Accelerating Earth System Modeling by Leveraging a Standardized Coupling Framework for Hybrid Physics-ML Models

Ufuk Turuncoglu
NCAR/CGD/CSEG/ESMF

April 16, 2026

About me

- Senior Research Software Engineer at NCAR
  - Climate and Global Dynamics Laboratory (CGD)
  - CESM Software Engineering Group (CSEG)
  - Application developer - Earth System Modeling Framework (ESMF) Team
- Mostly working on NOAA projects
  - Development of Unified Forecast System (UFS)
    - Model coupling infrastructure
    - Technical lead of UFS Coastal, NCAR
- Science Background: Air-sea interaction (Mediterenian and Black Sea), dynamical downscaling (Med-CORDEX etc.), model development (RegESM), in situ visualization

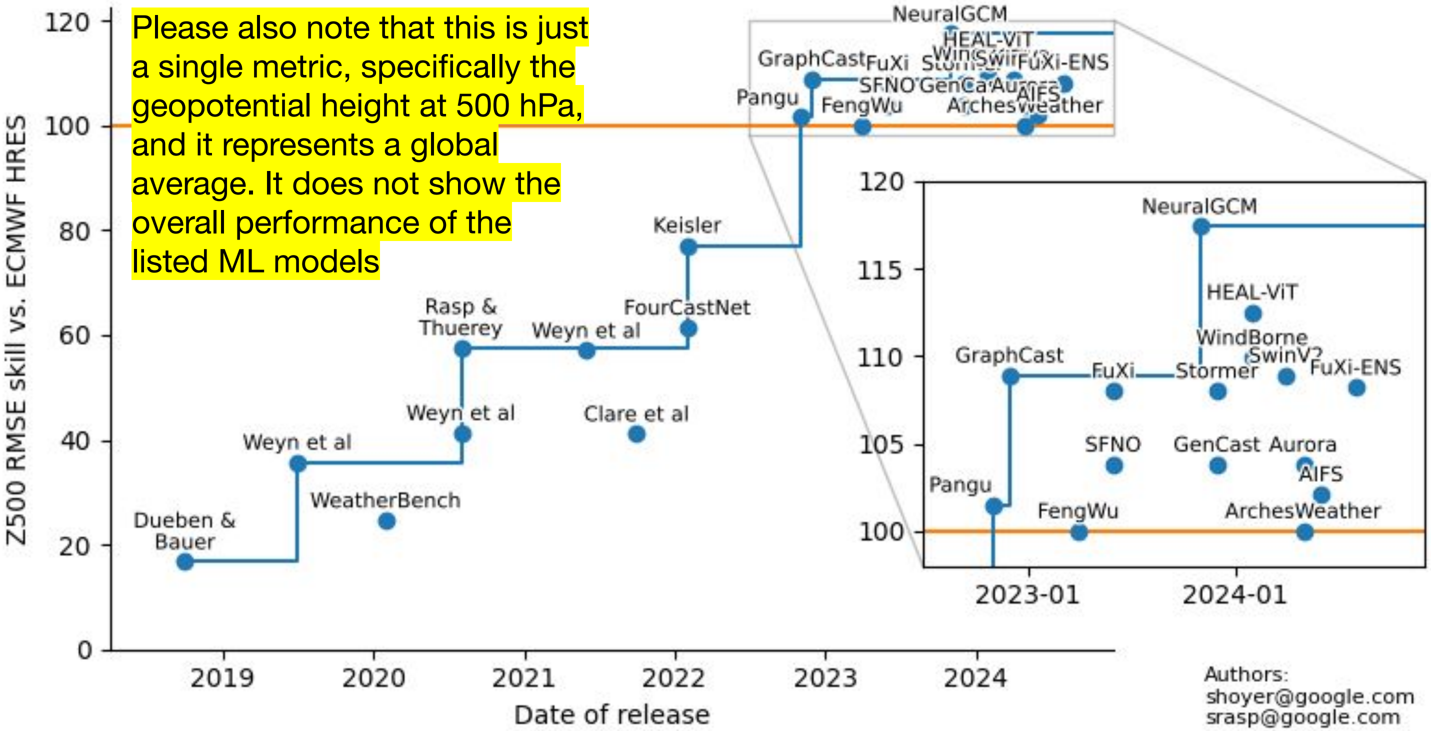
NSF | NCAR OPERATED BY UCAR | CGD Laboratory

Outline

- Some background and motivation
- Hybrid physics-ML coupled modeling approach
  - ESMF/NUOPC
  - New ESMF/NUOPC-complaint component - **GeoGate**
  - Its design and features
- The Aurora Foundational Model as an atmospheric model
- Realistic fully-coupled applications
  - Regional - *HAOCM-R* (Hybrid Atmosphere-Ocean Coupled Model)
  - Global - *HESM-S2S* (Hybrid Earth System Model for S2S Prediction)
- Q/A

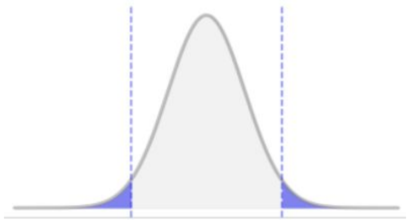
Evolution of ML Weather Prediction (MLWP) Models

Please also note that this is just a single metric, specifically the geopotential height at 500 hPa, and it represents a global average. It does not show the overall performance of the listed ML models

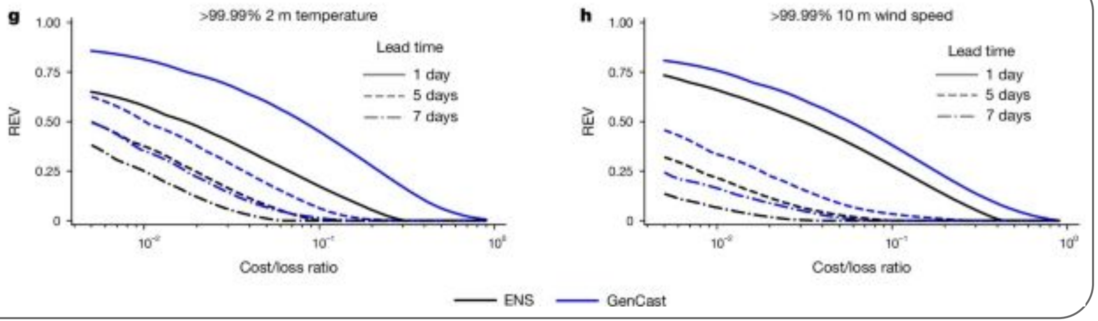


courtesy of Stephan Rasp

Emerging Challenges in ML-based models

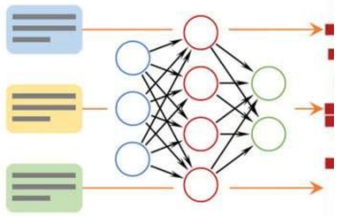


Unable to represent rare extreme events accurately



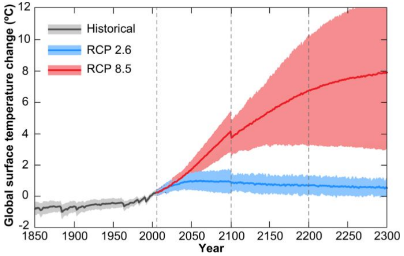
It is good in tracks, but not wind and pressure extremes and smoother precipitation (negative bias)

Price et al., 2024

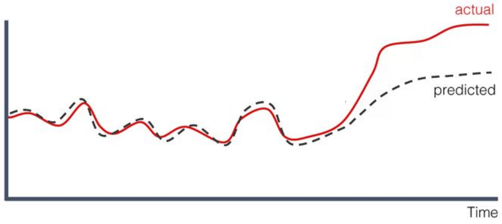


$$\frac{D\rho}{Dt} + \rho \nabla \cdot \boldsymbol{u} = 0$$

Insufficient or lack of physical constraints



Difficulties with extrapolating to new climate conditions



Numerical instabilities especially for longer time scales, higher temporal resolutions

Introducing additional components of the Earth System could help constrain AI/ML-based models and produce more physically realistic predictions.

NSF NCAR UCAR

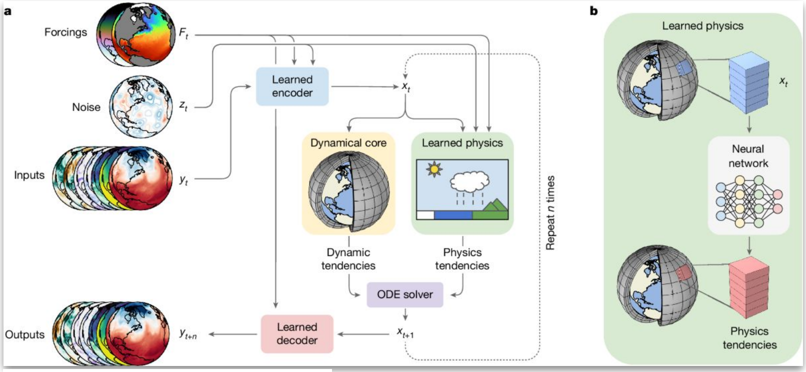
Hybrid Physics-ML Models?

- Its meaning varies significantly across different contexts
  - **ML corrected physics** (DA or pre- or post- processing)
    - Improving ICs, bias correction, and updating the model state
  - **ML-based subgrid scale parameterization** in traditional model
    - ML based physics component for subgrid parameterization
  - Completely **data-driven ML-based models**
    - Operational models that uses ML: NOAA's AIGFS, AIGEFS and, HGEFS
      - HGEFS combines physics-based ensembles with AI-based ones
  - Coupling **traditional physics-based models with ML models**

Hybrid Physics-ML Models?

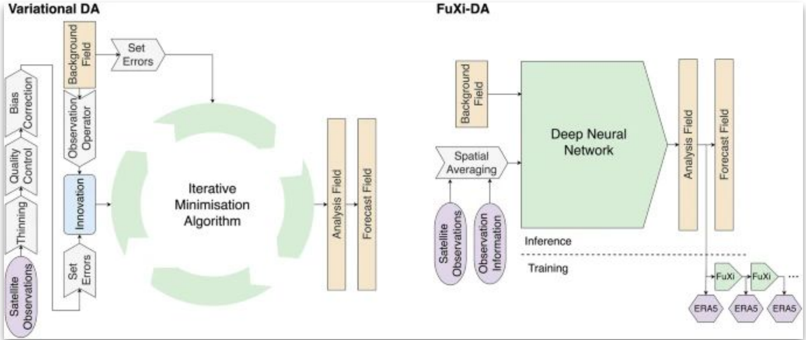
Kochkov et al., 2024: NeuralGCM

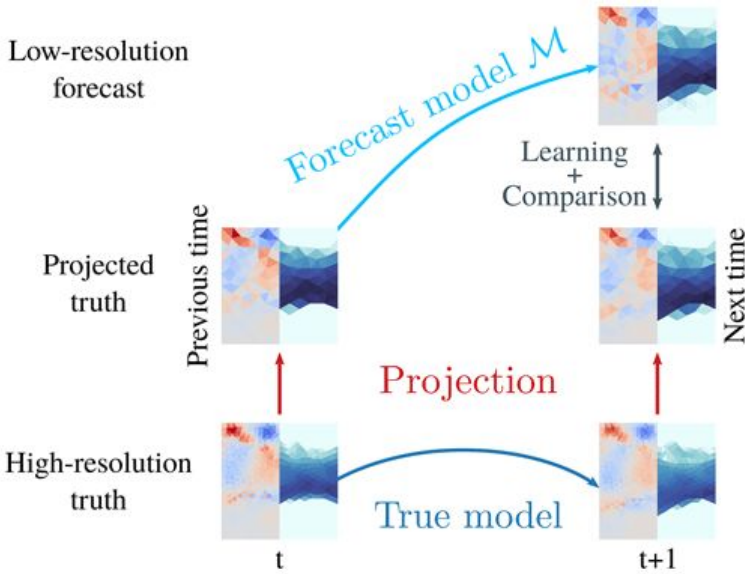
Traditional DYCORE coupled with AI/ML based physics



Xiaoze et al., 2025: FuXi-DA

DA framework for assimilating satellite observations





Finn et al., 2023: AI/ML based parametrization for sea-ice dynamics

Spencer et al., 2025: The Ai2 Climate Emulator coupled to a **slab ocean** accurately emulates temperature and precipitation sensitivity in a physics-based model

Our Motivation

- Adopting a hybrid modeling approach that combines ML-based with traditional physics-based models in Earth system science might leverage the strengths of both approaches
- Replacing any traditional physics-based model component with its equivalent might help to improve overall model accuracy, stability, and computational efficiency
    - Lead to more **physically constrained** coupled ML models
    - **Predictions for not only atmosphere but also other components** (i.e. ocean)
- Study the Earth as a system using hybrid physics-ML modeling systems
    - **Develop and test new ideas** in both physics and ML-based models
    - Opens a new way to perform controlled experiments → scientific innovations

ESMF/NUOPC

- The Earth System Modeling Framework (ESMF) is high-performance software infrastructure used in coupled Earth science applications.
- National Unified Operational Prediction Capability (NUOPC) is a software layer on top of ESMF that provides technical interoperability of model components so they can be shared across coupled systems.
- Focuses:
  - Interoperability
  - Reusability
  - Regridding
  - Performance

NCAR | COMMUNITY EARTH SYSTEM MODEL CESM

COAMPS

UFS

MOSSCO

GEOS

ICON
GETM

ESPC

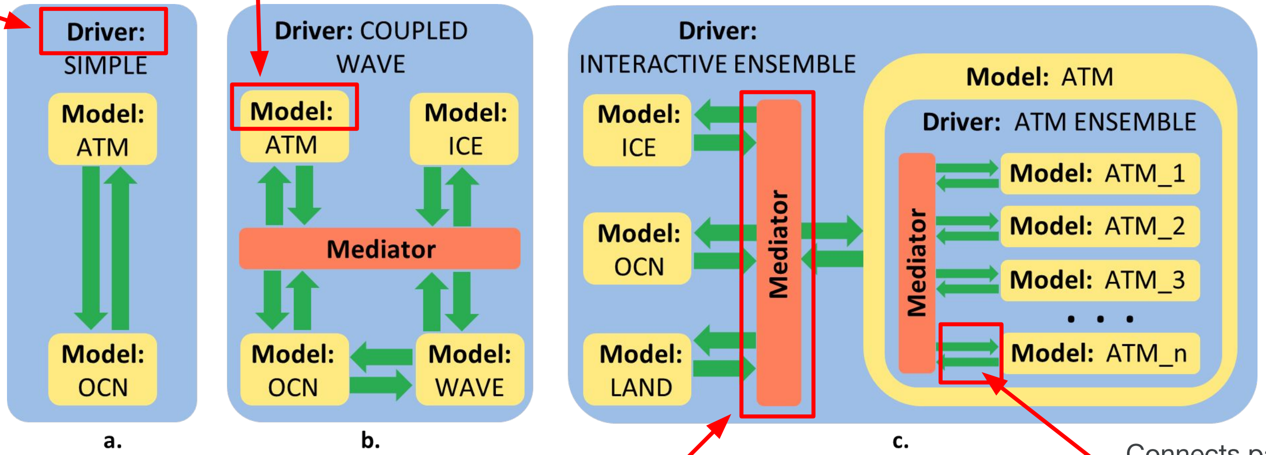
NorESM

ESMF/NUOPC

- The NUOPC layer includes building blocks to create new coupled applications

Typically implements a specific physical domain, e.g. atmosphere, ocean, sea ice, waves, etc. but not limited. **For example, CDEPS is also component**

Provides a harness for Models, Mediators, and Connectors, coordinating their initialization and driving them during the application time loop. Example: **E**arth **S**ystem **M**odel e**X**ecutable layer (ESMX)



Used for custom coupling code (flux calculations, averaging, etc.) between multiple Models. Example: The **C**ommunity **M**ediator for **E**arth **P**rediction **S**ystems (CMEPS)

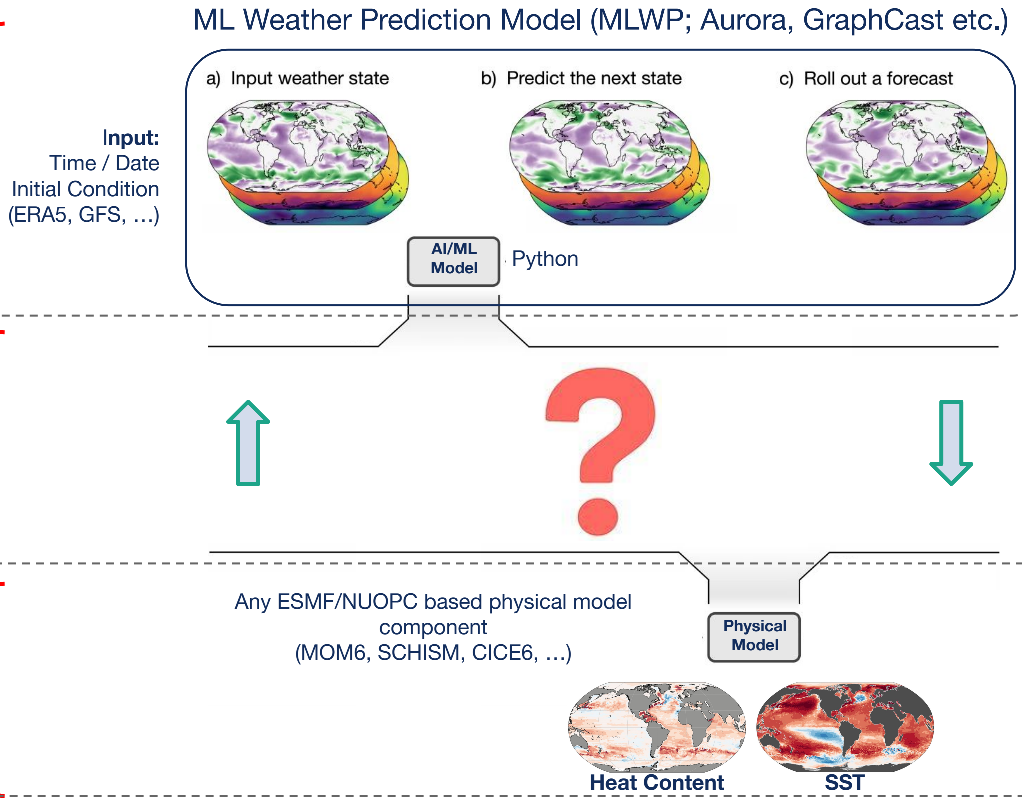
Connects pairs of components, e.g. Model to/from Model, or Model to/from Mediator, and executes simple transforms (i.e., regrid or redistribution).

More information

Hybrid Physics-ML Coupled Modeling System

Python
(PyTorch, …)
GPU enabled

Fortran/C/C++
(ESMF/NUOPC)



ML-based model component/s

Application specific vs. generic approach

Physics-based model component/s

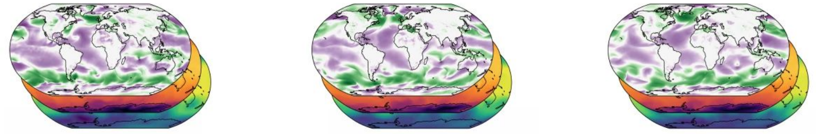
Example products from physical model (i.e ocean)

* It is application agnostic solution since other component/s will only see the generic ESMF/NUOPC component (GeoGate).

Bridging two worlds: GeoGate

ML Weather Prediction Model (MLWP; Aurora, GraphCast etc.)

a) Input weather state b) Predict the next state c) Roll out a forecast



ML Model Python

Input:
Time / Date
Initial Condition
(ERA5, GFS, …)

Running ML-based model

single forecast
or
ensemble

Two-way interaction:
The ML model might handle feedback from active model component such as surface state (fluxes, **SST**, …)

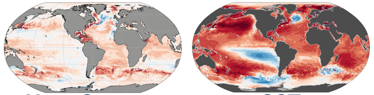
GEOGATE

If it is ensemble run, the results can be aggregated to drive physical component (ocean)

or multiple-instances can be triggered

Any ESMF/NUOPC based physical model component can be used
(MOM6, SCHISM, CICE6, …)

Physical Model

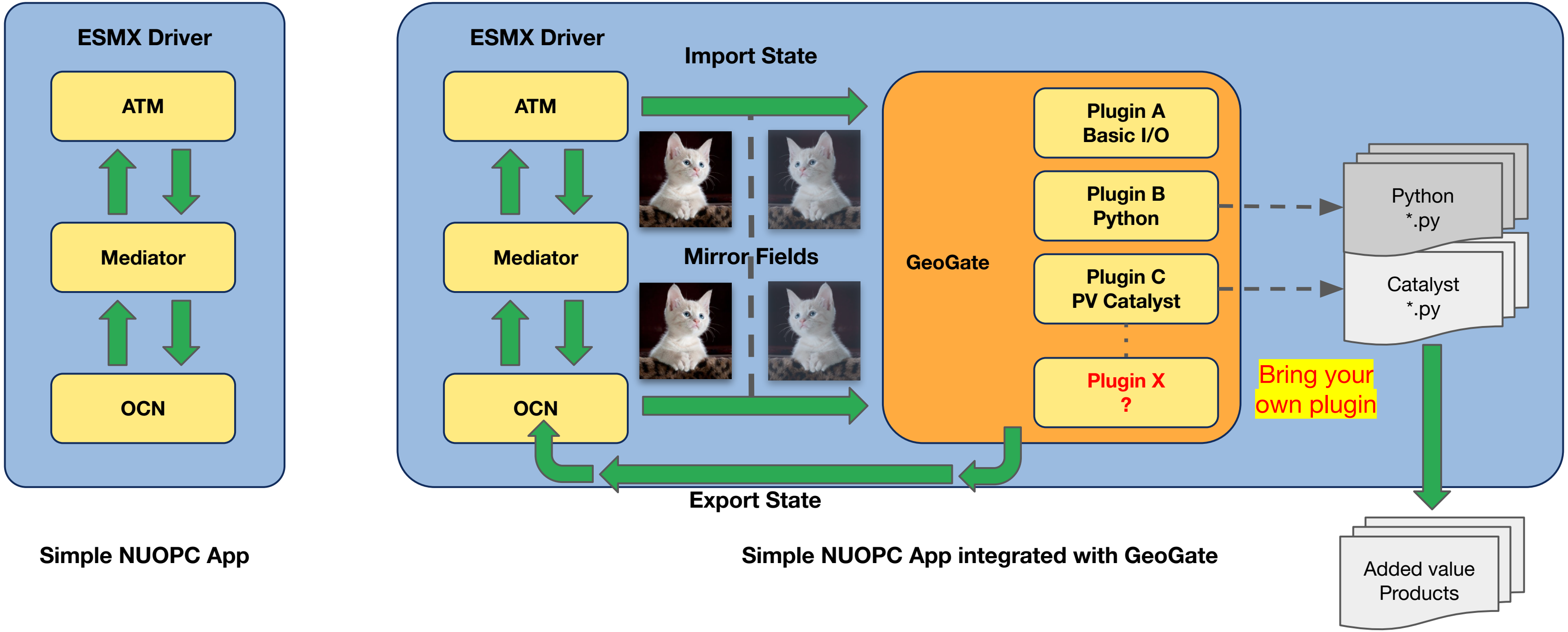


Heat Content **SST**

Example products from physical model (i.e ocean)

* It is application agnostic solution since other component/s will only see the generic ESMF/NUOPC component (GeoGate).

GeoGate: New ESMF/NUOPC-based Component



GeoGate is designed to become a generic & flexible ESMF/NUOPC-based **co-processing component**

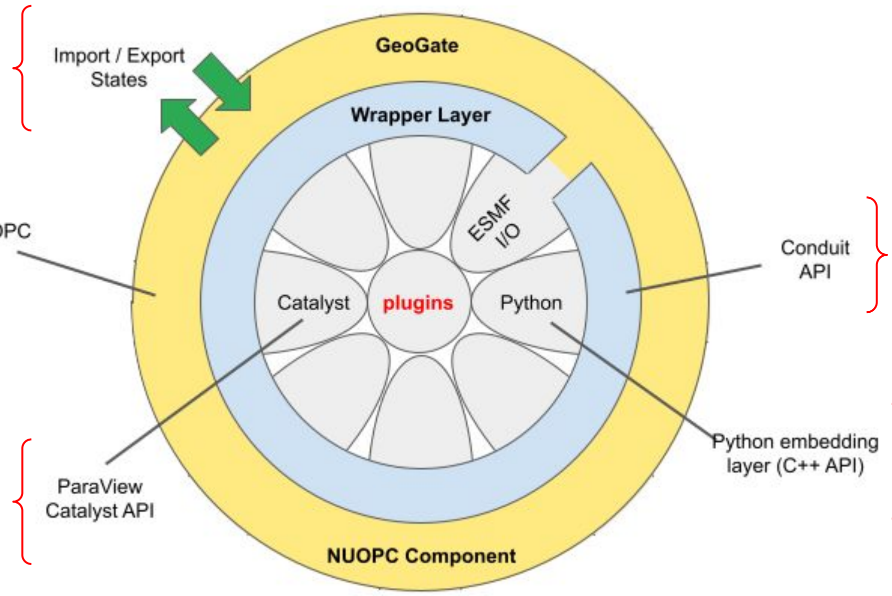
Specialized mediator like component with data processing and interaction - **plugin-based design**

GeoGate: Data Flow and Infrastructure

Interact with other components through its import/export states

NUOPC "cap" Fortran (mesh based)

In situ visualization and co-processing plugin



Wrapper leverages from Conduit to interact with Fortran, C, C++ and Python

Plugin for Python interaction

Conduit library passes data among layers in different programming languages (Fortran, C/C++, Python)

Zero copy between NUOPC "cap" and wrapper layer - no additional memory pressure

GeoGate: Use Cases

SUPPORT MODEL DEVELOPMENT ACTIVITIES	IN SITU VISUALIZATION	CO-PROCESSING	INTERACT WITH AI/ML MODELS	LANGUAGE INTEROPERABILITY
Support existing RT framework A set of diagnostics, plots and even added value products can be generated on-the-fly by running RT/s	Visualize data flowing from multi-component earth system model while it's being generated by a simulation It also supports live interaction and steering	Process data in highest possible temporal resolution while it's being generated Complex triggers can be used when certain threshold (i.e. entropy) is meet	Enable two-way interaction with AI/ML tools Training models and monitoring training process, coupling AI/ML with physical model components	Extensible plugin based design allows to interact with other programming languages with minimal effort Data processing workflows or model components written in Python, Julia etc.

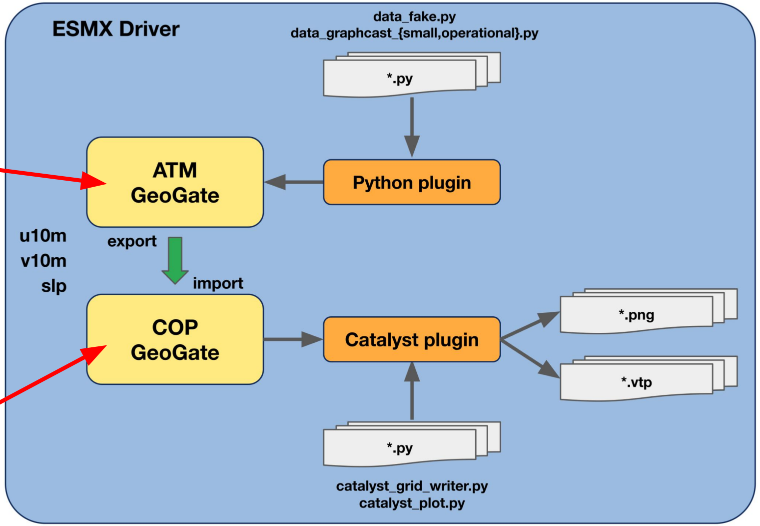
It can be extended through new plugins (i.e. interact with Julia, C/C++ based model components)

Current plugins could be configured to support all cases above from visualization to AI/ML interaction

Main Philosophy: Simple → Complex

- Two component prototype for development / testing

As a Producer
It passes information generated by Python



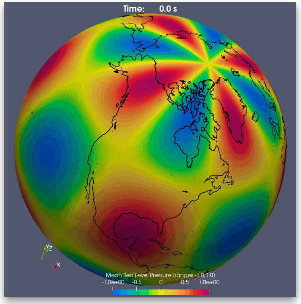
As a Consumer
The received information flowing from other component/s is processed by [ParaView Catalyst](#) (or Python)

Multi instance of GeoGate component can be used to bring more ML component or create complex configurations

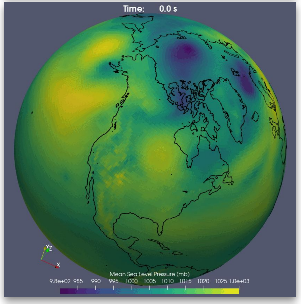
Flexibility
Same configuration with three different data source

NVIDIA's [Earth2studio](#)

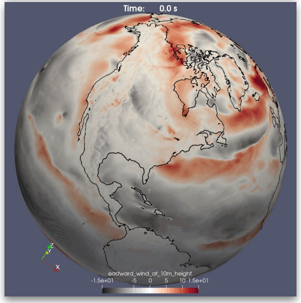
Custom Analytic data 0.25 deg ERA5 grid



Graphcast Small 1 deg



Graphcast Operational 0.25 deg



Difficulties to Use ML-based models for coupling

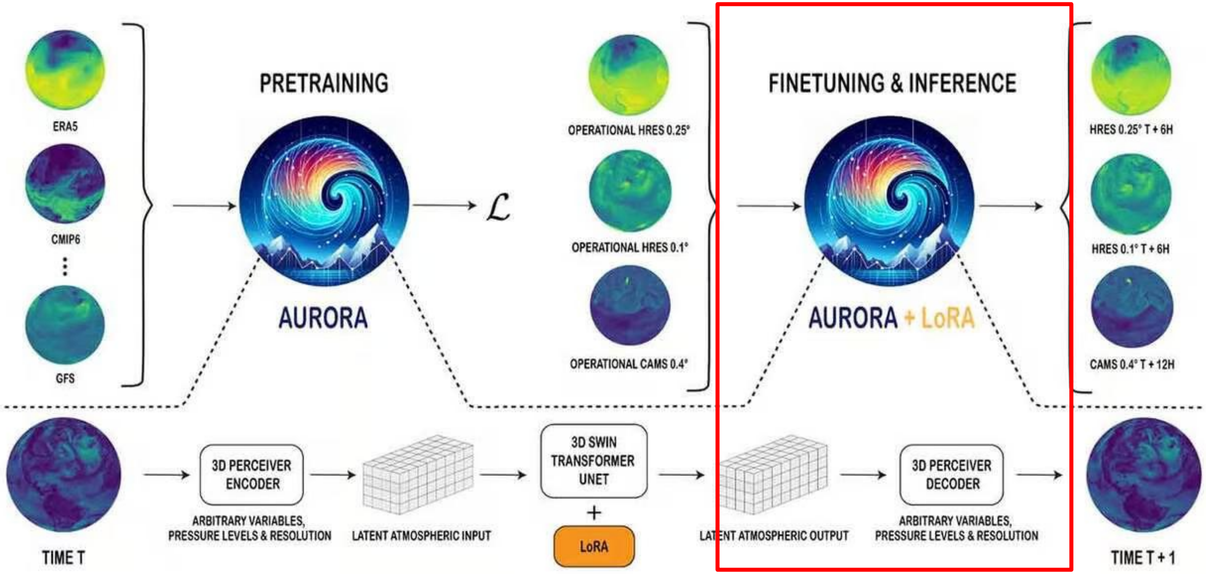
- Most of the ML-based model is only predicting a small set of variables
    - Basic surface variables, atmospheric state through the column (pressure levels)
    - Lack of all the variables required for coupling
        - humidity, radiation and flux components
- They are running generally in a coarse resolution in space and time
    - 0.25 deg. resolution
    - 6 or 12-hourly intervals
- Mostly designed to predict only atmospheric state
    - Other Earth system model components → wave, ocean, river routing, sea-ice

Bringing Required Surface Variables to ML-based Model

- The ML model needs to be flexible enough to adapt itself to provide more variable with acceptable error range to allow to drive traditional physics based models
  - If the predictions are not physically consistent that it could create stability issues in the underlying physical model
    - Nice way to test the stability and physical consistency of the ML-based models → standardized coupled configurations for benchmark
- Used ML-based model: **Microsoft's Aurora Foundational Model**
  - Relatively easy to understand and modify
  - Initially trained for atmosphere
    - fine tune + transfer learning → use in wave and chemistry or add variables
  - Also able to get support (Wessel P. Bruinsma @ The Alan Turing Institute)

The Aurora Foundational Model

- Aurora is a foundation model for the Earth system



Use to add missing variables

We used 0.25° resolution pretrained version in this study and fine-tuned to include new surface variables for coupling

2012-2017 training with ERA5 6h data

More information about available models - link

Main paper

Performance of the Aurora Model - w/ New Variables

- Testing period is selected as 2022 (5-day prediction for entire year, every 6h)

Included to create two-way interaction with physics based ocean model

More variables can be added like surface flux components etc.

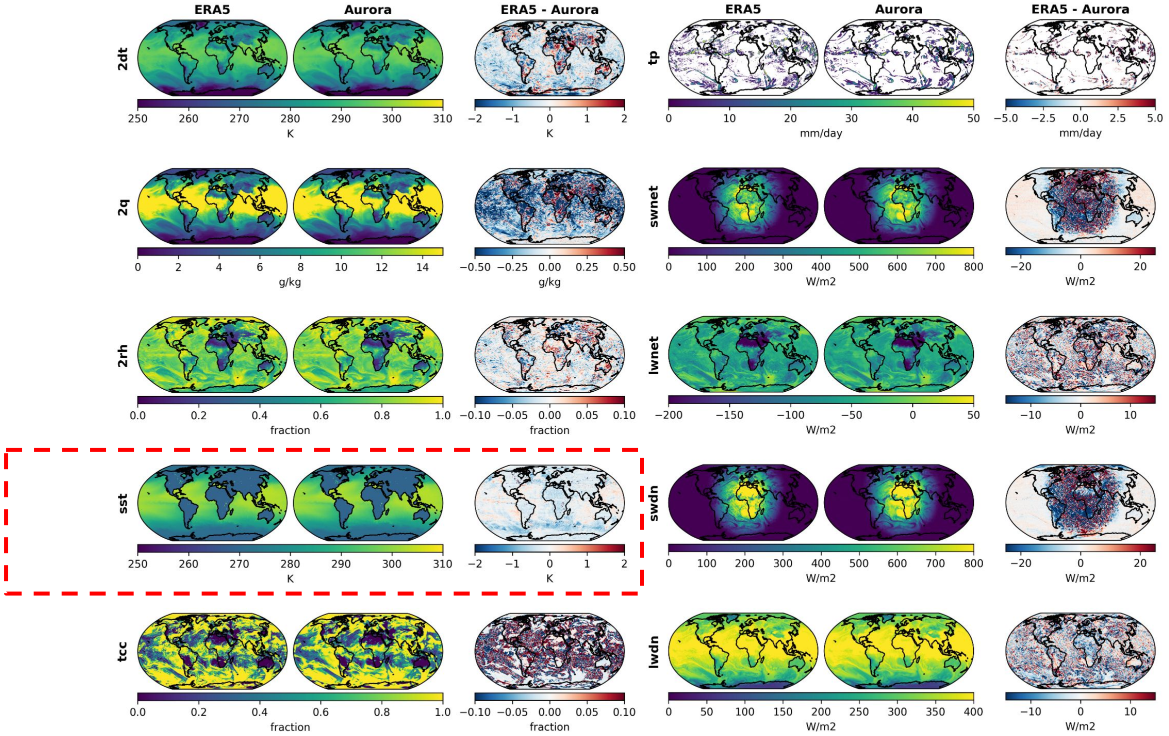
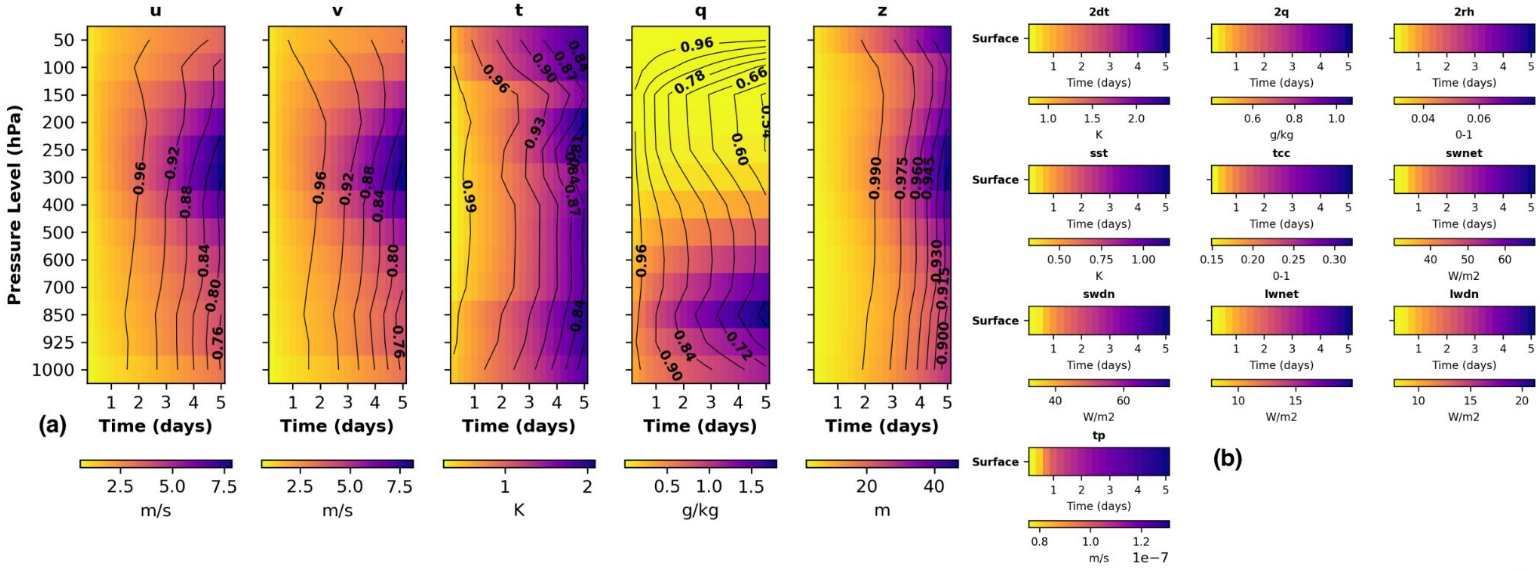


Fig: Selected surface variables (relative humidity at 2 meters, sea surface temperature, total precipitation, net and downward component of shortwave and longwave radiation) of their difference from ERA5 dataset for 2022-08-21 00Z and 12h lead time.

We are also performed test for entire 2021

Performance of the Aurora Model - w/ New Variables

- Testing period is selected as 2022 (5-day prediction for entire year, every 6h)

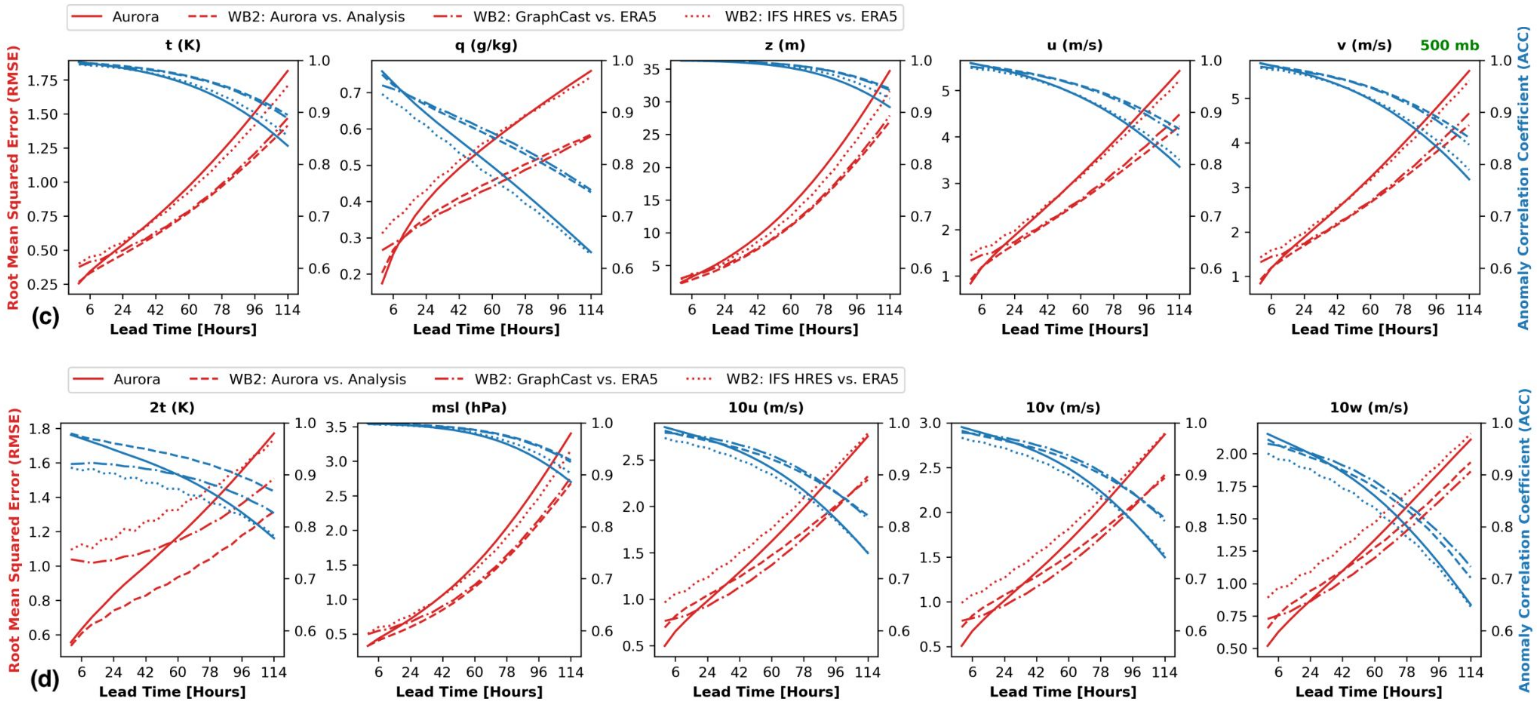


In general the performance of the model is comparable with the ones found in **WeatherBench 2 database**

Also note that we did not follow two step fine-tuning approach (additional training with HRES-T0) due to lack of data

Performance of the Aurora Model - w/ New Variables

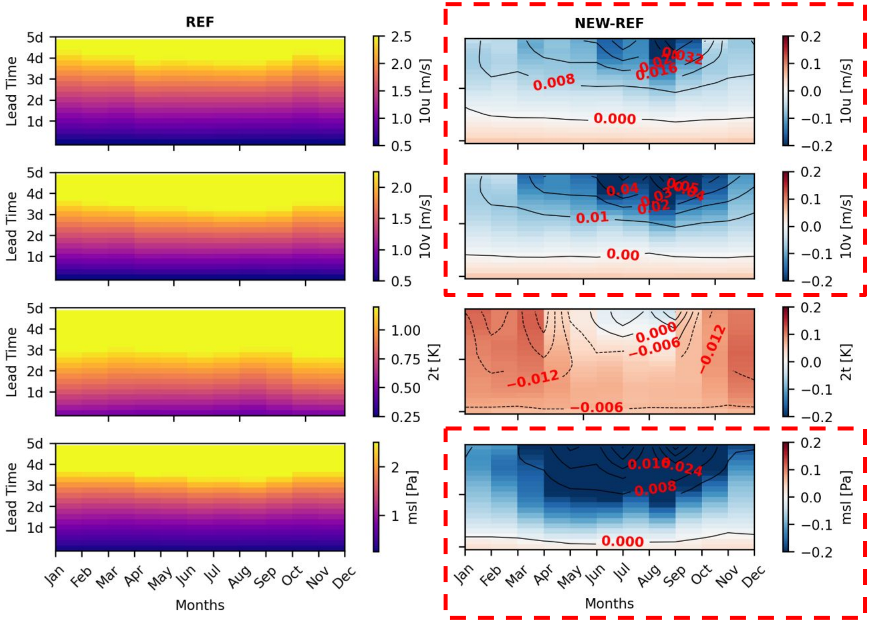
- Comparison with WB2 database (GraphCast vs. ERA5 and IFS HRES vs. ERA5)



Note that we only trained between 2012-2017 and not to use IFS HRES T0 for extra fine tuning

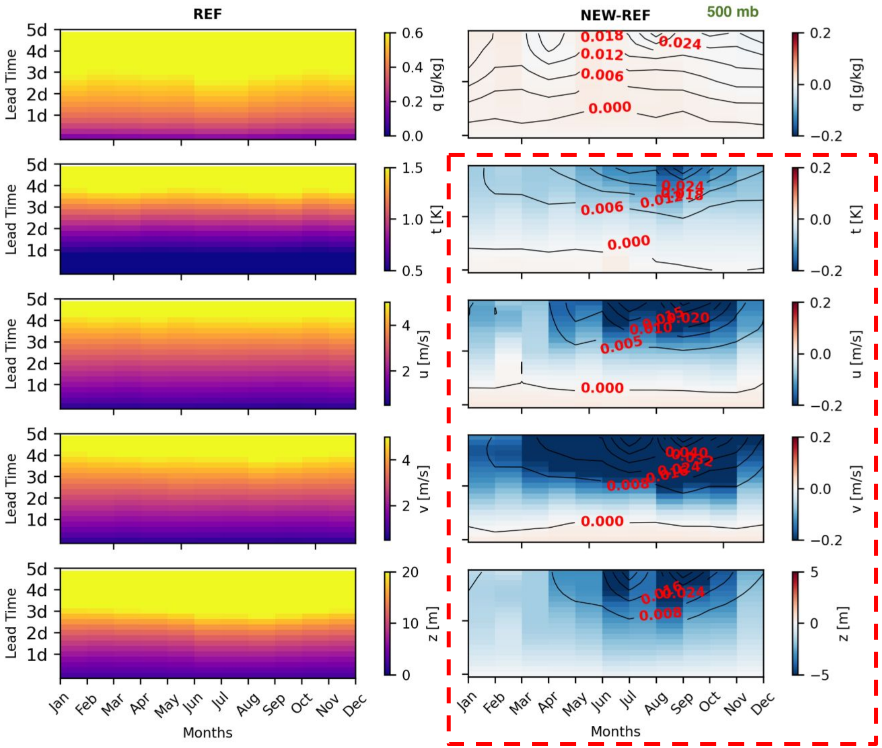
Added Value: w and w/o Newly Introduced Variables

- Comparison with no extra variable case



Model is trained without any new variable (REF) and results are compared with the model that has new variables (NEW).

Reduced bias during warm season

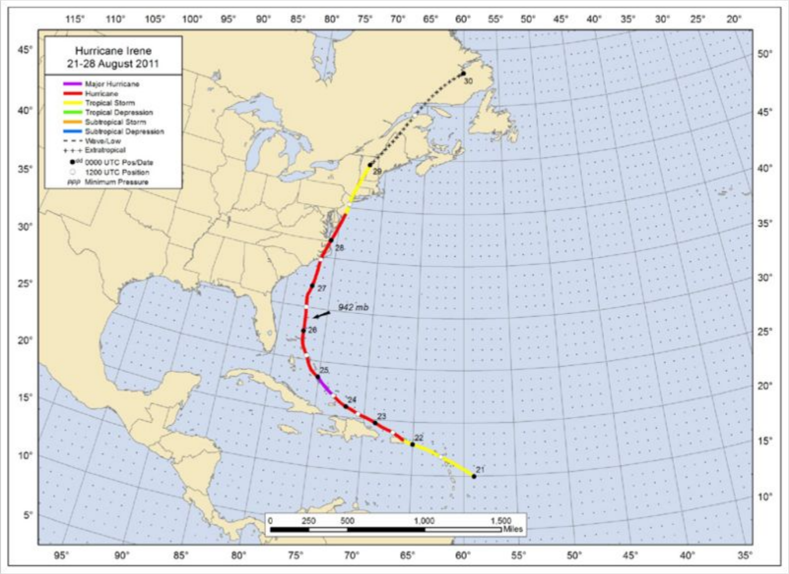


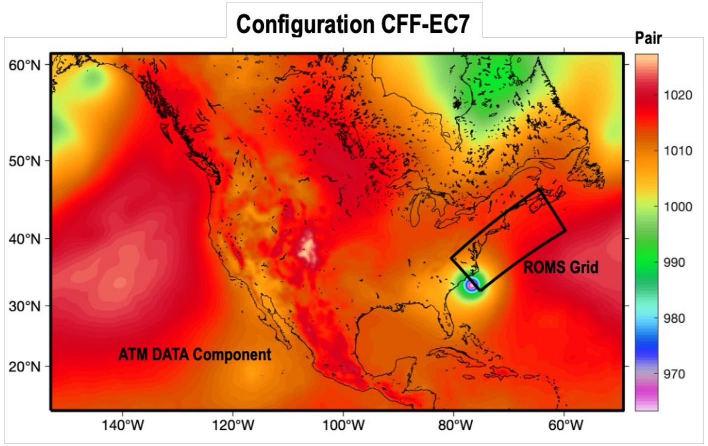
Summary: The Aurora Model Performance

- The fine-tuned Aurora Model runs as expected
  - RMSE and correlation values are comparable with WB2 database (see deterministic scores)
- What is new?
  - **New variables** are added to allow realistic coupling
  - **Ready to force underlying physics-based model (w/ help of new variables)**
    - Still lack of physical constraints
      - negative shortwave, cloud cover or relative humidity greater than 1
  - **Ready to be used as a component** in existing Earth system models
    - As a part of the overall ESMF/NUOPC based coupled system

Realistic Application (1): Regional + Hurricane Irene

- Two component coupled system → Data component (CDEPS or AI/ML) + ROMS
- Hurricane Irene configuration - August 27, 2021 -> 42 hours





ROMS Grid: 7km (40 L)

Mixing: GSL (k-kl)

Forcing: ERA5

Time step: 60 s

Hurricane Irene: The Aurora Model Performance

- Hurricane Track

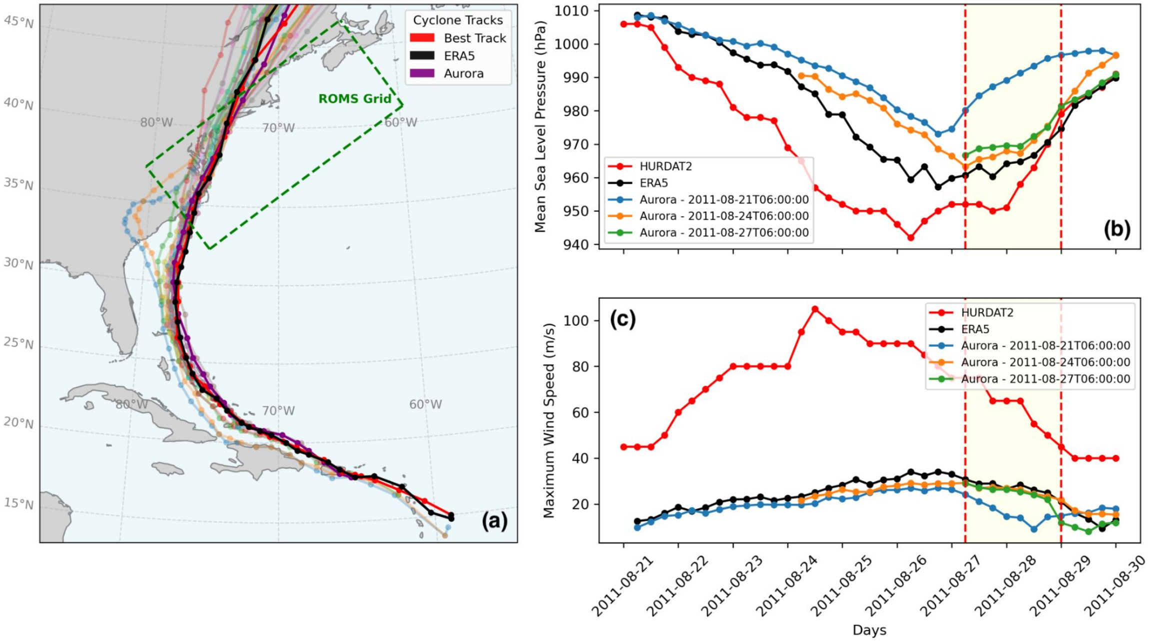


Fig X. Evaluating Hurricane Irene prediction from the Aurora model. (a) the list of storm tracks predicted by the Aurora model, best track from HURDAT2 and ERA5, (b) Daily average time series of mean sea level pressure from the Aurora model initialized from different dates, ERA5 and HURDAT2, and (3) Daily average time series of maximum wind speed from the Aurora model initialized from different dates, ERA5 and HURDAT2

Relatively good in MSLP predictions but not good in maximum wind speed (incl. raw ERA5). This is also seen in other ML-based models.

It is consistent with the results of Price et al., 2024

Hurricane Irene: The Aurora Model Performance

- Total Precipitation

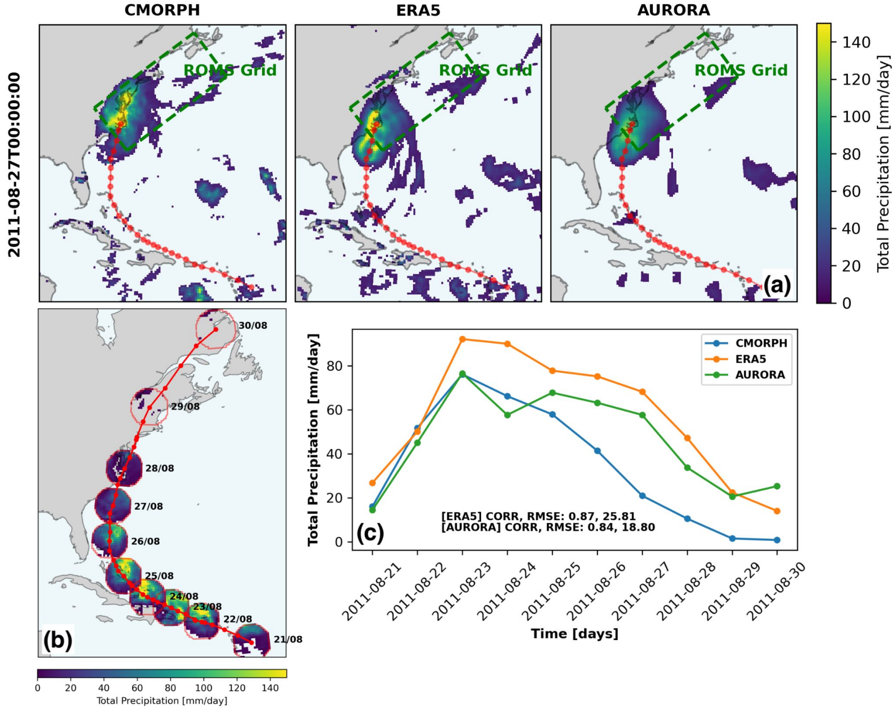
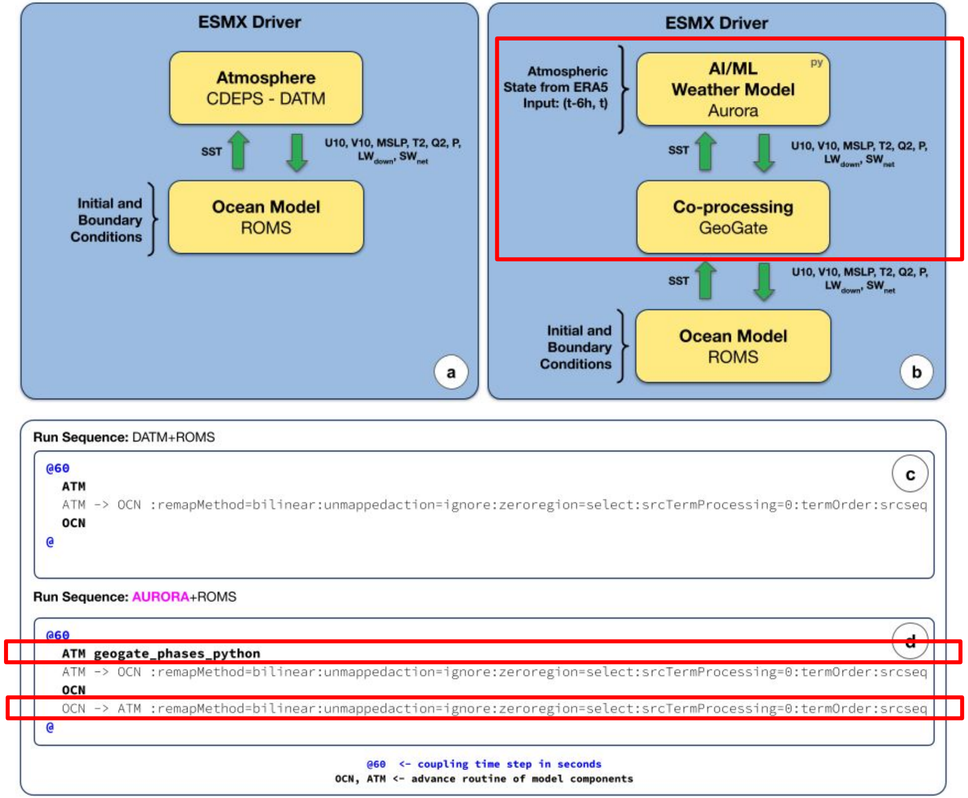


Fig X. Comparison of total precipitation produced by the Aurora model (t+6h) and available observation (CMORPH) and reanalysis dataset (ERA5) used to drive the model. **(a)** The spatial distribution of total precipitation over the region at 2011-08-27 just before Hurricane Irane entered ROMS ocean model domain, **(b)** CMORPH observational total precipitation along the best track over circular regions in a 250 km radius. Which is used to create time series, and **(c)** daily time series of observation, reanalysis dataset and the Aurora predictions along the strom track shown in (b).

The Aurora model is able to represent the location of the maximum precipitation with some degree of negative bias in intensity.

$\text{Aurora}_{\text{RMSE}} < \text{ERA5}_{\text{RMSE}}$ @ 1st lead time

HAOCM-R: The Aurora Model + **GeoGate** + ROMS



Replaces CDEPS - DATM

The Aurora model require t-6h and t to predict t+6h and t+12h

Runs GeoGate Python plugin

Defined as a NUOPC phase

NUOPC connector used to interact with the Aurora Model

Experiments

- A set of experiments are conducted with *HAOCM-R* to measure the performance of the Aurora coupled configurations

#	Forcing Component	Input	Temporal Interpolation	Coupling Type	Feedback
R1	CDEPS	ERA5 (6h)	coszen*	one-way	No
R2	CDEPS	ERA5 (1h)	coszen*	one-way	No
R3	CDEPS	ERA5 (6h)	linear	one-way	No
R4	CDEPS	ERA5 (1h)	linear	one-way	No
R5	Aurora	ERA5 (6h)	linear	one-way	No
R6	Aurora	ERA5 (6h)	linear	two-way	Yes

Directly comparable configurations

R3 and R5

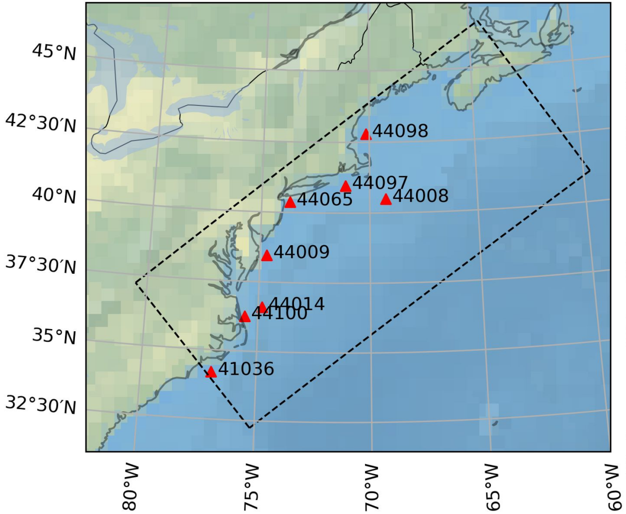
Two-way coupled configuration

The Aurora model updates its SST (over the coupling region) provided by ROMS and uses ERA5 SST in the rest

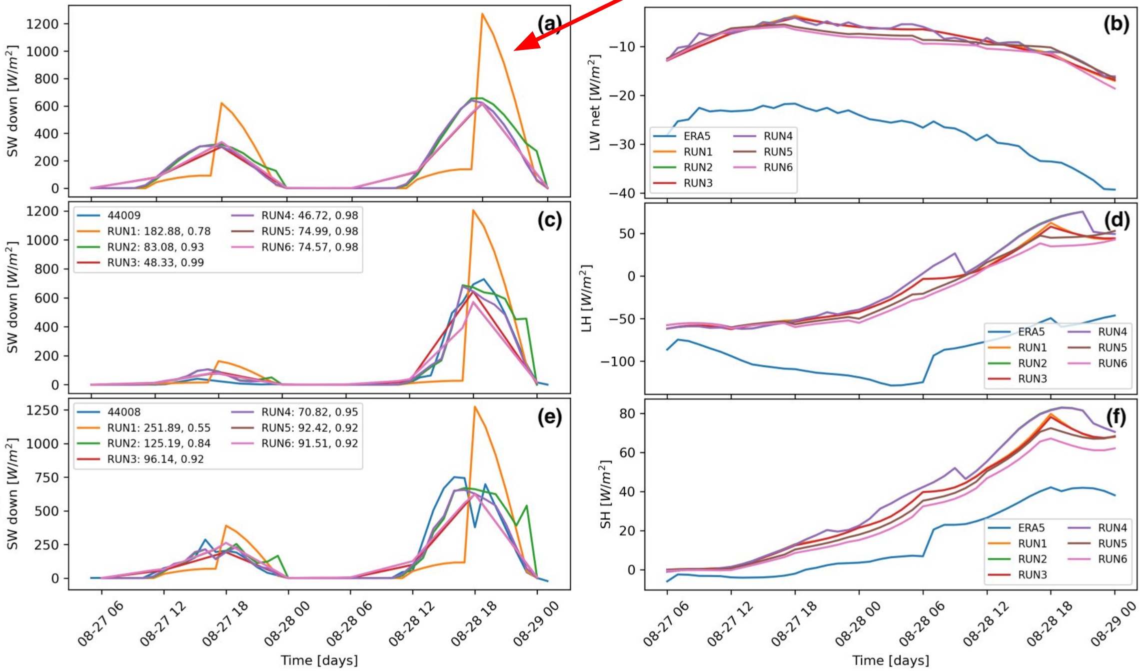
Results

- Flux components are compared with the data from buoys

Unrealistic peak in swrad
RUN1: 6h ERA5+coszen (CDEPS)

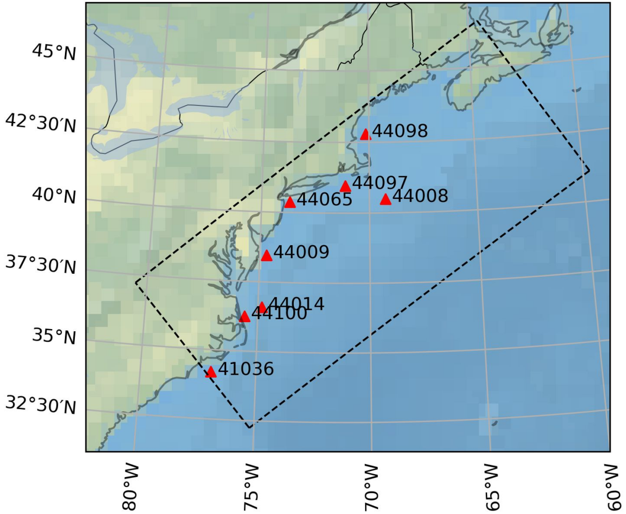


National Data Buoy Center (NDBC) buoys
41036, 44009, 44065, 44100, 44014, 44097, 44008, and 44098

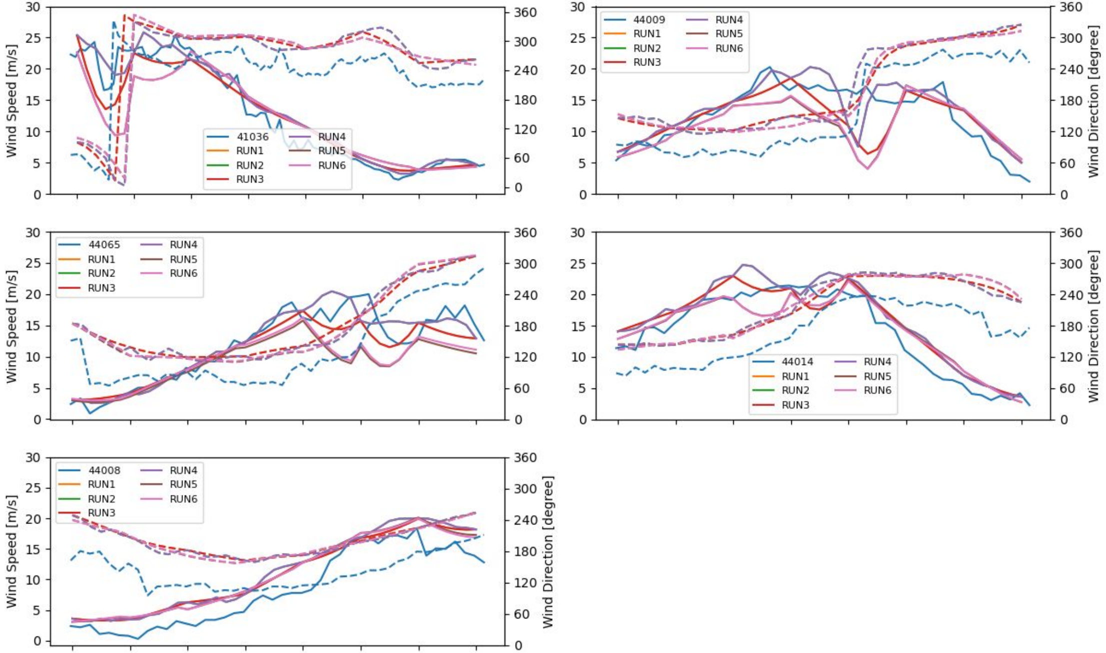


Results

- Wind components (speed and direction) are compared with the data from buoys

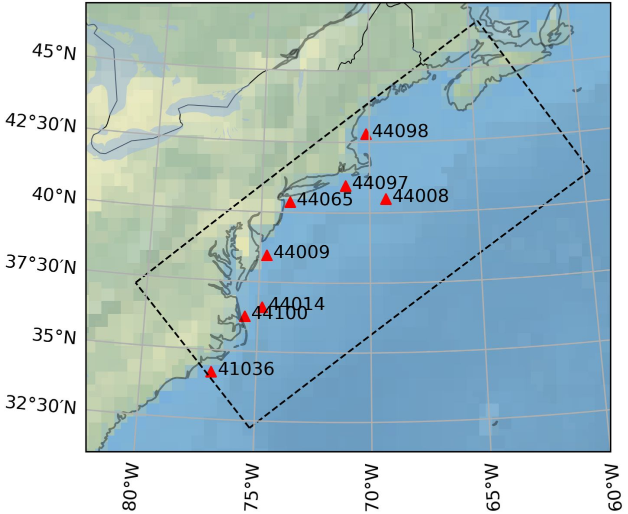


National Data Buoy Center (NDBC) buoys 41036, 44009, 44065, 44100, 44014, 44097, 44008, and 44098

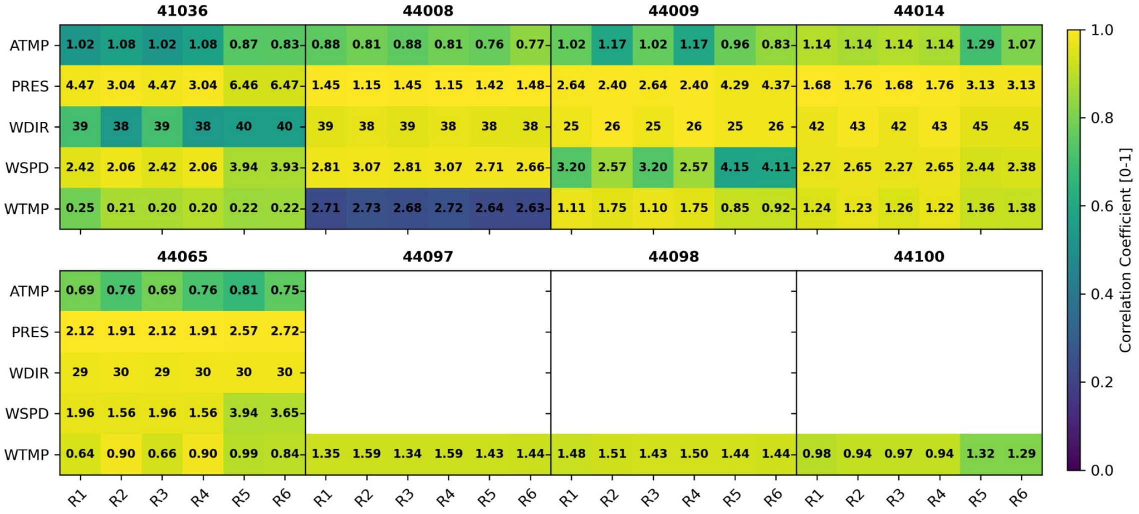


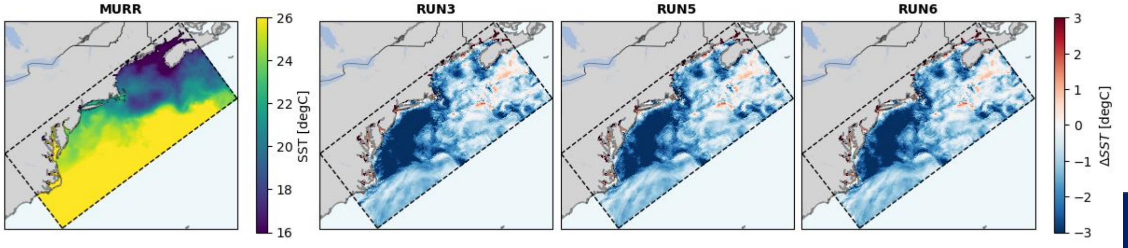
Results

- Air Temperature, Pressure, SST are compared with the data from buoys + obs



National Data Buoy Center (NDBC) buoys 41036, 44009, 44065, 44100, 44014, 44097, 44008, and 44098





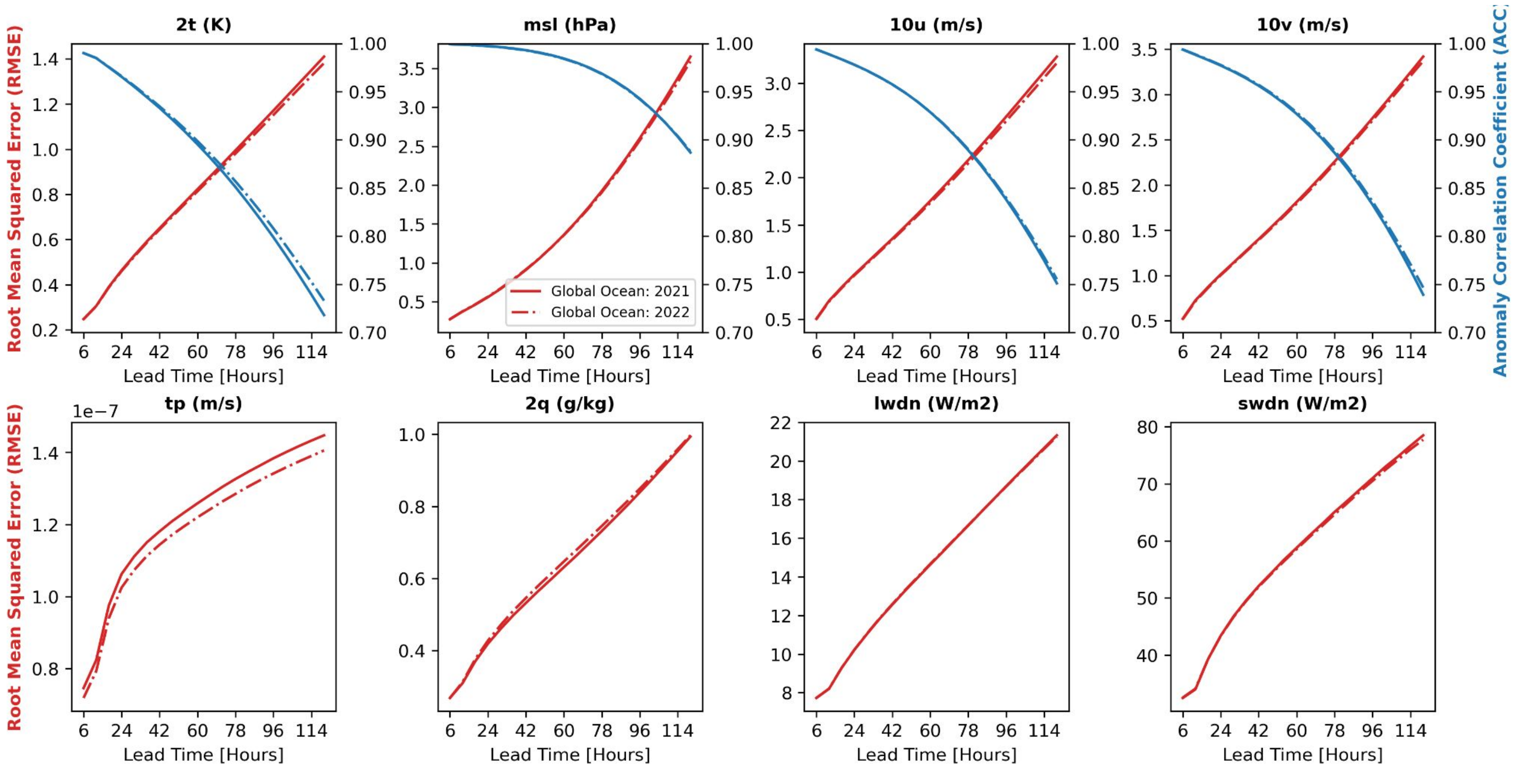
HAOCM-R Coupling Summary

- The Aurora Model is able to drive underlying ocean model realistically
  - The results show comparable performance with ERA5 driven configuration
  - It demonstrates lower RMSE and higher correlation than ERA5 in some cases
  - No significant sensitivity to SST (feedback) in Aurora model
    - The simulation duration short (42 h)
    - The coupling process just updating SST over active coupling region and rest of the SST input is coming from ERA5
      - Aurora model is still running globally
    - Coarse temporal interval (t-6h, t → t+6h, t+12h)
    - Coupling time step used in here is 60 s. The linear interpolation is applied between t+6h and t+12h to match with coupling interval

Realistic Application (2): Global S2S Setup

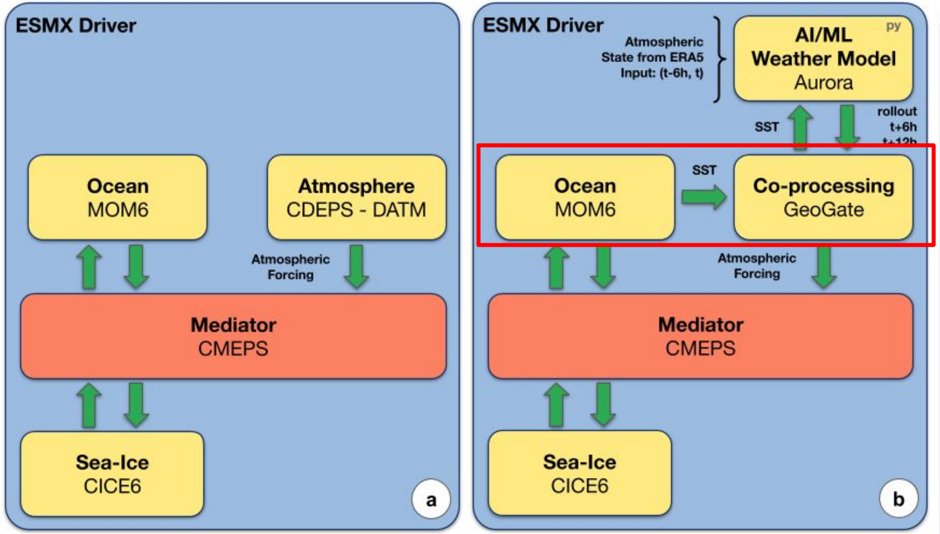
- Four component coupled system → Data component (CDEPS or AI/ML) + GeoGate + MOM6 + CICE6 + CMEPS
- Based on UFS WM regression test (RT): *datm_cdeps_mx025_cfsr*
  - The initial conditions for MOM6 and CICE6 are updated for selected year (thanks to Denise Worthen and Nick Szapiro from NOAA/EMC)
  - The data atmosphere and the Aurora coupled configurations
    - 35-days long runs
    - Starting from Jan, Apr, Jul, Oct 2021 (to cover all seasons)
  - Performance of the coupled modeling system and conducted experiments (data atmosphere, the Aurora with one-way and two-way coupling) are compared against the available observations (sea-ice, altimetry etc.) and ERA5 dataset

Performance of the Aurora Model: 2021 vs. 2022



Note that we are comparing only over ocean since we would like to drive MOM6 and CICE6

HESM-S2S: The Aurora + **GeoGate** + MOM6 + CICE6 + CMEPS



Interacts with the global ocean model (MOM6) via NUOPC connectors

Run Sequence: DATM+MED+MOM6+CICE6

```
@3600
   MED med_phases_prep_ocn_avg
   MED -> OCN :remapMethod=redist
   OCN
   @900
     MED med_phases_prep_ice
     MED -> ICE :remapMethod=redist
     ATM
     ICE
     ATM -> MED :remapMethod=redist
     MED med_phases_post_atm
     ICE -> MED :remapMethod=redist
     MED med_phases_post_ice
     MED med_phases_ocnalb_run
     MED med_phases_aofluxes_run
     MED med_phases_prep_ocn_accum
   @
   OCN -> MED :remapMethod=redist
   MED med_phases_post_ocn
@
```

c

Run Sequence: AURORA+MED+MOM6+CICE6

```
@3600
   MED med_phases_prep_ocn_avg
   MED -> OCN :remapMethod=redist
   OCN
   @900
     MED med_phases_prep_atm
     OCN -> ATM :remapMethod=redist
     ATM geogate_phases_python
     ATM -> MED :remapMethod=redist
     MED med_phases_post_atm
     MED med_phases_prep_ice
     MED -> ICE :remapMethod=redist
     ICE
     ICE -> MED :remapMethod=redist
     MED med_phases_post_ice
     MED med_phases_ocnalb_run
     MED med_phases_aofluxes_run
     MED med_phases_prep_ocn_accum
   @
   OCN -> MED :remapMethod=redist
   MED med_phases_post_ocn
@
```

d

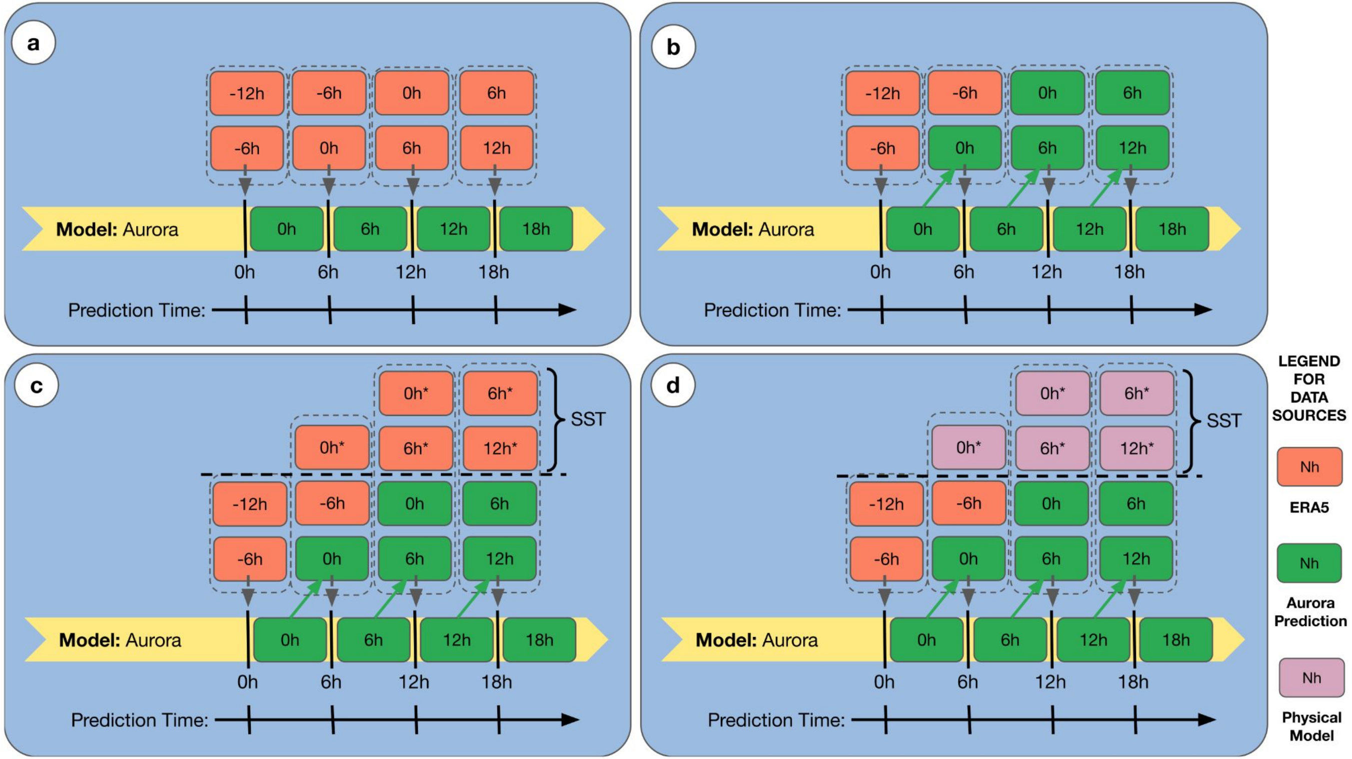
```
@3600 <- slow coupling time step in seconds
@900  <- fast coupling time step in seconds
OCN, ATM, ICE <- advance routine of model components
MED med_phases_aofluxes_run <- mediator atmosphere-ocean fluxes calculation
```

The newly developed component can be used with slow/fast coupling loops

Inference Strategies

Forced prediction ERA5

Perfect prediction



Perfect prediction constrained with ERA5 SST

Perfect prediction constrained with MOM6 SST

Experiments

- A set of experiments are conducted with *HESM-S2S* to measure performance of the Aurora coupled configurations

Table 3: *List of performed sensitivity simulations. Each simulation is performed in four seasons and the model initialized at month January (1), April (4), July (7) and October (10). Each run is initialized from the first day of the month and runs 35-days in forward.*

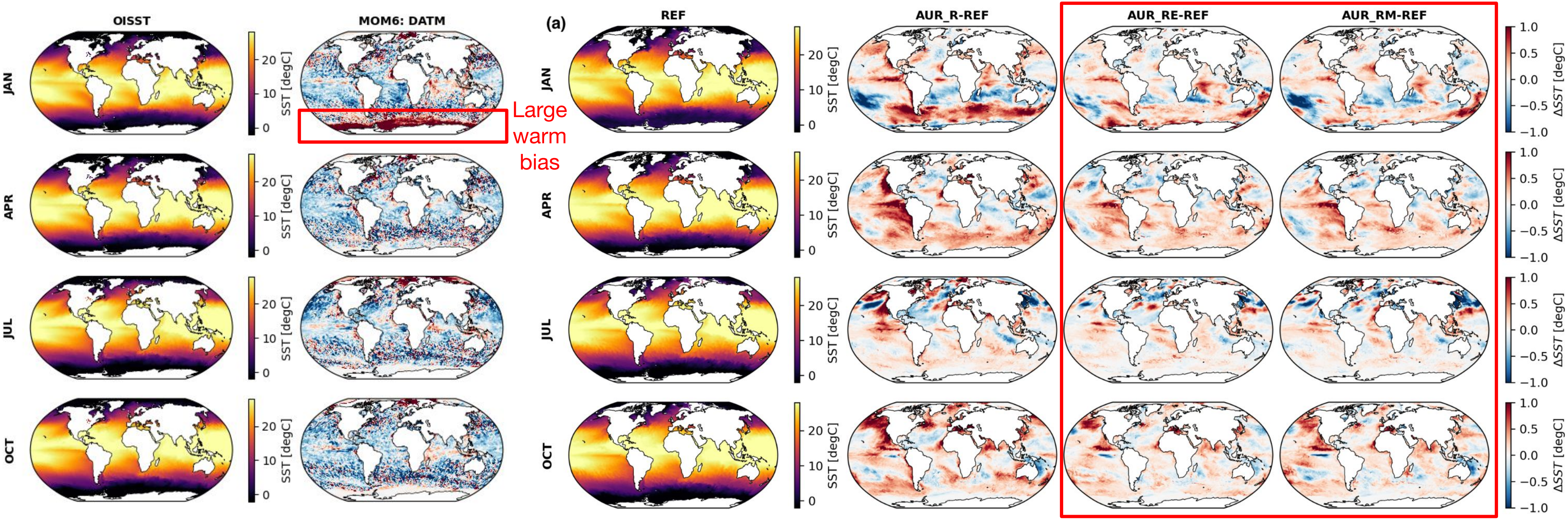
#	Forcing Component	Input	Temporal Interpolation	Coupling Type	Feedback
REF	CDEPS	ERA5	linear	one-way	No
AUR_F	Aurora	Forced ERA5 (Fig. 15a)	linear	one-way	No
AUR_R	Aurora	Rollout (Fig. 15b)	linear	one-way	No
AUR_RE	Aurora	Rollout + ERA5 SST (Fig. 15c)	linear	one-way	No
AUR_RM	Aurora	Rollout + MOM6 SST (Fig. 15d)	linear	two-way	Yes

- The used coupled configuration is based on one of the existing UFS WM RT
  - *datm_cdeps_mx025_cfsr* (CDEPS+CMEPS+MOM6+CICE6)

Sea Surface Temperature from MOM6

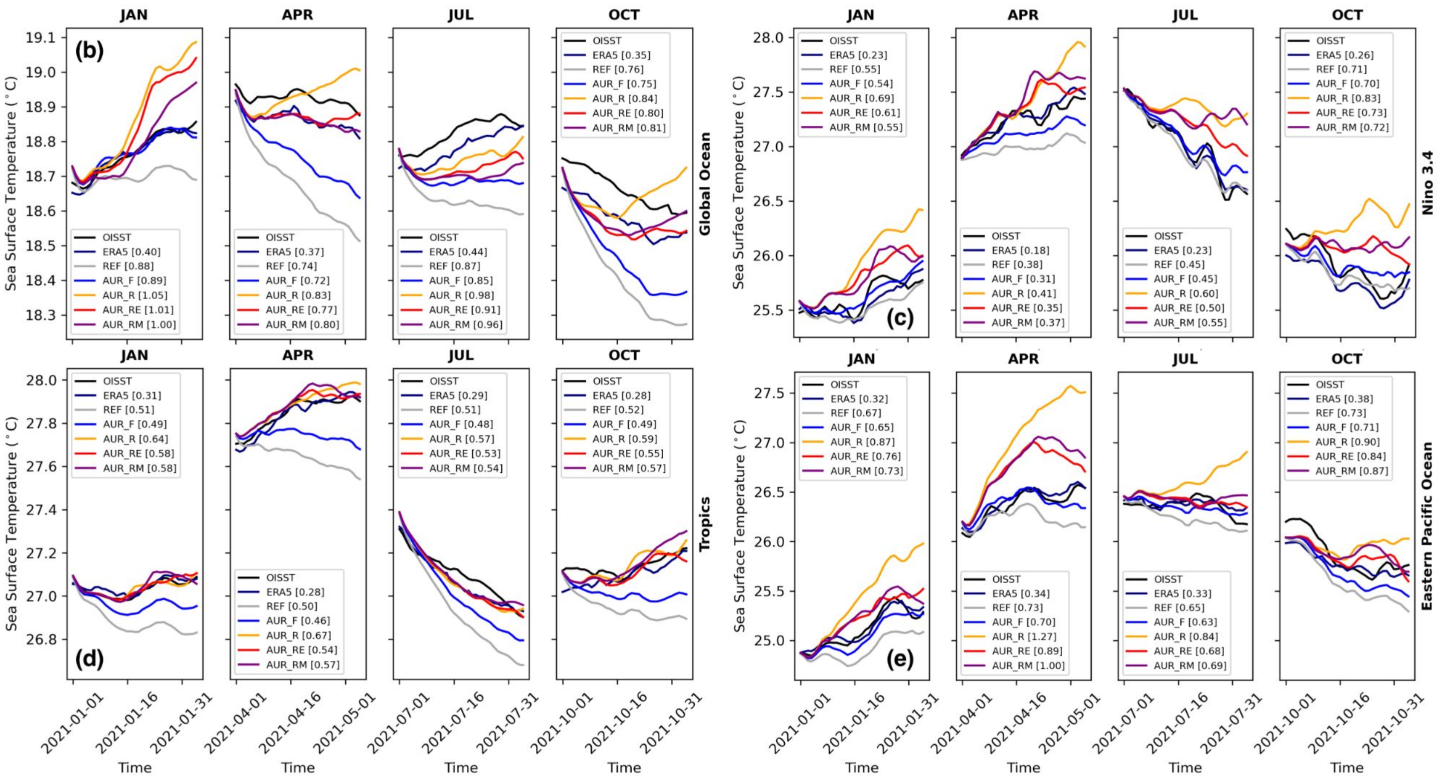
- Comparison against REF and OISST

Reduced bias respect to AUR_R



In general the performance of the Aurora coupled configurations are comparable with the REF simulation.
Some issues are found in MOM6 and CICE model configurations and ICs

Sea Surface Temperature from MOM6

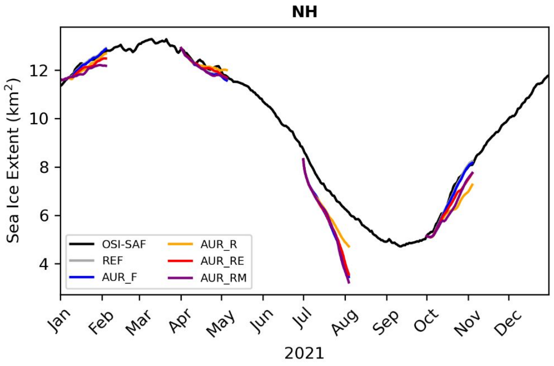


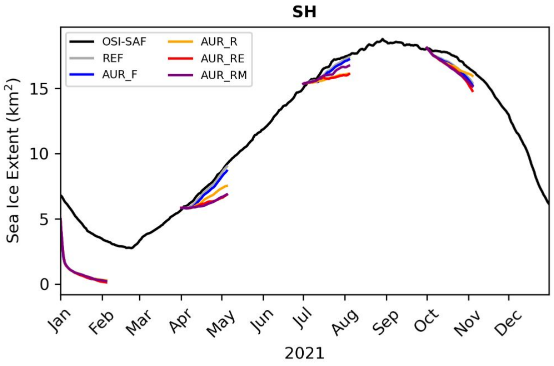
NCAR
UCAR

Sea Ice Extent and Coverage from CICE6

Very similar large bias in each hemisphere summer is appeared in all simulations incl. CDEPS DATM

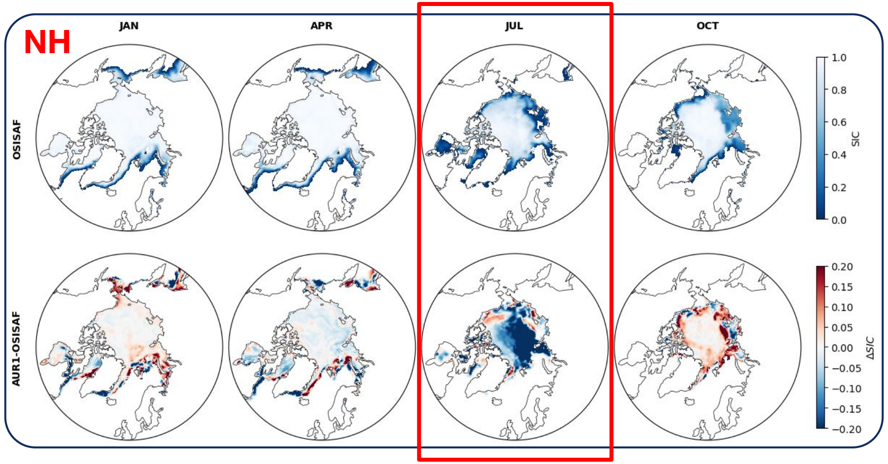
It is related with model configuration!!!

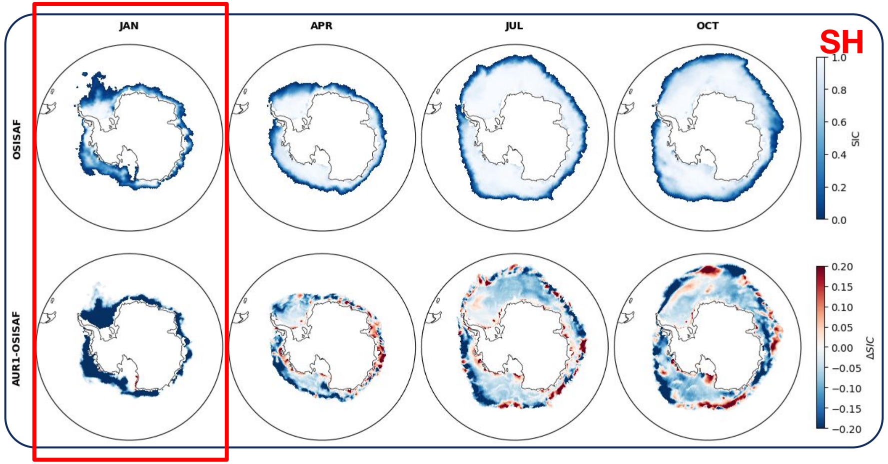




ERA5 has also issues in the polar regions with large bias in surface temperature

Batrak and Müller, 2019 and Michalezyk et al., 2026





HESM-S2S Coupling Summary

- A new ESMF/NUOPC based co-processing component is developed to interact with machine learning-based models to enable coupled hybrid modeling applications.
- The new approach is demonstrated to be flexible enough to construct various model configurations in a modeling system-agnostic way.
    - *ESMX* driver
    - *HAOCM-R* and *HESM-S2S* realistic applications
    - Seamlessly works with *CMEPS (mediator)* and *CDEPS (data components)*
- The fine-tuned version of the Aurora model is able to drive underlying model components to represent spatiotemporal variations

What is next?

- Ready to couple with home grown ML-besed models (CAM emulator etc.)
- It might be used to test ideas like online bias correction, data assimilation etc.
    - model/s → AI/ML (bias correction or DA) → model/s
- We are looking into options to integrate with UFS WM
    - It is already part of UFS Coastal (fork of UFS WM)
    - Ensemble approach might be important for UFS Coastal and also HAFS applications
- We have still room for optimization in GeoGate
    - C++ layer for Python embedding → initialization cost of model?
- The Aurora model can be also improved
    - physical constraints, longer training, more variable for coupling (flux components)

Dedicated GitHub Organization

GeoGate IO

Develops a set of novel toolset for interacting with fully-coupled Earth System Models.

1 follower · United States of America

Unfollow

To support co-processing needs and AI/ML interaction

README.md

The GeoGATE IO GitHub organization is dedicated to developing a state-of-art set of tools designed to handle the massive amount of data generated by fully coupled Earth system models. These novel toolset focuses on:

Co-processing: Enabling computations and analysis to happen concurrently with the model simulation, rather than post-processing it.

In situ visualization: Allowing for the real-time visualization/instrumentation of model outputs as they are being produced, providing immediate insights.

Bridging Earth System Models and cutting-edge AI/ML tools: Facilitating the application of advanced artificial intelligence and machine learning techniques to effectively analyze and extract valuable information from the immense Earth system data and provide interface to couple AI/ML models with prognostic Earth System Model components.

Essentially, GeoGate IO aims to provide researchers with the capabilities to efficiently process, interact, visualize, and learn from the vast and complex data produced by Earth System Models, pushing the boundaries of Earth science research.

Pinned

Customize pins

Core functionality

GeoGate Public

GeoGate is an open source project that provides a generic way to interact with the existing earth system models and applications coupled using ESMF/NUOPC library.

Fortran · 2 · 2

GeoGateApps Public

Demonstrate capabilities of GeoGate component using simplified configurations

Python

Prototype applications such as two component, multi-instance configuration

View as: Public

You are viewing the README and pinned repositories as a public user.

Get started with tasks that most successful organizations complete.

Discussions

Set up discussions to engage with your community!

Turn on discussions

People

Invite someone

Top languages

Fortran · Python

Realistic applications

HAOCM-R and HESM-S2S

are still in private repository and will be available under same GitHub organization after submission of the initial paper

Questions?

Collaborators:	
Wessel P. Bruinsma (co-author)	AI for Science
Robert Oehmke, Daniel Rosen, Gerhard Theurich, and Ann Tsay	ESMF Team
Brian Dobbins	CSEG
Saeed Moghimi	NOS
Denise Worthen and Nick Szapiro	NOAA/EMC
Gustavo Marques and David Bailey	CGD
Brian Vanderwende and Negin Sobhani	CISL