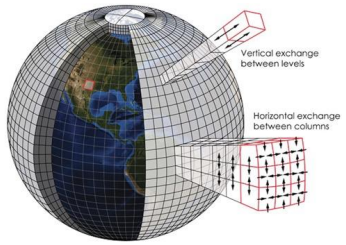
❖ Our error-consistent UQ method prevents overconfidence and ensures reliable predictions under changing conditions.

- Lu et al., *JHM*, 2022; Liu, Lu, Painter, Griffiths, and Pierce, *Frontiers in Water, 2023*.

From physics-based to data-driven, now to AI foundation models

3rd Paradigm: Computational Model

Computational models, simulating complex Earth system

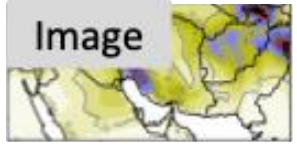


4th Paradigm: Data-driven ML model

ML models simulate the Earth system from data

Data

Image



Spatiotemporal

Time series

Database

Training

AI Foundation Model

Adaptation

Applications

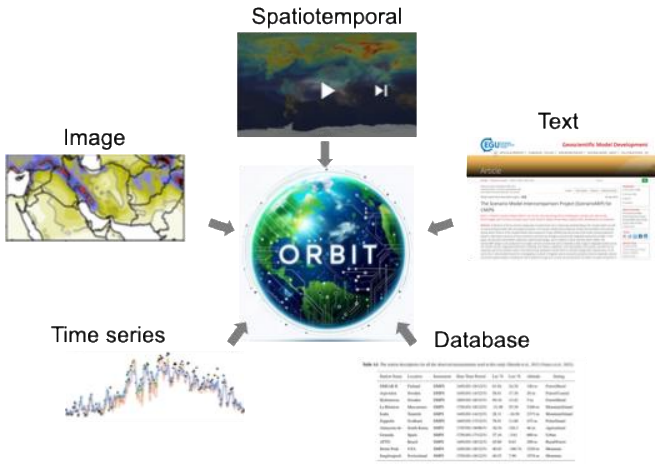
- Weather, climate prediction
- Climate projection
- Climate simulation downscaling
- E3SM simulation acceleration

❖ An AI foundation model is a large-scale neural network trained on extensive, diverse datasets and adaptable to a variety of modeling tasks.

AI foundation model can advance Earth system modeling

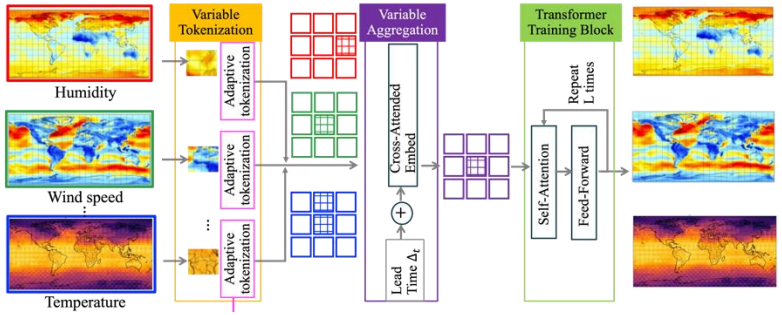
Heterogeneous Data

- Observations from lab, field, and satellite
- Model simulation data
- Data have multiple types, scales, and resolutions.
- These heterogeneous data cannot be fully integrated by numerical models and task-specific ML models.



Scalable Model

- Vision Transformer model
- Integrate heterogenous data
- Scale with data size and resolution

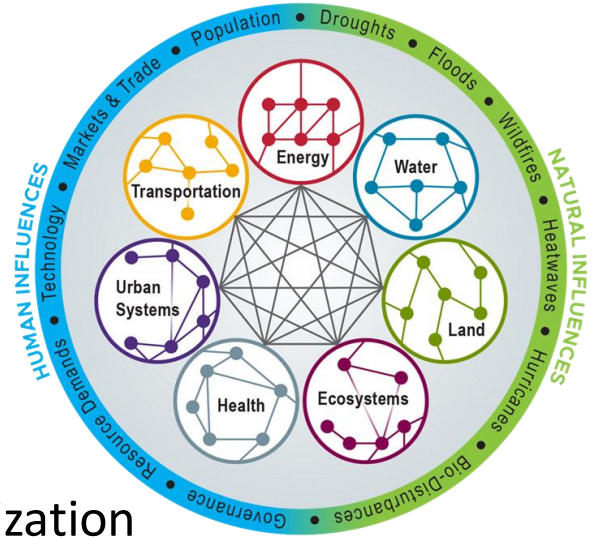


Foundation model:

- Integrate rich, multimodal data
- Reduce reliance on labeled data
- Improve accuracy, efficiency, and generalization
- Ensure high versatility

Various Applications

- Earth system is a coupled system.
- Its simulation advances various scientific applications and impacts multiple sectors.
- Foundation models can save effort, cost, and energy.



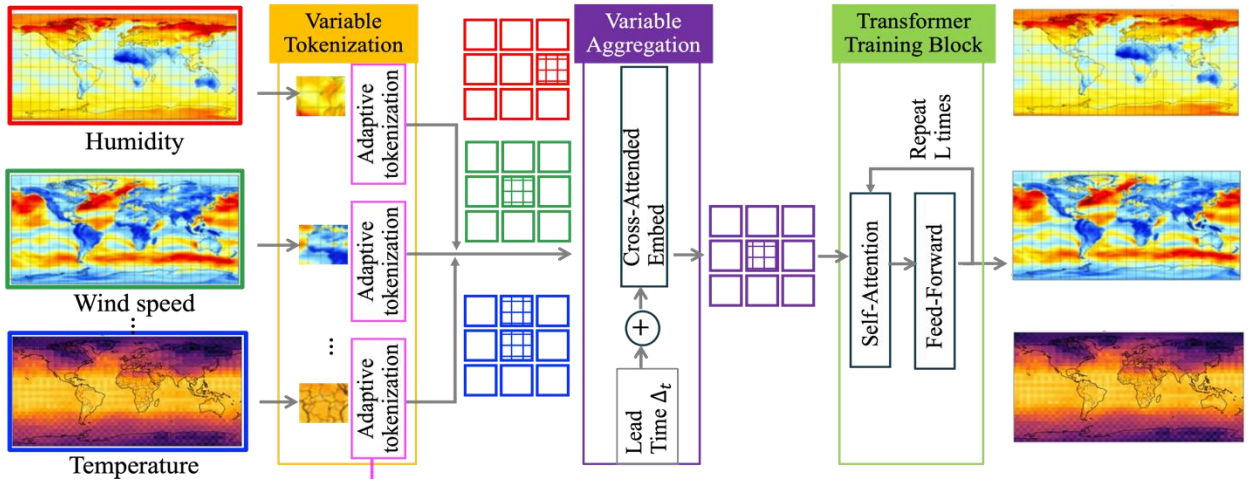
ORBIT: our AI foundation model for Earth system modeling

Pre-train on CMIP6 simulation dataset

- Simulation data from 10 CMIP6 models;
- Each model provides 65 to 100 years of data at 6h interval;
- Consider 91 variables with spatial-res of 128*256;
- 1.2 million data point and 223.6 billion tokens.

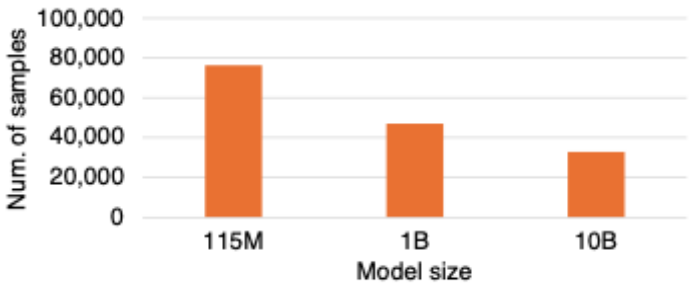
- Use ESGF to access data and PMP to select quality data.

Develop large ViT models to enable effective learning of Earth systems from extensive data



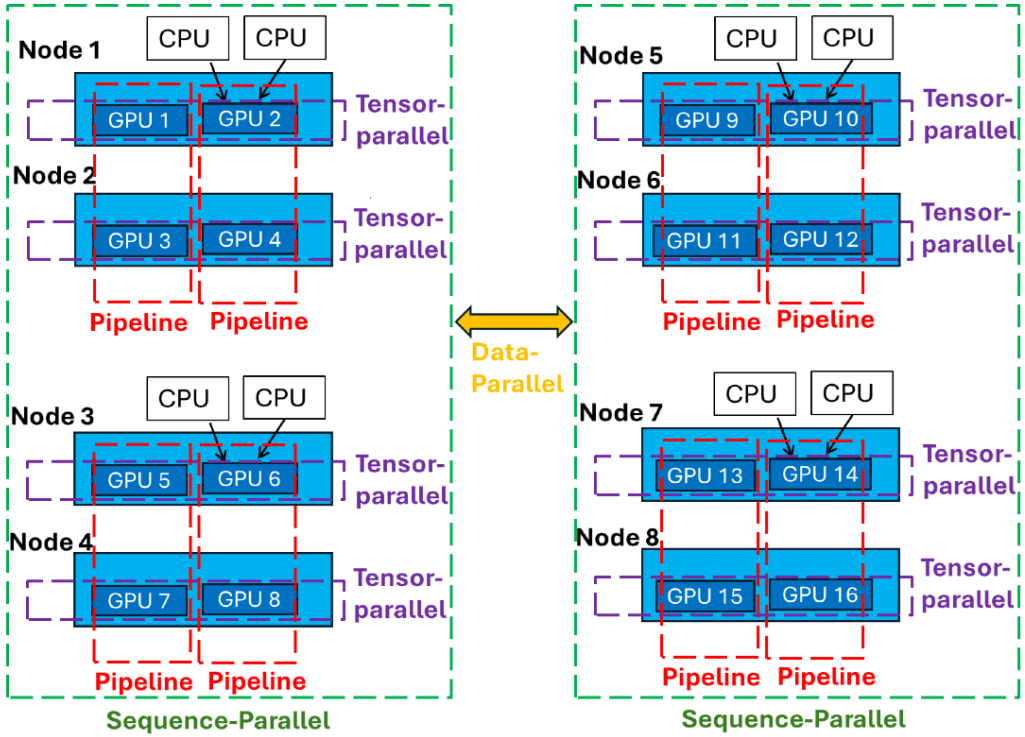
- ORBIT has four model sizes with 115M, 1B, 10B, and 113B parameters.
- It is the largest AI model for Earth system.

Larger models are more effective in Earth system modeling



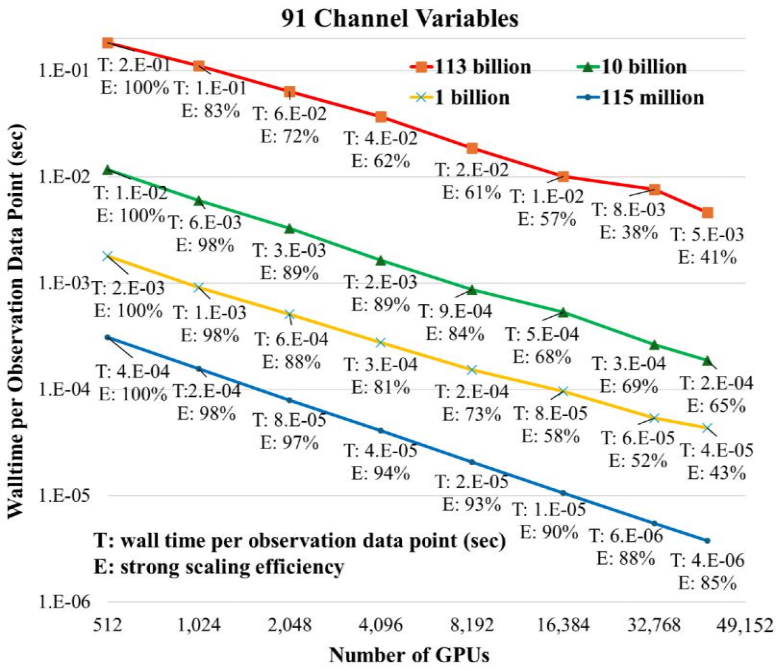
- As model size increases, the required training samples decreases in Earth system modeling fine-tuning tasks;
- This data efficiency can lead to significant cost and time savings in various Earth system modeling applications.

ORBIT achieves strong scaling efficiency on Frontier supercomputer



- We develop a novel hybrid model-data-sequence parallelism that merges
  - Tensor
  - Pipeline
  - Data
  - Sequence

  parallelism orthogonally to accelerate ORBIT training.



- ORBIT achieves 1.6 exaflop sustained computing throughput on 6,144 Frontier nodes (49,152 GPUs), with strong scaling efficiency between 44% to 85% for model sizes of 100M, 1B, 10B, and 113B.

Collaborating with

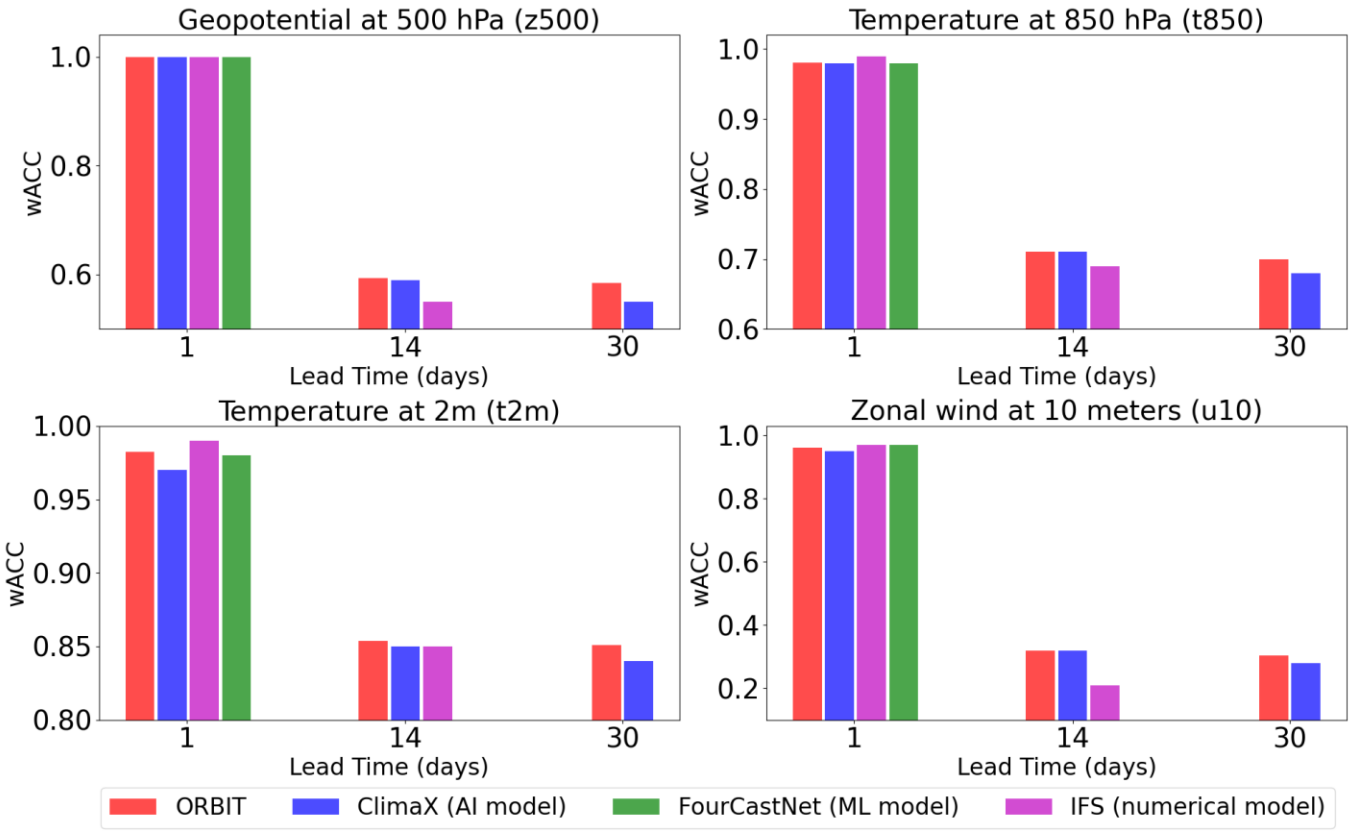
- Microsoft DeepSpeed4Science Team
- AMD Team on Frontier platforms for AI

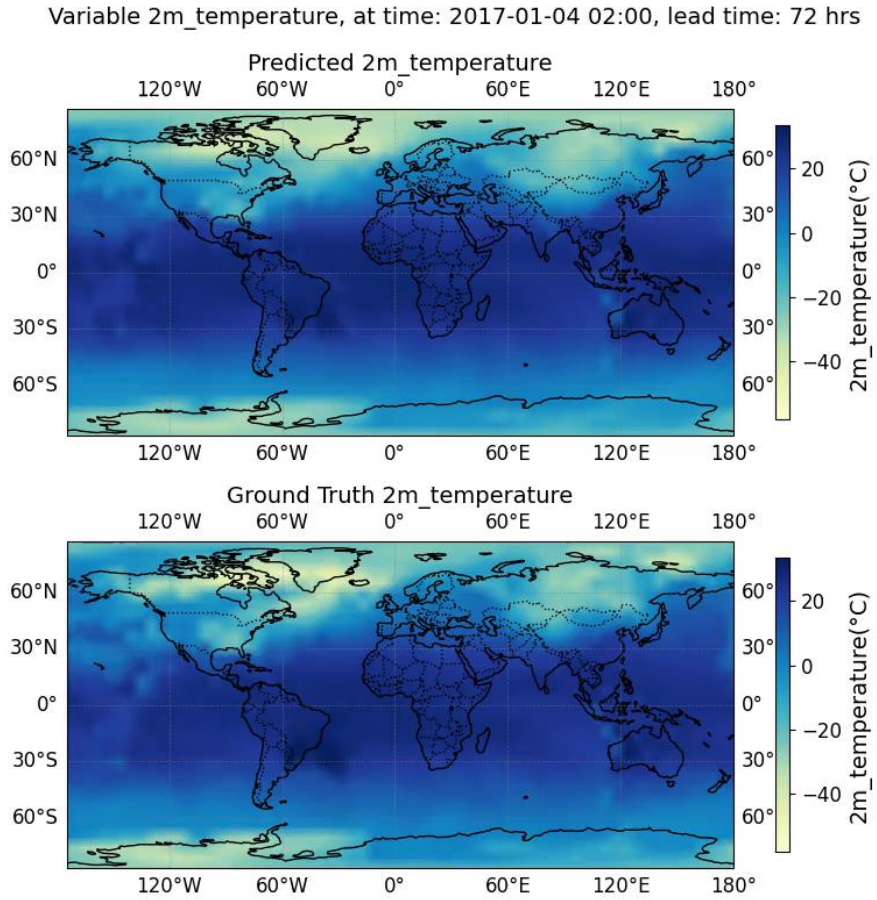
Microsoft AMD

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

ORBIT provides fast and accurate weather forecasts

- Finetune ORBIT using ERA5 data for weather forecast





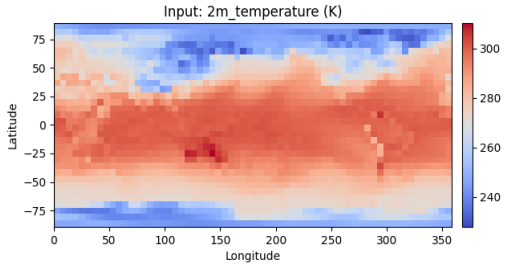
❖ ORBIT achieves competitive performance in weather forecasting, matching or surpassing state-of-the-art numerical, machine learning, and foundation models.

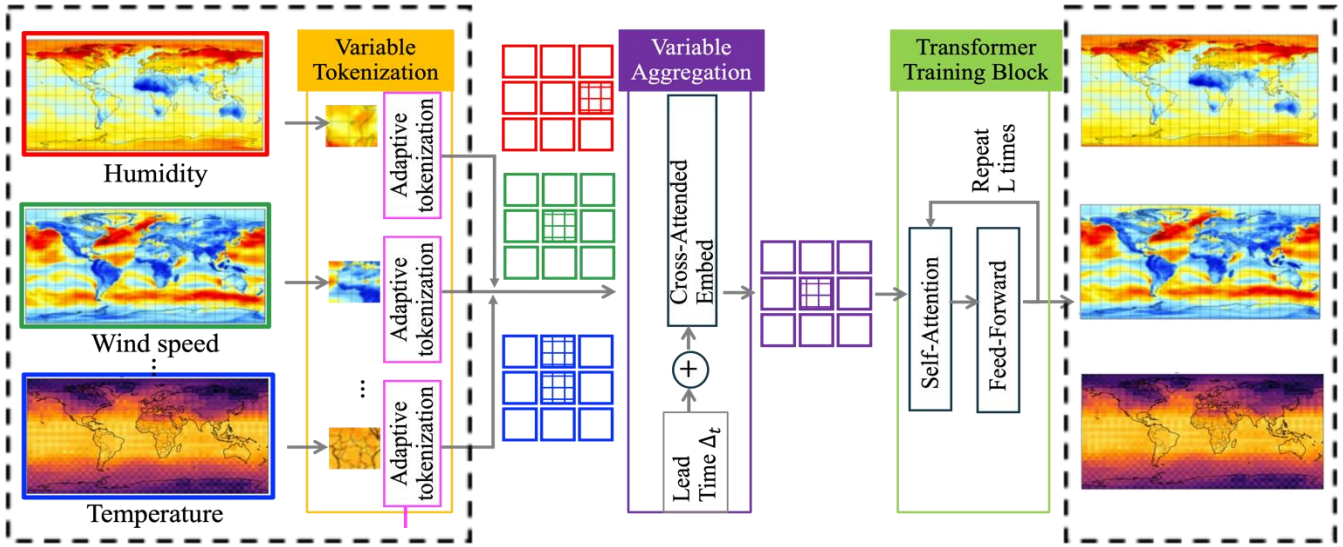
Model Size	115 million
GPUs	1 GPU
Forecast Time	0.04 sec

ORBIT can be used for weather/climate downscaling

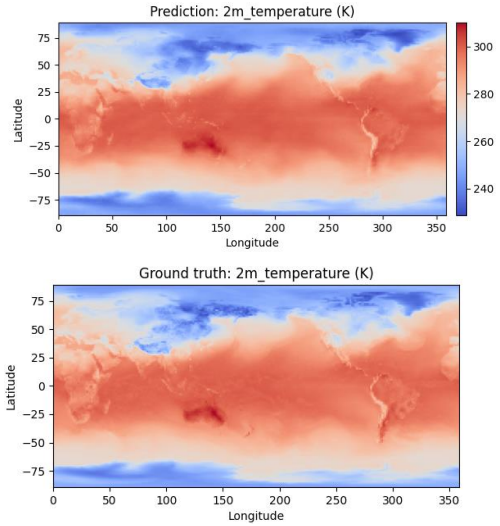
- Finetune ORBIT using pairs of low-resolution and high-resolution data for downscaling

Input: low-res data

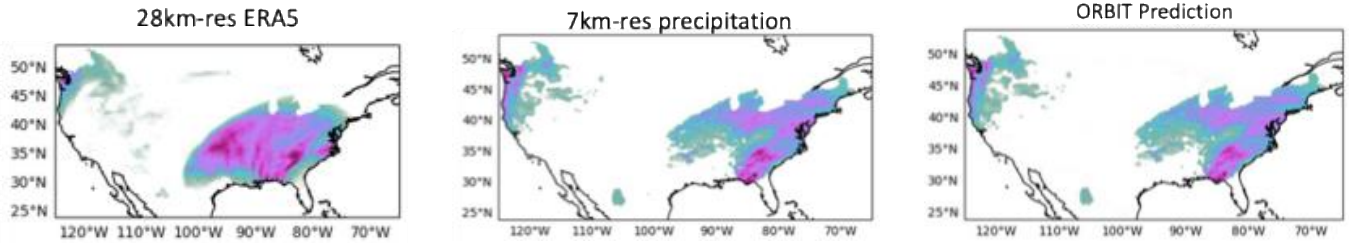


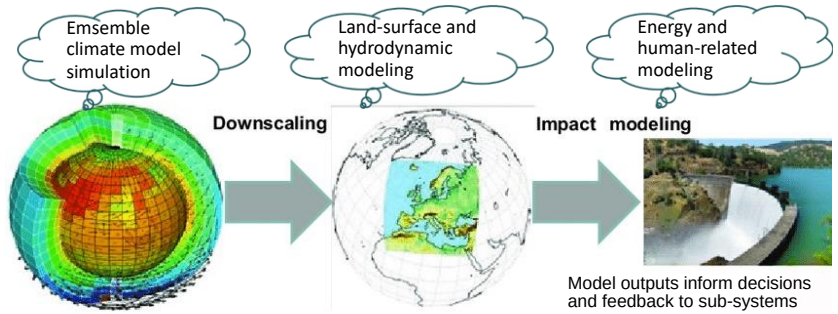


High-res data



- We adapted ORBIT for climate downscaling by replacing its embedding layers and prediction heads, while retaining its attention layers and variable aggregation module.





❖ ORBIT has potential to enhance high-resolution climate modeling and support critical decision-making.

AI foundation model has potential to transform Earth system modeling

- ORBIT has potential to advance Earth system modeling by leveraging diverse datasets and multi-model analysis.

Fine-tuning forecasts: ORBIT brings long-range weather prediction within reach

November 13, 2024

Researchers at Oak Ridge National Laboratory used the Frontier supercomputer to train the world's largest AI model for weather prediction, paving the way for hyperlocal, ultra-accurate forecasts. This achievement earned them a finalist nomination for the prestigious Gordon Bell Prize for Climate Modeling.

Gordon Bell Prize for Climate Modeling Finalist
Top Supercomputing Achievement Award

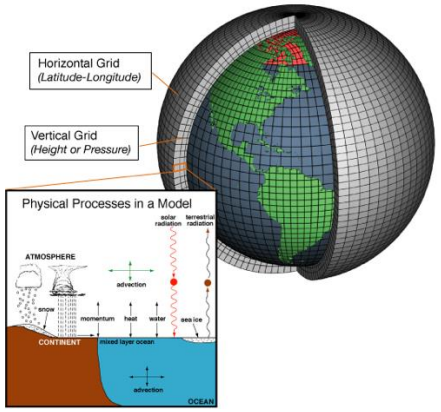
Oak Ridge National Laboratory receives honors in 2024 HPCwire Editors' Choice award

November 19, 2024

ORNL has been recognized in the 21st edition of the HPCwire Readers' and Editors' Choice Awards, presented at the 2024 International Conference for High Performance Computing, Networking, Storage and Analysis in Atlanta, Georgia.

Advancing Earth system prediction through data-model integration

Numerical Model



Challenges:

- High computational costs;
- Large, multi-uncertainty;
- Cannot integrate diverse data;

Our study:

- Emulators; multiscale modeling
- UQ; data assimilation
- Physics-ML hybrid modeling

Data-Driven ML Model

Challenges:

- Lack of explainability
- Lack of energy conservation
- Need trustworthiness

Our study:

- Interpretable ML
- Physics-informed ML
- Reliable ML; UQ for ML

AI Foundation Model (FM)

- A foundation model is a large NN trained on massive data at scale that can be adapted to broad applications
  - Integrate ”big data” and knowledge
  - Use for a wide range of modeling tasks
  - Save cost, effort, and energy
  - Improve performance, understanding, and generalizability

❖ ML at scale requires scalable models and advanced techniques tailored to the model, data, and HPC systems, to ensure efficient modeling and science advancement.