CAMulator: Fast Emulation of the Community Atmosphere Model

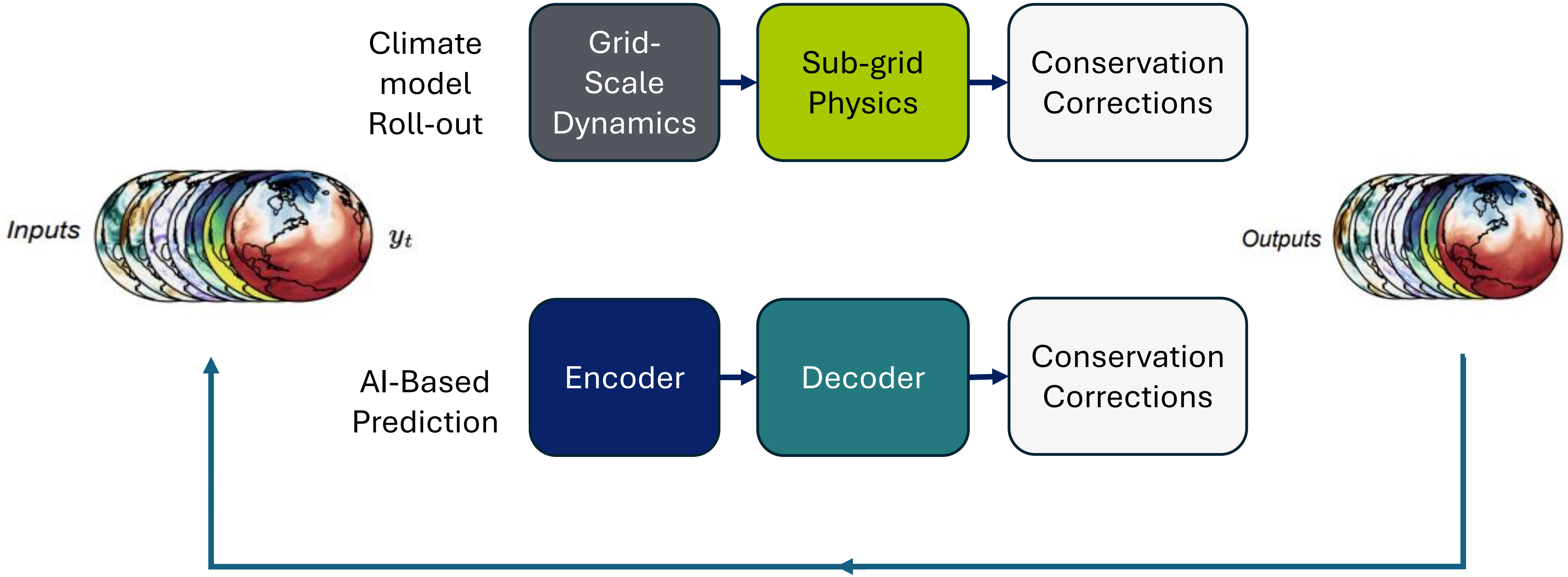
CESM wg, Feb 4th, 2025

Will Chapman[1], John Schreck[1], Yingkai Sha[1], David John Gagne[1], Dhamma Kimpara[1,2], Judith Berner[1]

[1] National Center for Atmospheric Research, Boulder Co.
[2] University of Colorado, Boulder Co.

Additional thanks to the M²LInES team, and NCAR Software Engineers

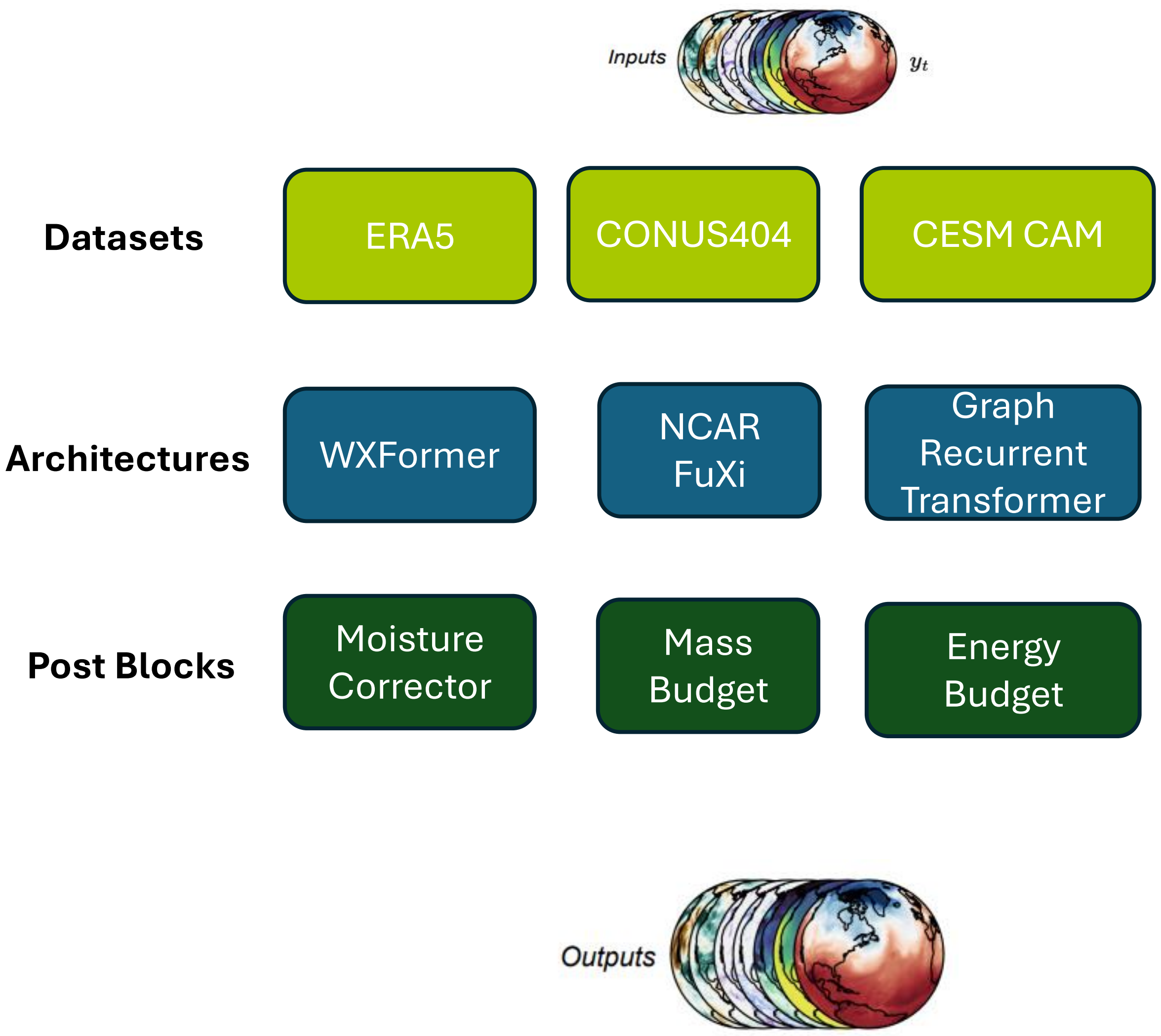
CAM vs. CAMulator workflow



NSF NCAR ***CREDIT*** PLATFORM

Schrek, Sha, Chapman, Gagne, Kimpara, Berner, Kazadi, Sobhani, Kirk

- Research platform for understanding of best practices for training and operating global and regional AI weather prediction / climate models
- Platform Features
  - Integrated pre-processing for reanalysis, reforecast, or model data
  - Library of existing and new PyTorch neural network weather prediction architectures
  - Scalable training and inference on NCAR HPC
  - Analysis tools and plotting
- Novel advances
  - WXFormer architecture
  - More physically informed inputs
  - Stable hourly global model out to **10** days
  - Spectral normalization and new padding stabilizes multiple architectures
  - Physical constraints improve precipitation and other state variables



Training Challenges / Solutions:

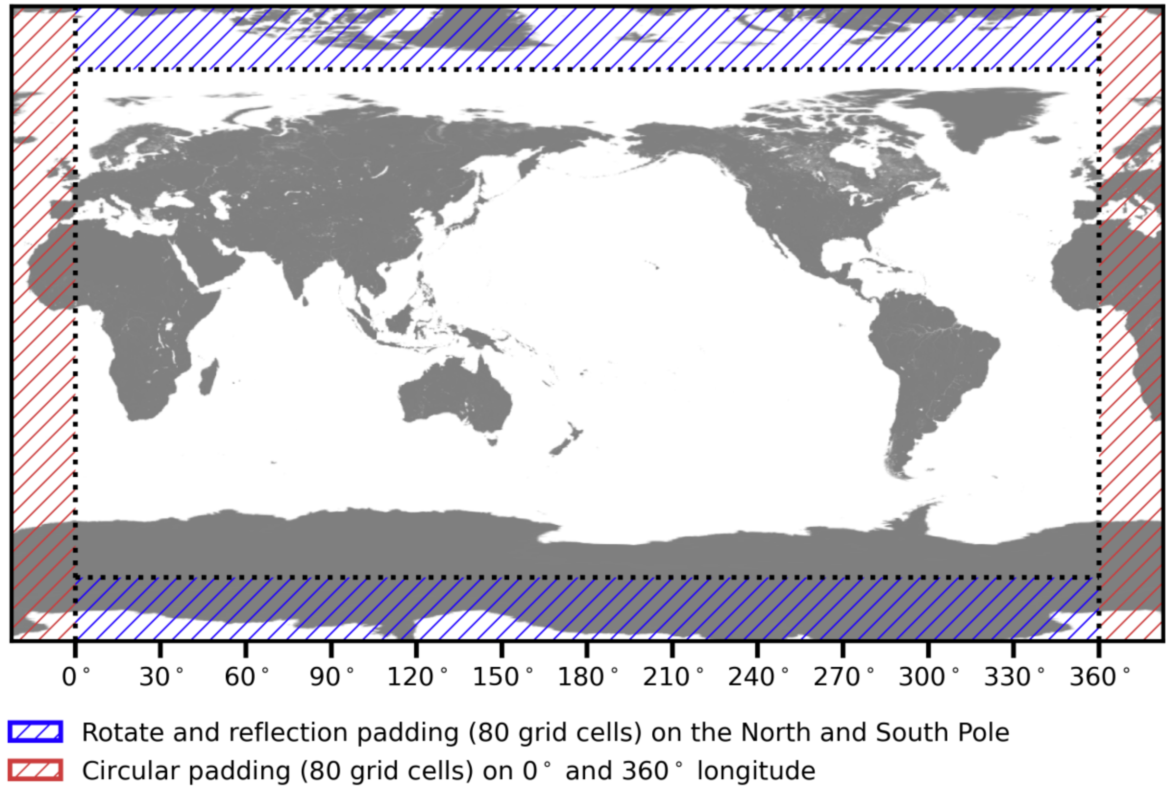
- **Single-step training**
  - Predict the next timestep (1 or 6 hours) and backpropagate
  - 1 sample per GPU; 16 GPUs total
  - Loss: latitude-weighted MSE
  - Gradient checkpointing used to reduce memory usage (at expense of computation)

- **Multi-step training**
  - Run model forward in time multiple steps and accumulate gradient against truth data along trajectory
  - Backpropagate gradients through time to update weights
  - Start with shorter steps and extend length up to 3 days (6-hour model) or 2 days (1-hour model)
  - Necessary for stable and accurate rollouts

Training Challenges

- Setting up PyTorch to use MPI and NCCL
- Chunking input data shapes
- Correct padding for across latitude and longitude
- Stabilized autoregressive prediction via spectral normalization

Boundary padding operations in CREDIT models



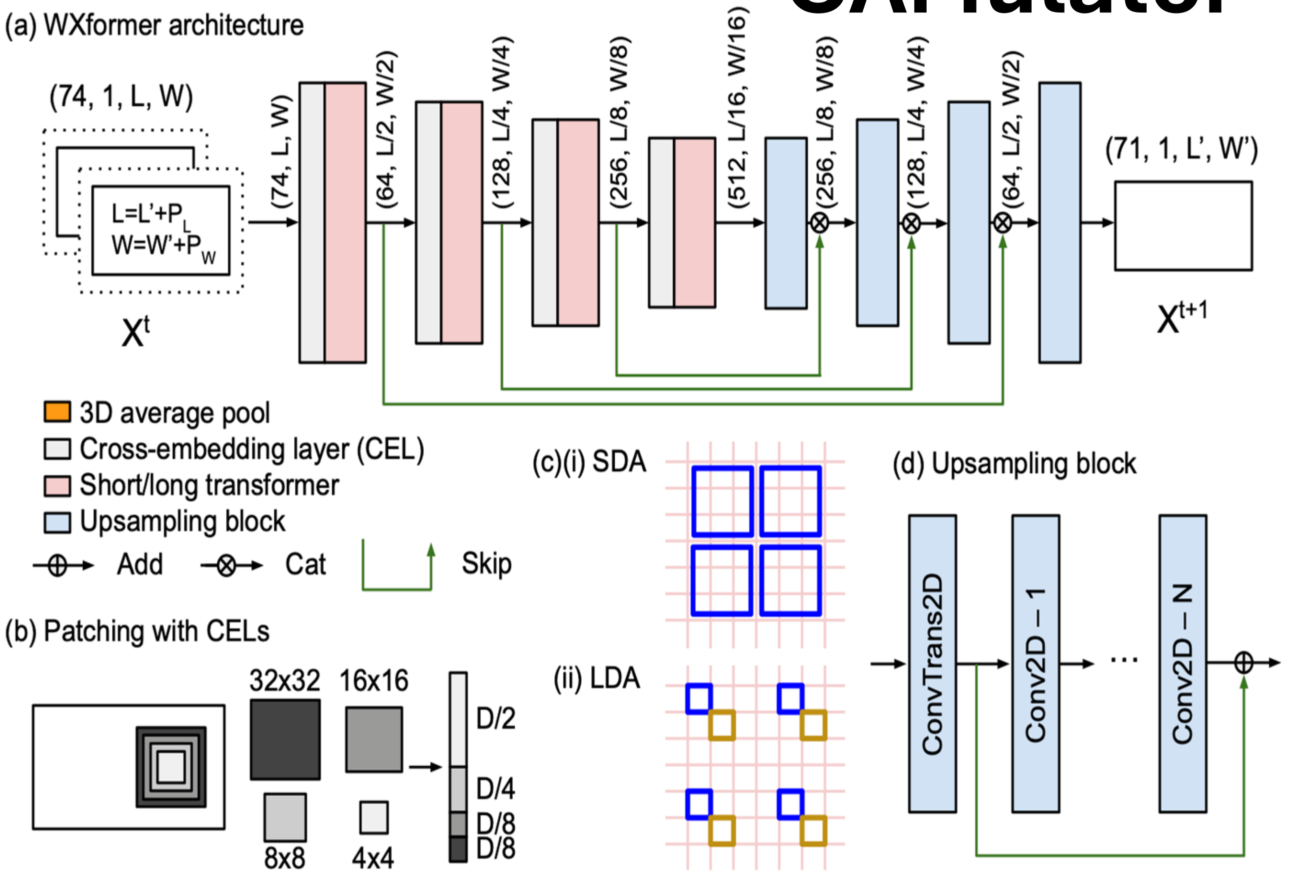
Emulating CAM 6 (CESM 2.1.5)

<u>Data Prep Workflow</u>

1) Conduct a 35-year FHIST run (1979-2014) of CAM6 in CESMv2.1.5
2) Collect 6-hourly data
3) Compute Total Water + Convert to Flux forms
4) Gather Static Forcing Data
5) Package in yearly Zarr data structures

Variable	Description	Units	Single Level/Levels	I/O
	Prognostic Variables (Input and Output)			
U	Zonal Wind	m/s	32 levels	Input/Output
V	Meridional Wind	m/s	32 levels	Input/Output
T	Temperature	K	32 levels	Input/Output
Qtot	Specific Total Water	kg/kg	32 levels	Input/Output
	Diagnostic Variables (Output Only)			
PRECT	Precipitation Rate	m	Single Level	Output
CLDTOT	Total Cloud Cover	fraction	Single Level	Output
CLDHGH	High Cloud Cover	fraction	Single Level	Output
CLDLOW	Low Cloud Cover	fraction	Single Level	Output
CLDMED	Medium Cloud Cover	fraction	Single Level	Output
TAUX	Zonal Wind Stress	N/m^2	Single Level	Output
TAUY	Meridional Wind Stress	N/m^2	Single Level	Output
U10	10m Wind Speed	m/s	Single Level	Output
QFLX	Surface Moisture Flux	m	Single Level	Output
FSNS	Net Solar Flux at Surface	J/m^2	Single Level	Output
FLNS	Net Longwave Flux at Surface	J/m^2	Single Level	Output
FSNT	Net Solar Flux at TOA	J/m^2	Single Level	Output
FLNT	Net Longwave Flux at TOA	J/m^2	Single Level	Output
SHFLX	Sensible Heat Flux	J/m^2	Single Level	Output
LHFLX	Latent Heat Flux	J/m^2	Single Level	Output
	Surface Variables Prognostic (Input and Output)			
PS	Surface Pressure	Pa	Single Level	Input/Output
TREFHT	Near-Surface Air Temperature	K	Single Level	Input/Output
	Dynamic Forcing Variables (Input Only)			
SOLIN	Incoming Solar Radiation	J/m^2	Single Level	Input
SST	Sea Surface Temperature	K	Single Level	Input
	Static Forcing Variables (Input Only)			
Surface Geop.	Normalized Surface Height	m^2/s^2	Single Level	Input
Land-Sea Mask	Land Mask × Cosine Latitude	unitless	Single Level	Input

CAMulator

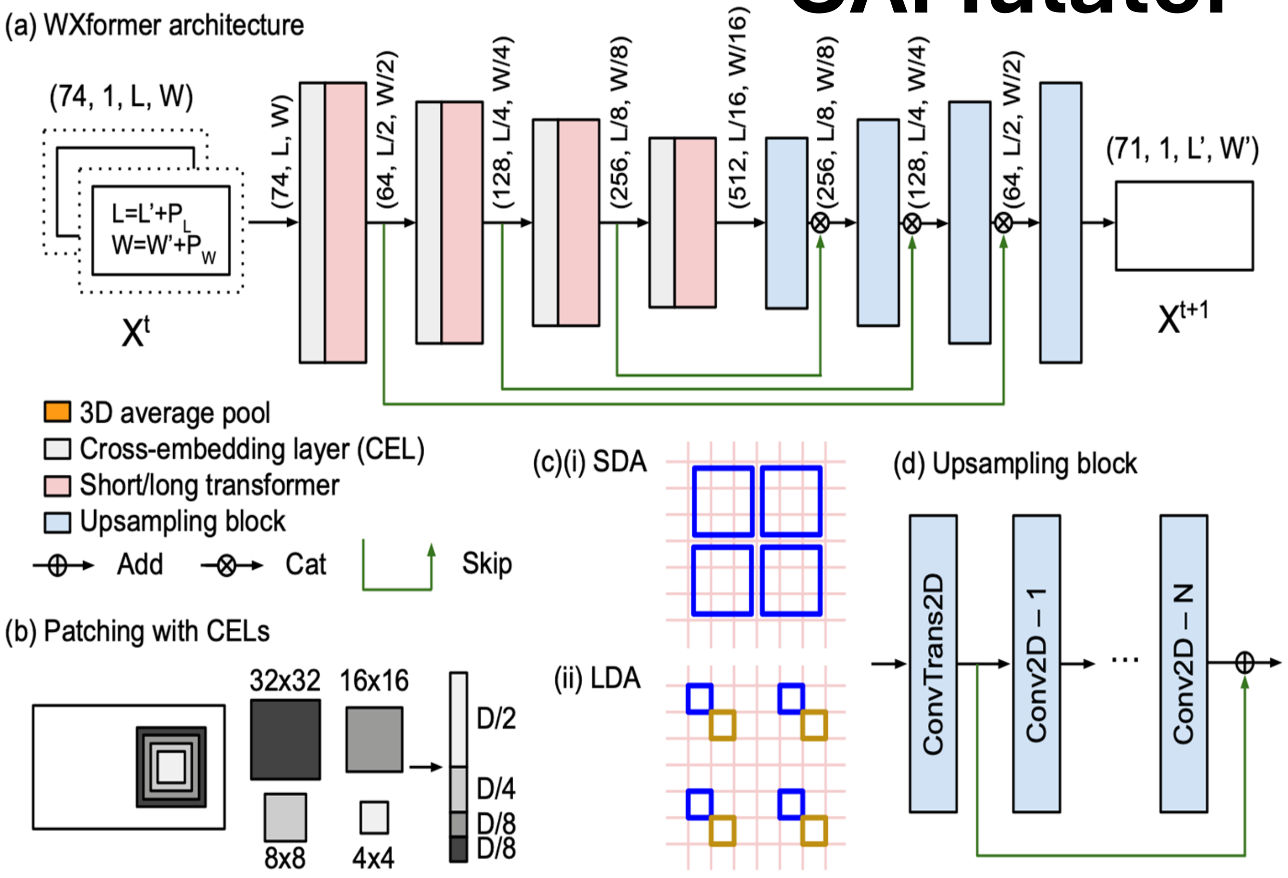


- O1000 lines of code
- 751,134,146 parameters
- 0.428775321% of GPT-3
- 2.86535 GB

Training Routine:

- Train on ***1 step*** prediction for 150 epochs on 16 Derecho GPUS (~1.5 days)
  - Do not enforce physical constraints
- Train & Fine-Tune on ***2 step*** prediction for 90 "*epochs*" on 1 Derecho GPU (~.5 days)
  - Activate physical constraints
- Initially validate on 1 or 2 step prediction skill using RMSE as a metric.
- Fine-tune validation on a 1-year climatology run conducted between *epochs* compared to a 10-year climatology.

CAMulator

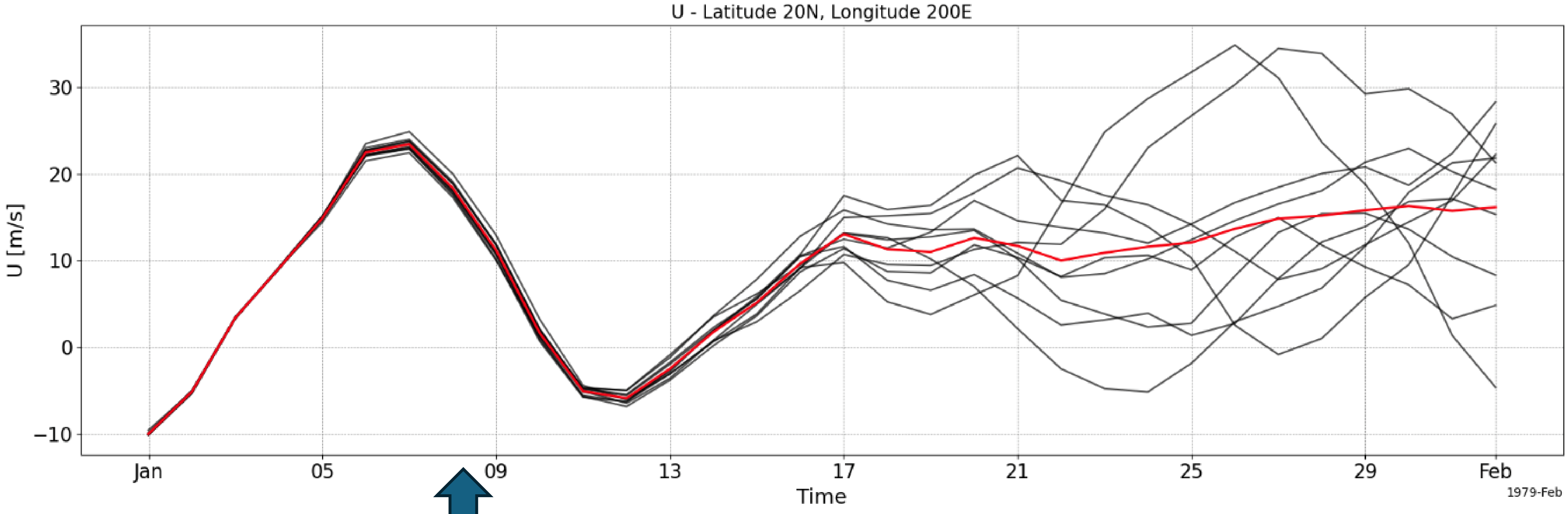


- O1000 lines of code
- 751,134,146 parameters
- 0.428775321% of GPT-3
- 2.86535 GB

Model Inference

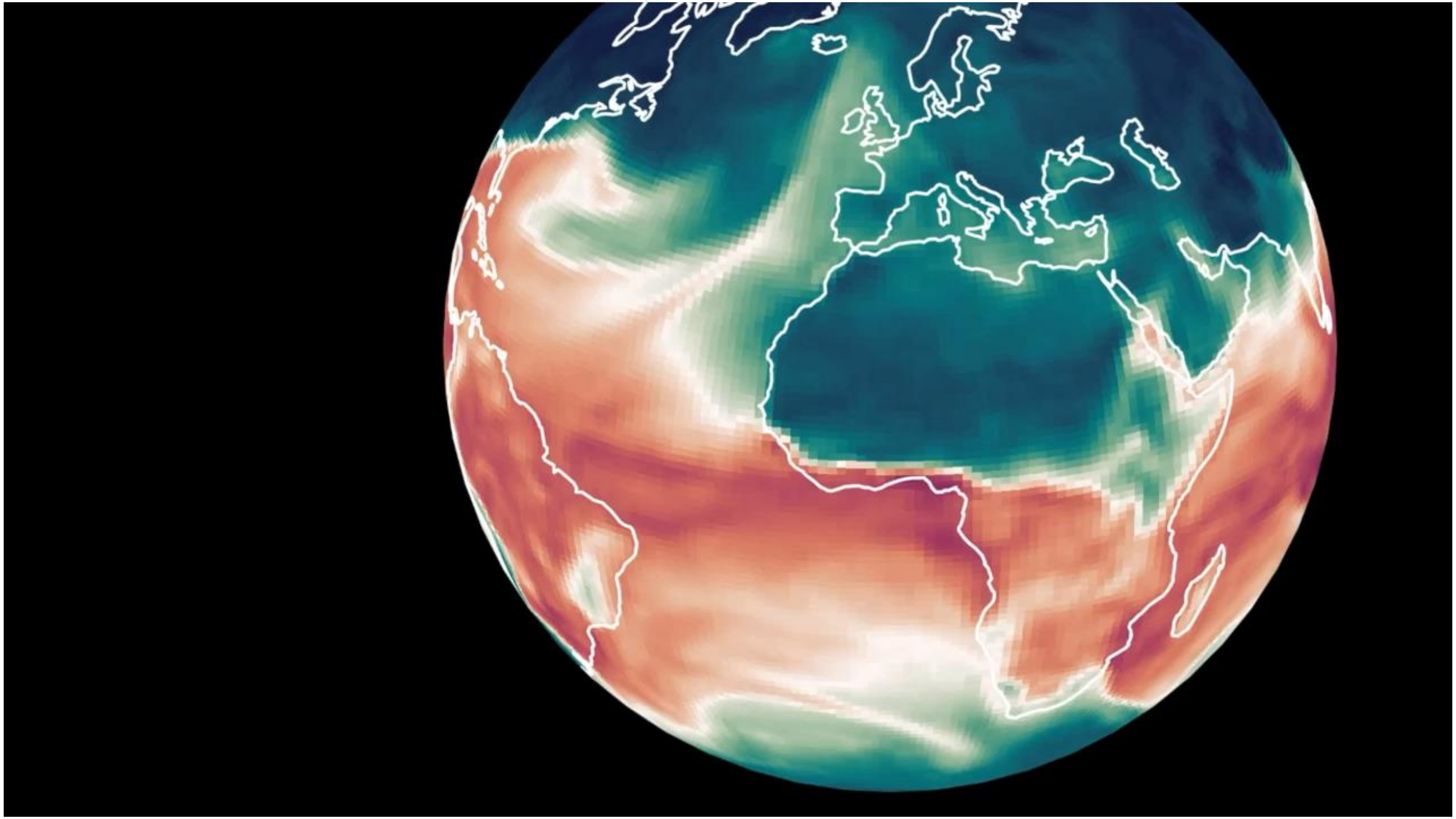
- **480 Model years / day** on a single CASPER GPU node (with 6-hourly I/O)
- CAM6 @ identical resolution **~14 model years / day** on 10 nodes.
- ~400x speed up

Let's Look at some 35 year runs [~90 minutes to complete]:

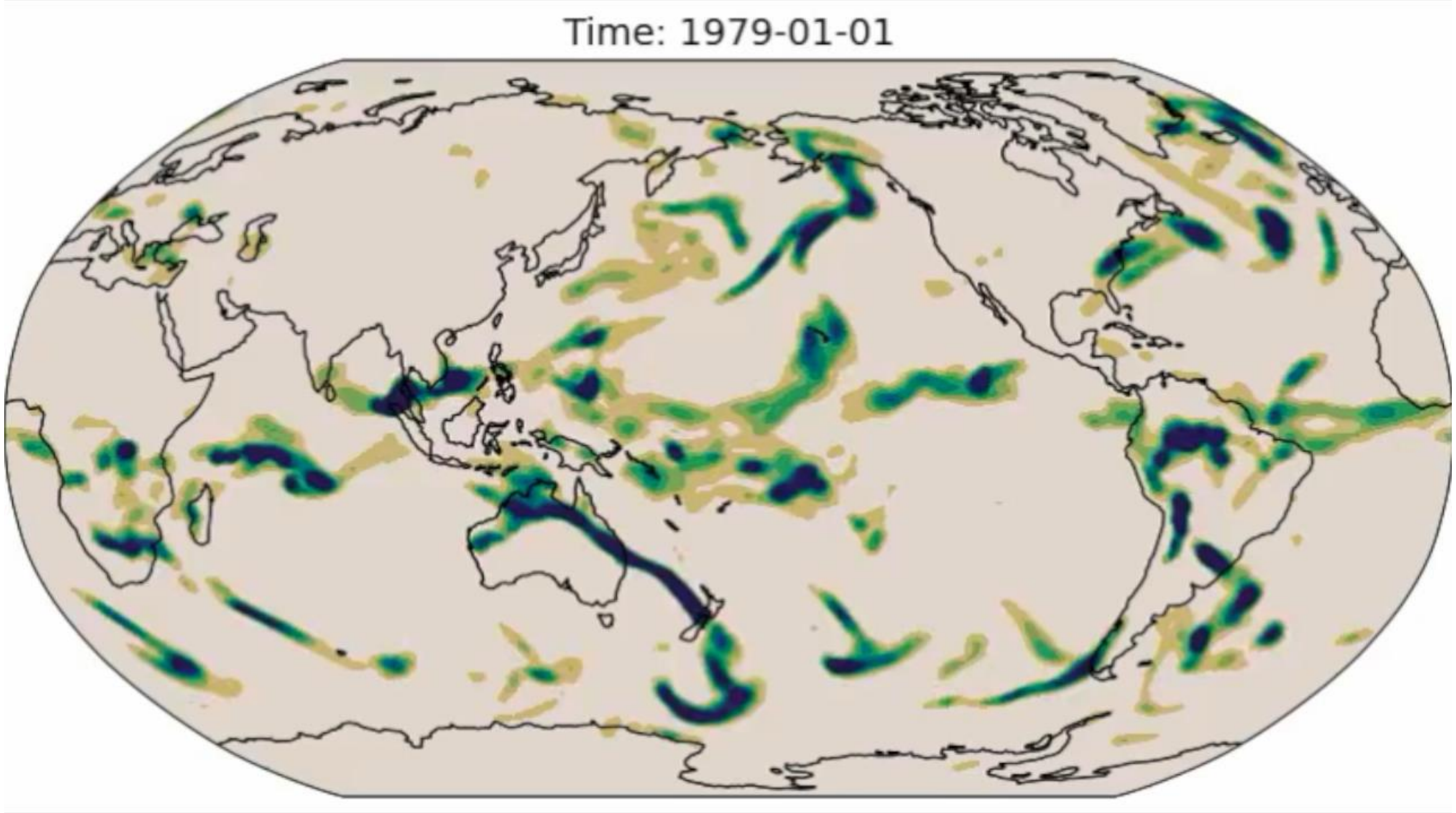


Adding a 'pertlim'

CAMulator



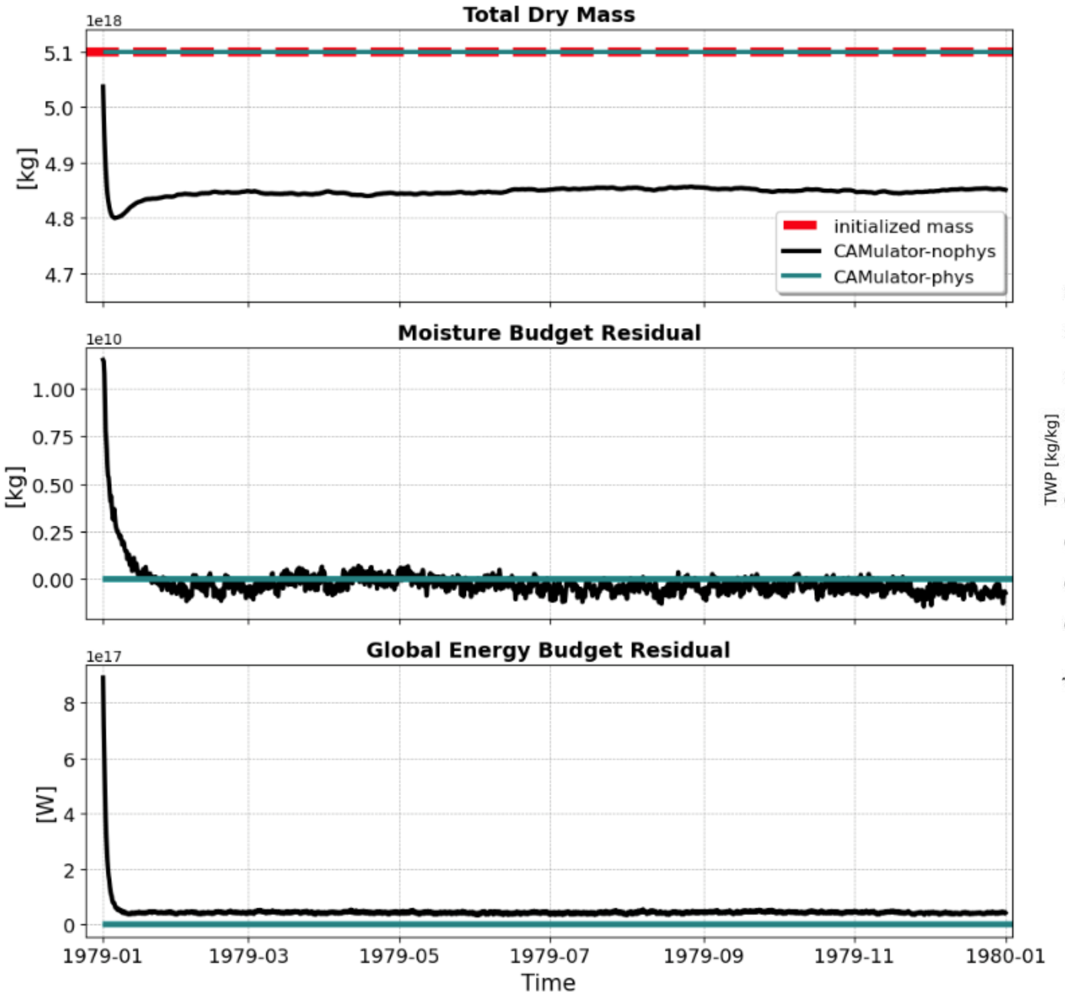
Developing Precip. Climatology

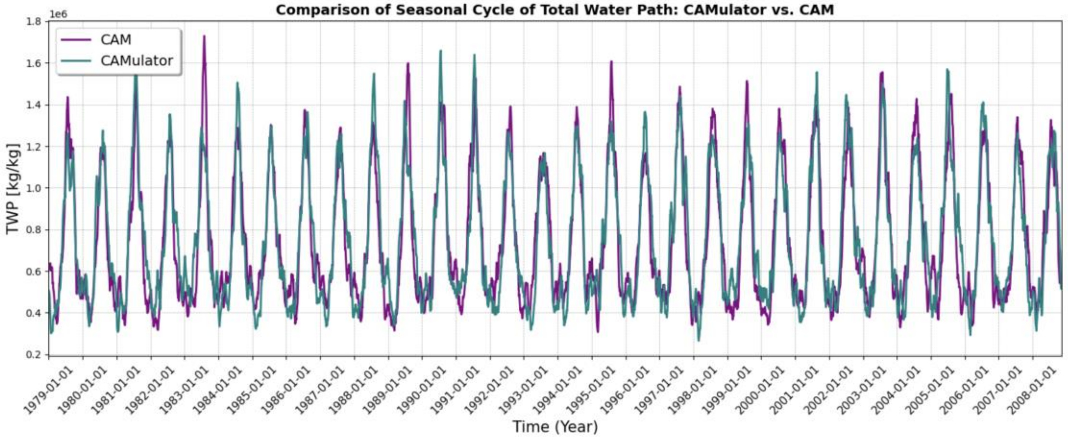


Imposed Physics Constraints

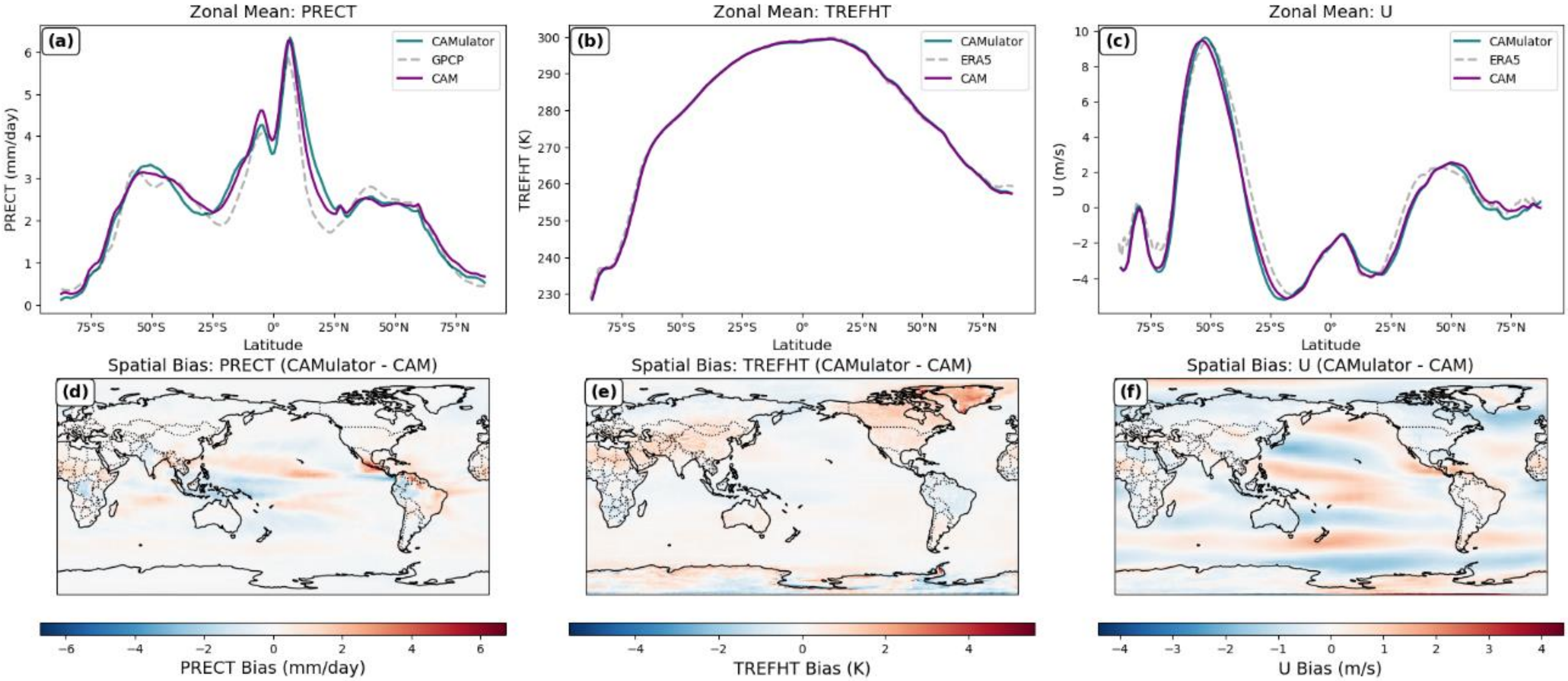
Model state [Δt]

WXformer → **Non-Negative** correction → **Global Dry Air Mass** conservation → **Global Moisture Budget** conservation → **Global total atmospheric** energy conservation → Compute Training Loss

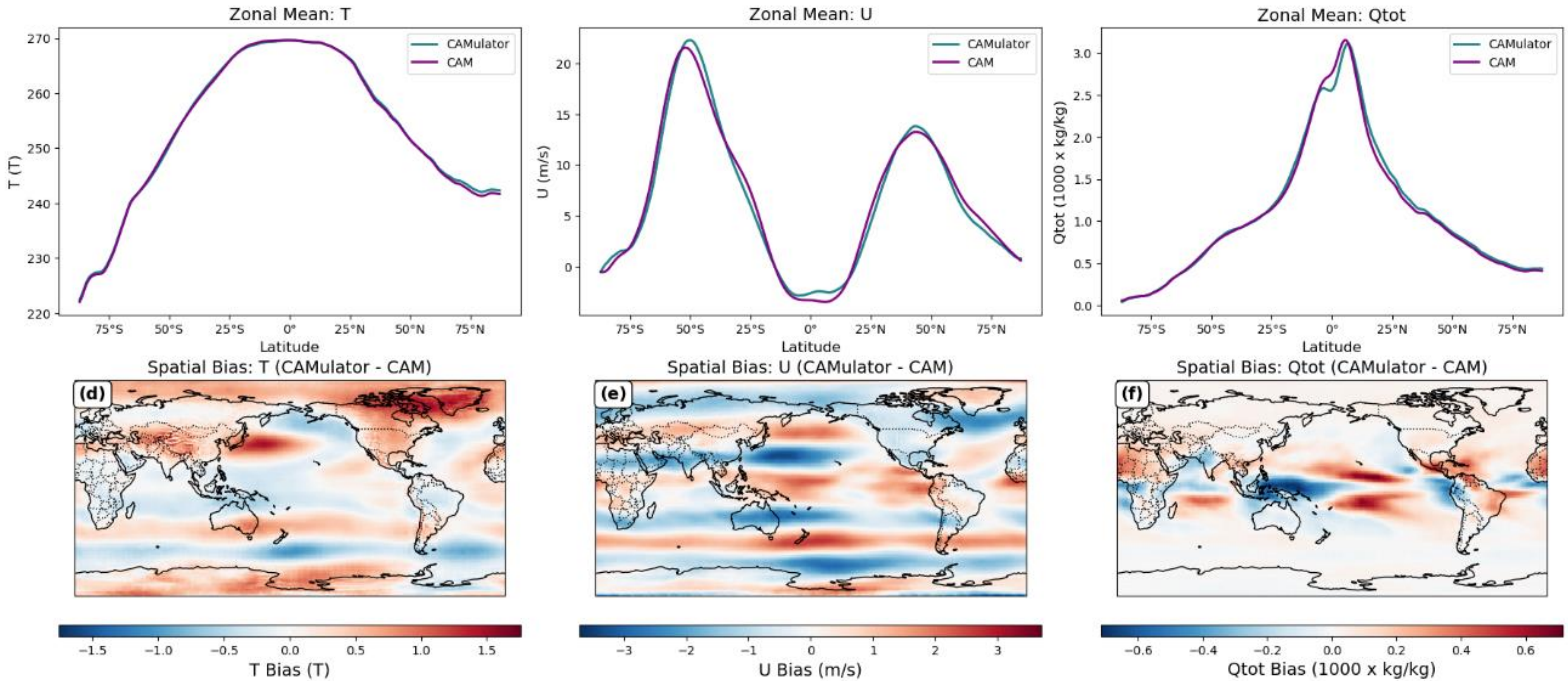




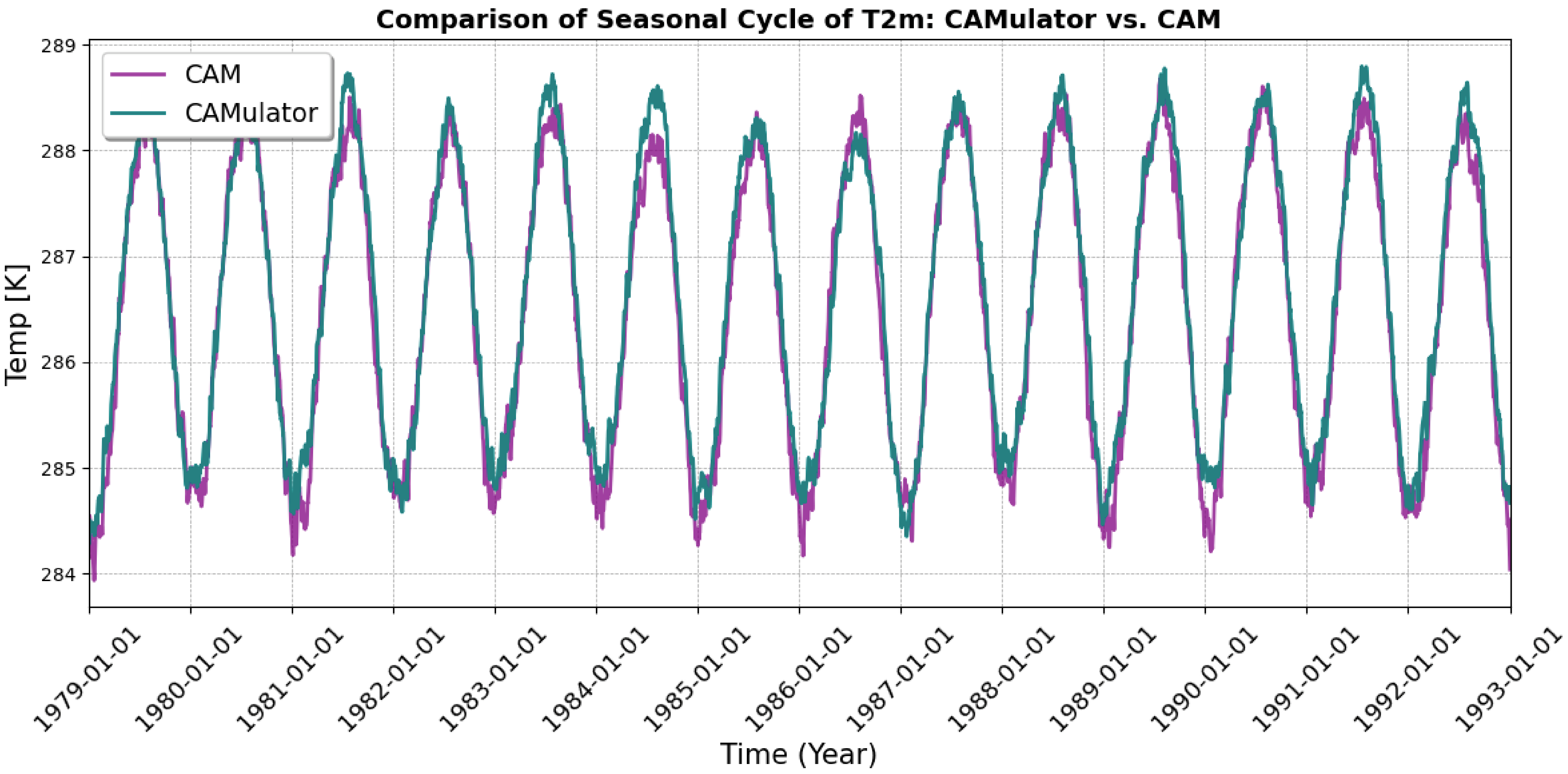
Annual Biases:



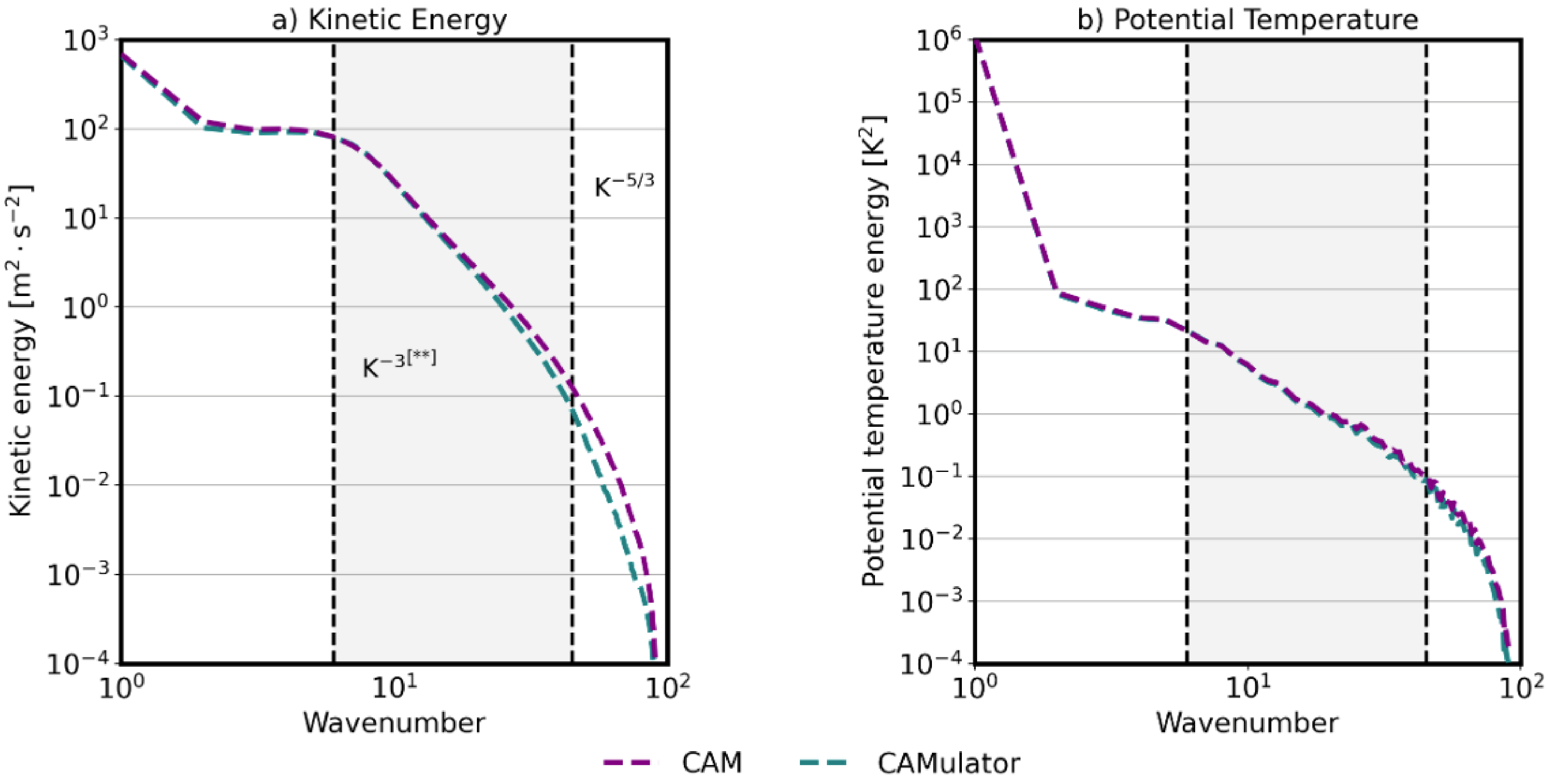
Annual Biases:



Seasonal Cycle T2m:



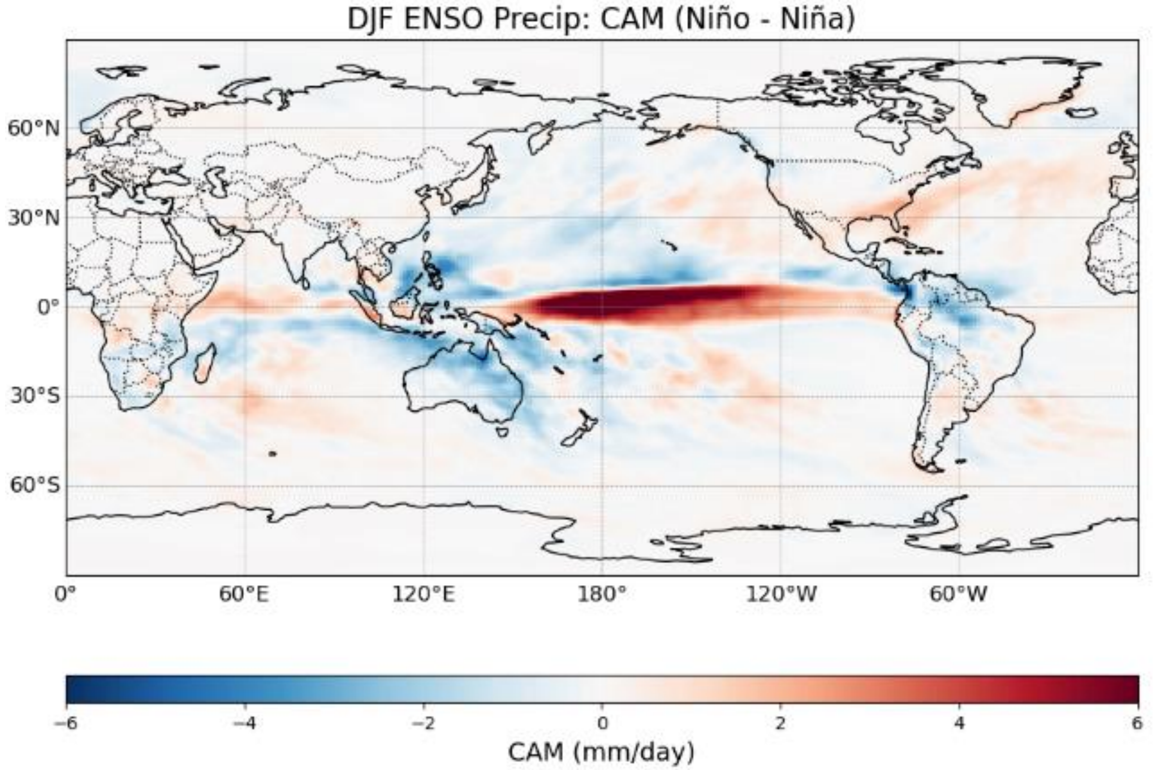
Effective Resolution (model smoothing) – 500mb:

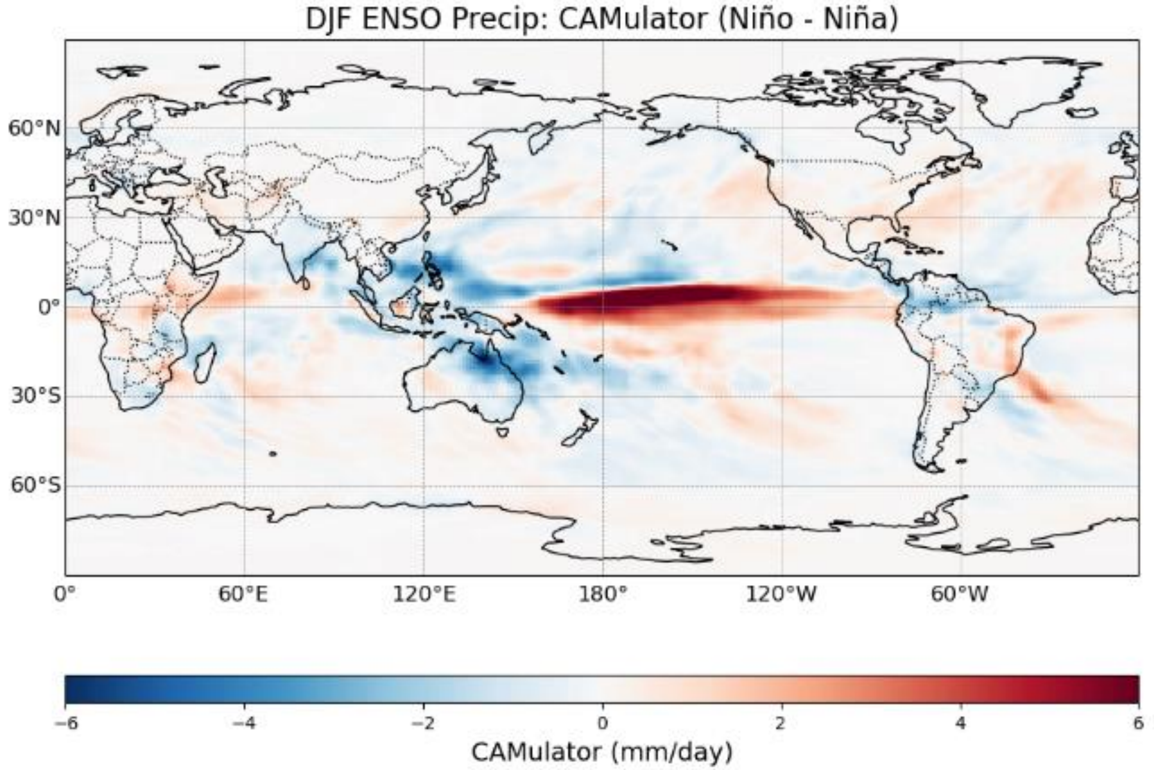


[*] Energy spectrum was computed using spherical harmonic transform on each initialization and forecast lead time, and averaged as mean values.

[**] K^{-3} and $K^{-5/3}$ regions were estimated based on the 1000-100 km scale wavelength on 45 ° N and 45 ° S.

ENSO Precipitation Response; ***Composite of 8 strongest Niño's – Niña's***

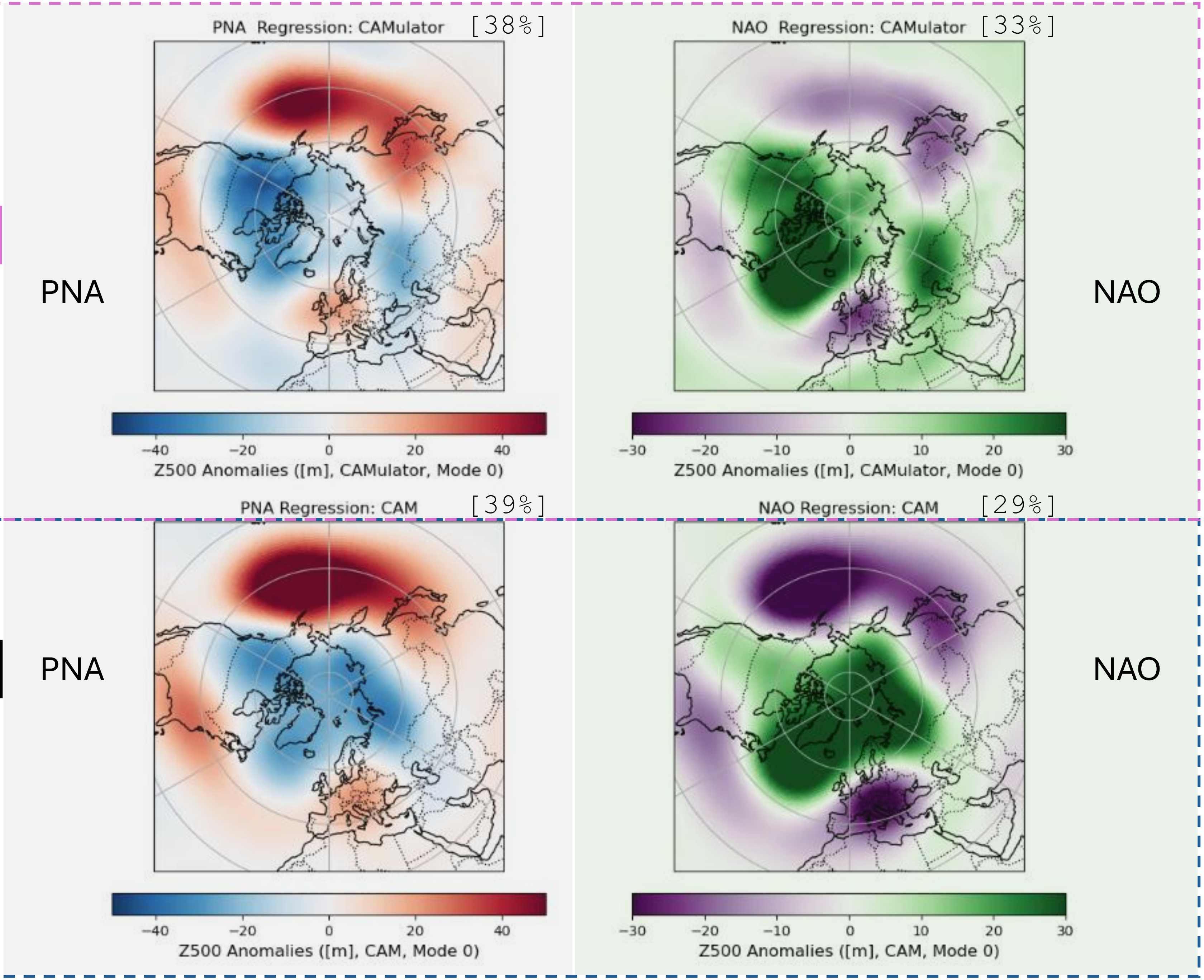




NSF NCAR Operated by UCAR

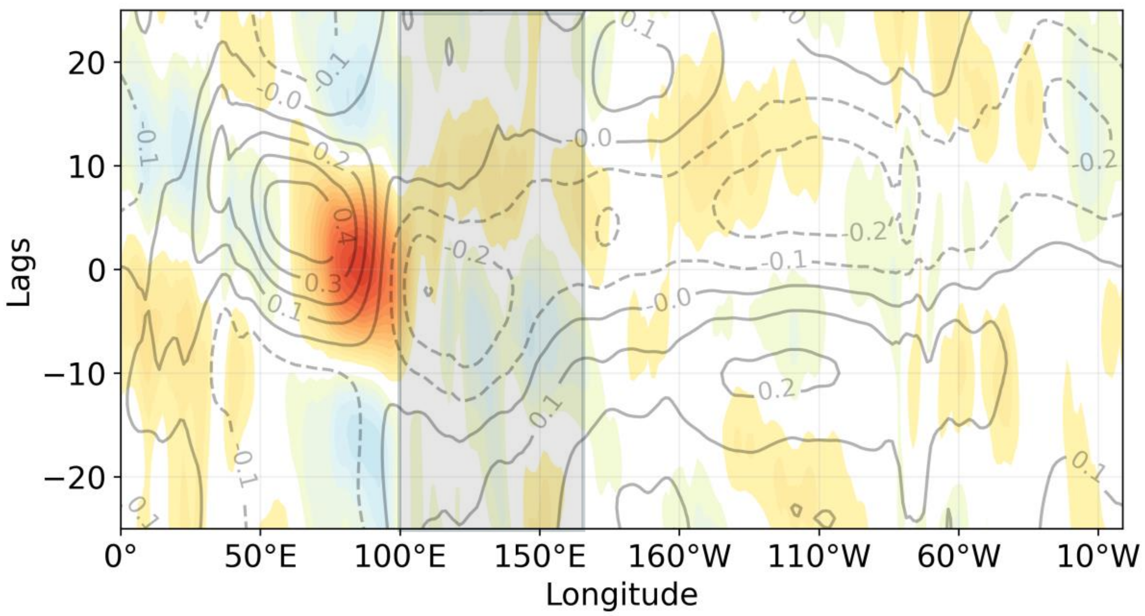
CAMulator →

CAM →

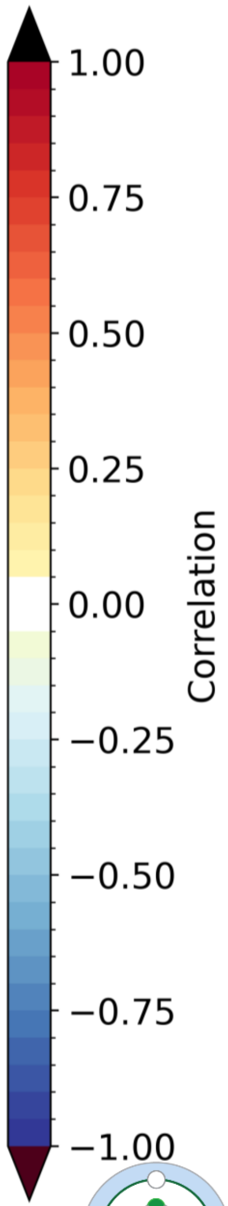


MJO propagation over Maritime continent

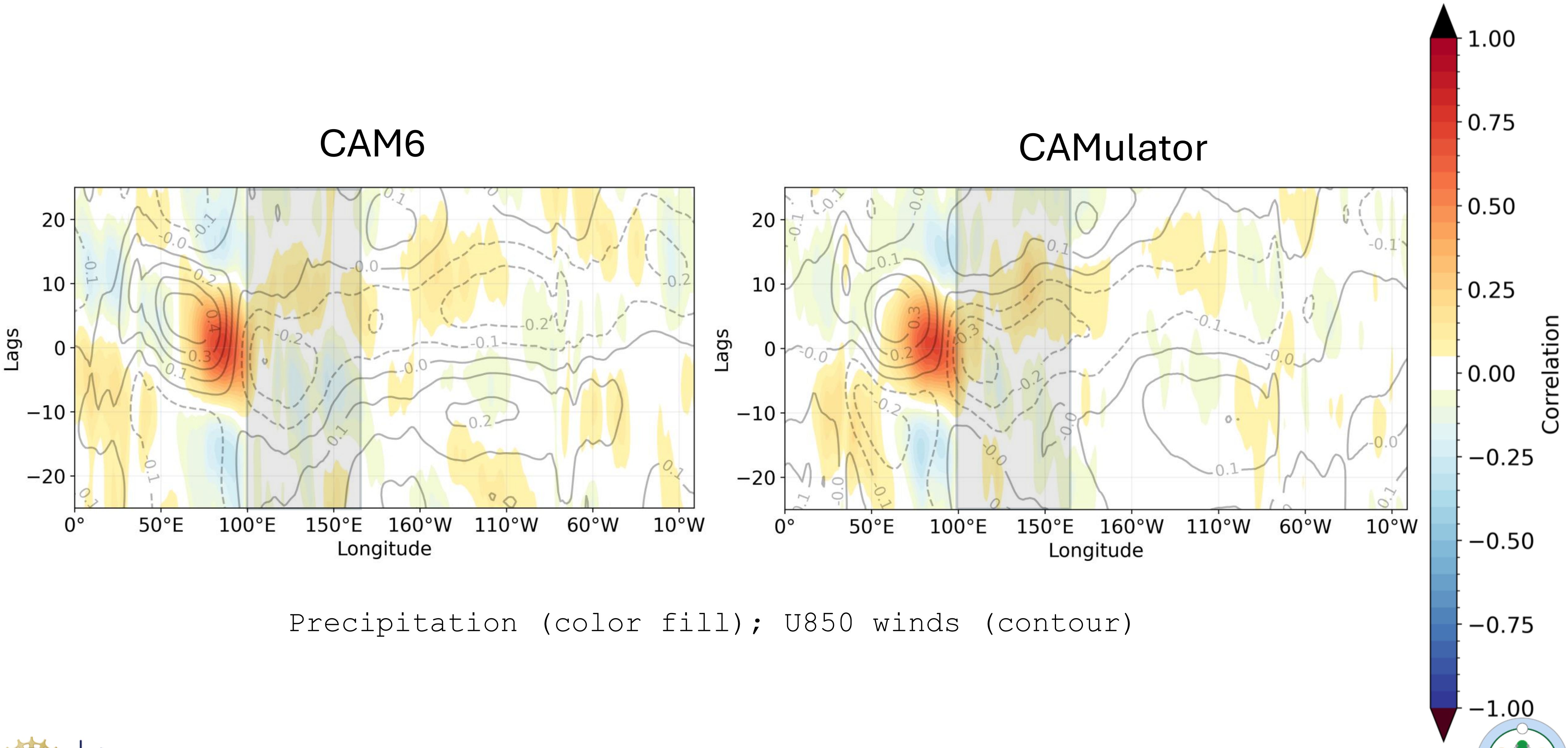
CAM6



CAMulator



MJO propagation over Maritime continent



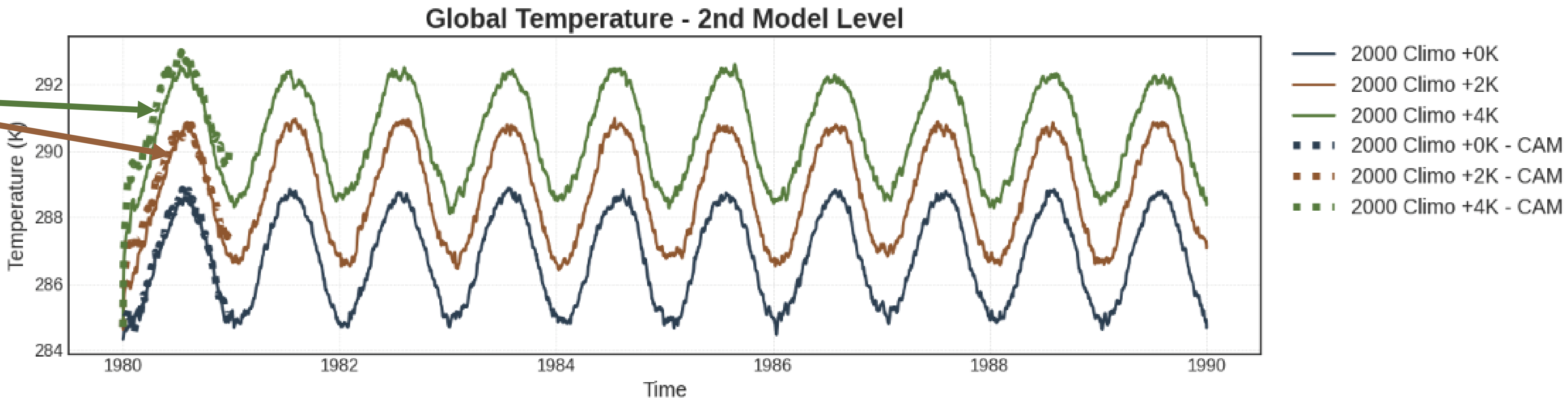
Precipitation (color fill); U850 winds (contour)

Current Configured Run Scenarios (SST driven):

- 1979, 2000, 2010 Climatology Run
- 1979-2014 historical SST case
- Coupled SST cases (35 years)
- **Year 2000, +2K, +4K runs**

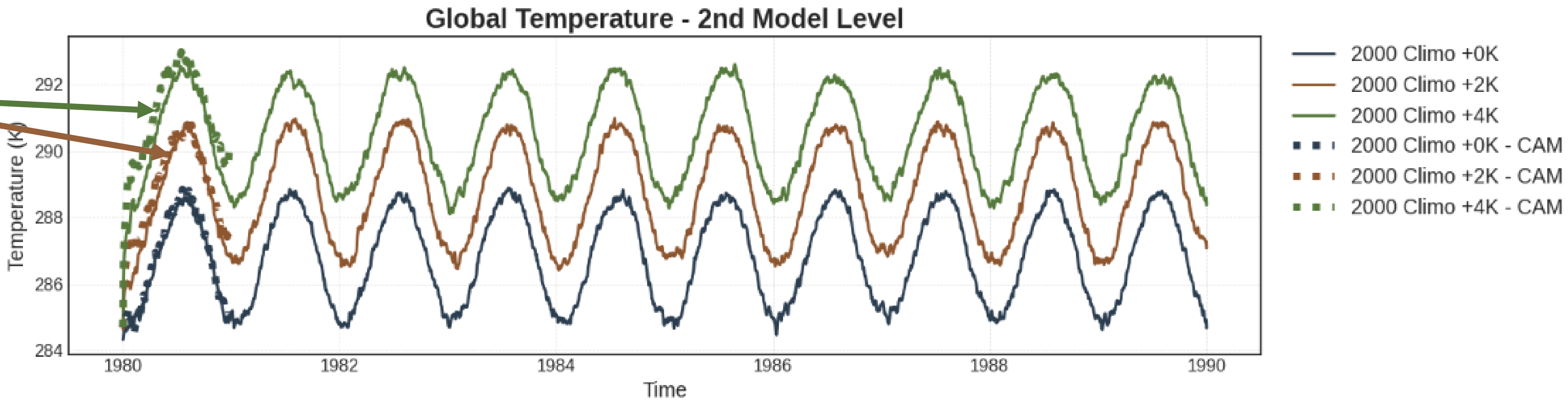
Working out of training sample:

What if we force the model with increased SST runs?



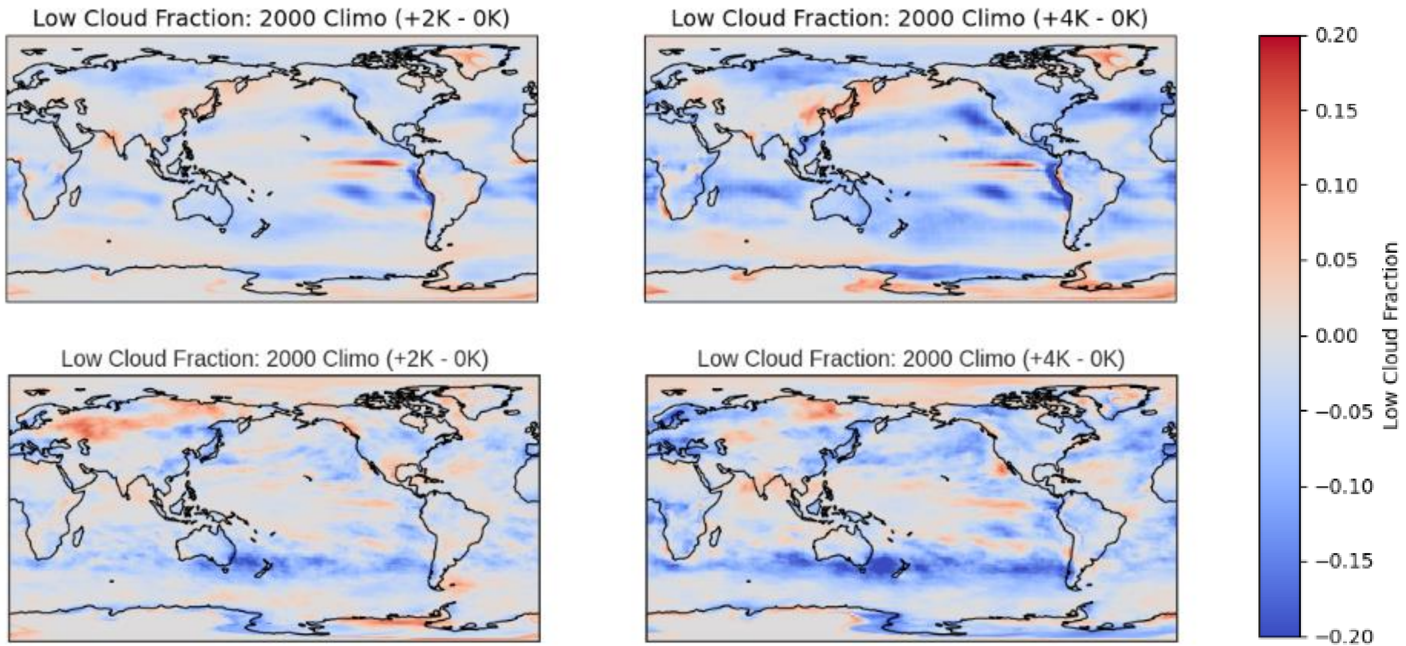
Working out of training sample:

What if we force the model with increased SST runs?



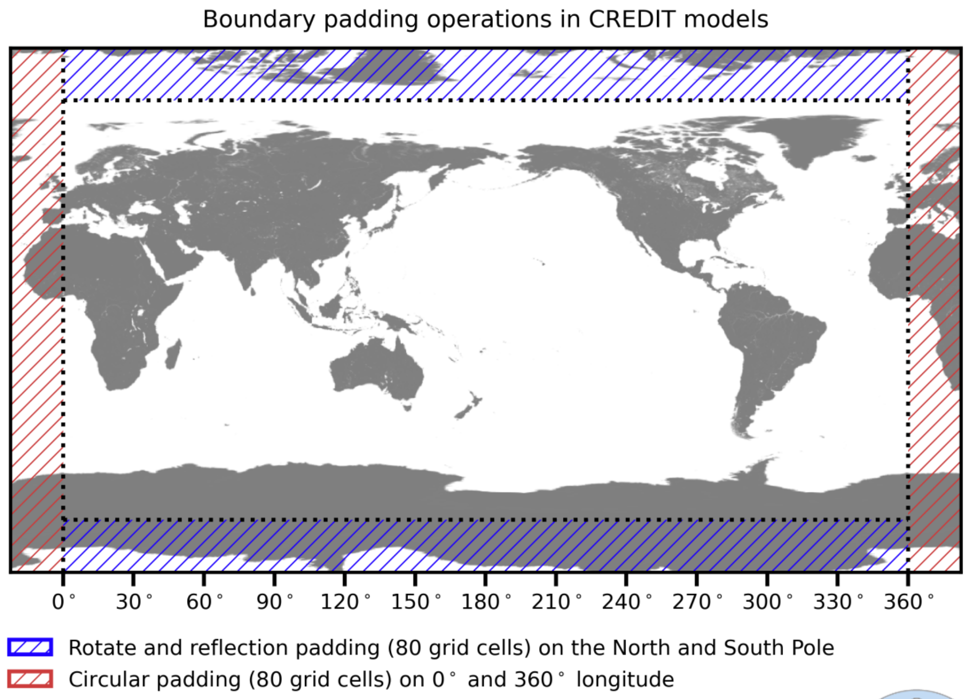
CAMulator (10 years):

CAM (1 year):



Current / Future Projects

- Easy / Calibrated weather→ seasonal ensemble creation
  - Stochastic backscatter + CRPS based training. ***Thanks Dhamma!***
  - this gets us up to 50 members per 40g GPU → ~1000 model years per day
- Plugging into existing NCAR/community verification suites (cupid, ADF, mdtf)
- CO^2 forcing scenarios
- Score-based diffusion GenAI fields.
- ERA5 nudging/improvement
- Super modeling
- Handle ice/sea-ice better
- Ocean, Land, Sea-Ice emulation
- Dynamic model coupling
- Improvements to validation, visualization routines
- Improvement to grid representation



Questions?