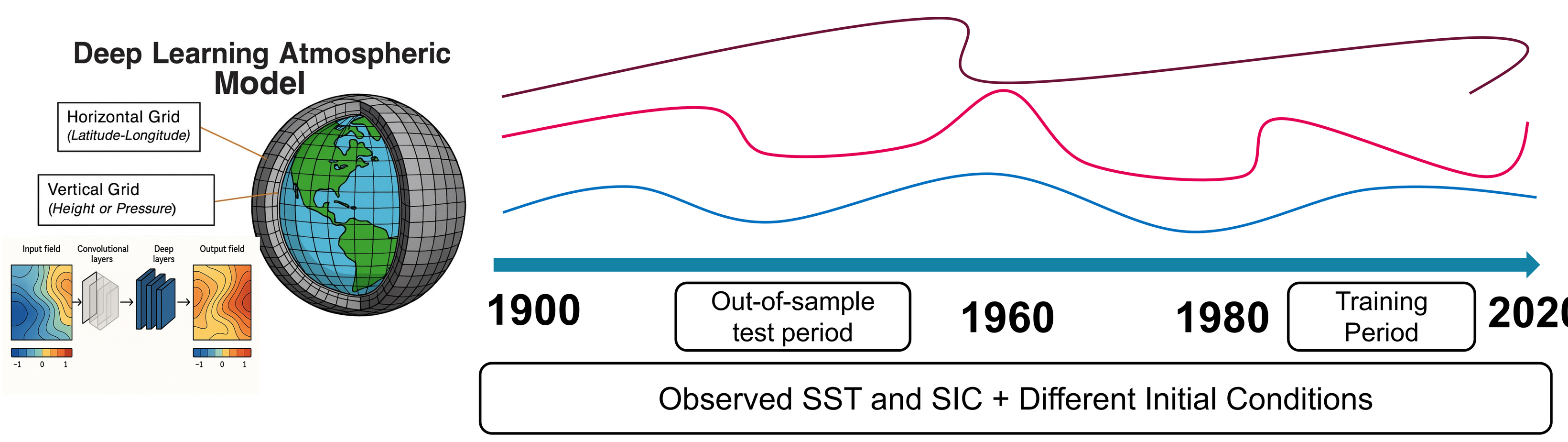
Deep Learning Atmospheric Models Reliably Simulate Out-of-Sample Temperature Extremes and Blocking Frequencies

Email: zilumeng@uw.edu
Homepage: zilum.github.io
Preprint: arxiv.org/abs/2507.03176

Zilu Meng[1,*], Gregory Hakim[1], Wenchang Yang[2], and Gabriel Vecchi[2]; [1]Department of Atmospheric and Climate Science, University of Washington; [2]Department of Geosciences, Princeton University

Background and Motivation



DLESyM (Cresswell-Clay et al. 2025)	Neural GCM (Kochkov et al. 2024)	HiRAM (Zhao et al., 2009)
Full Neural Network (DLWP U-Net) 100 members	Physical Equation + Learned Physics 100 members	Physical High-Res Model 5 members

Real World Observations = Internal atmospheric variability+ Signal from Boundary Conditions

Large Ensemble AMIP Simulations Mean = Signal from Boundary Conditions (SST)

Scientific Goals

- **Validation**: To rigorously test if Deep Learning (DL) atmospheric models (DLESYM, NGCM) can reproduce historical extreme weathers (1900-1960) outside their training period as reliably as physical models.
- **Attribution**: To utilize the speed of DL models to generate massive ensembles (100 members), allowing us to isolate the influence of Sea Surface Temperature (SST) variability on atmospheric blocking and temperature extremes.

Models & Experiment Design & Data

1. The Model Hierarchy

Attribute	NGCM	DLESyM	HiRAM
Model Type	Physics–DL Hybrid	Purely Data-Driven	Physical GCM
Horizontal Resolution	~ 2.8°	~ 1°	~50 km (C96)
Boundary Forcing	SST & SIC	SST	SST & SIC
Training Period	1979–2017 (ERA5)	1983–2016 (ERA5/ISCCP)	N/A (Physics-based)
Ensemble Size	100	100	5
Running Time	6 hours/100yrs	1.9 hours/100yrs	10,00 hours/1yr
Device	1 CPU + 1 GPU	1 CPU + 1 GPU	1 CPU

2. Simulation Protocol (AMIP-style)

- **Forcing**: All models were driven by historical observed Sea Surface Temperatures (SST) and Sea Ice Concentration (SIC) from HadISST with random initial conditions.
- **Timeline**: Simulations covered 1900–2020.
- **The Generalization Test**: In-Sample (1980–2020): Period used to train the AI models; Out-of-Sample (1900–1960): The critical validation period. Can AI models simulate weather patterns from a climate they have never "seen"?

3. Defining the Extremes – Heatwave, Cold wave and Atmospheric Blocking

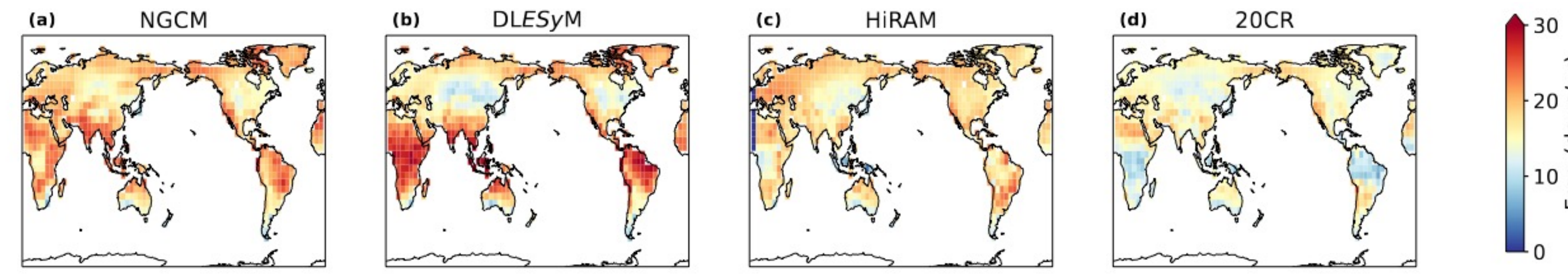
- **Land Heatwave and Cold Wave**: For each grid point, daily temperature exceedance is computed as the difference between the daily mean 2-meter air temperature and the 90th-percentile temperature for that calendar day, where the percentile is estimated using a 15-day moving window across all available years. A heatwave is identified when this exceedance remains positive for at least three consecutive days. Cold waves are defined analogously, using the 10th percentile and requiring negative exceedance for three days.
- **Atmospheric Blocking**: A 2D gradient-based index applied to 500-hPa geopotential height (Davini et al. 2020). For each grid point from 30°–75°N, we compute three meridional height gradients using latitudes 15° and 30° to the south and 15° to the north. A location is classified as blocked when these gradients indicate a reversal to easterly flow, with westerlies to the north and south.

4. Dataset

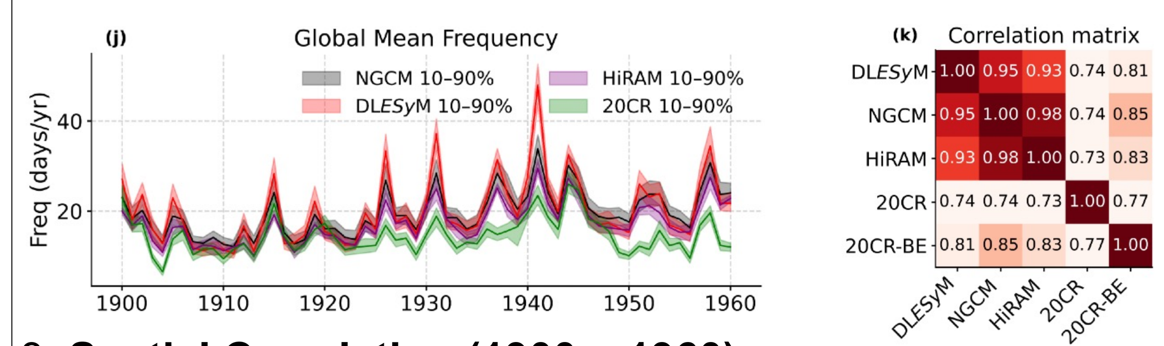
- **20th Century Reanalysis Version 3** (20CRv3; 1836–2015): 80 members, SST&SIC as boundary conditions and assimilate pressure data.
- **Berkeley Earth** (BE): Surface Temperature Instrumental Observations.

Land Heatwaves and Cold Waves

1. Heatwave Climatology – Models tend to overestimate frequencies

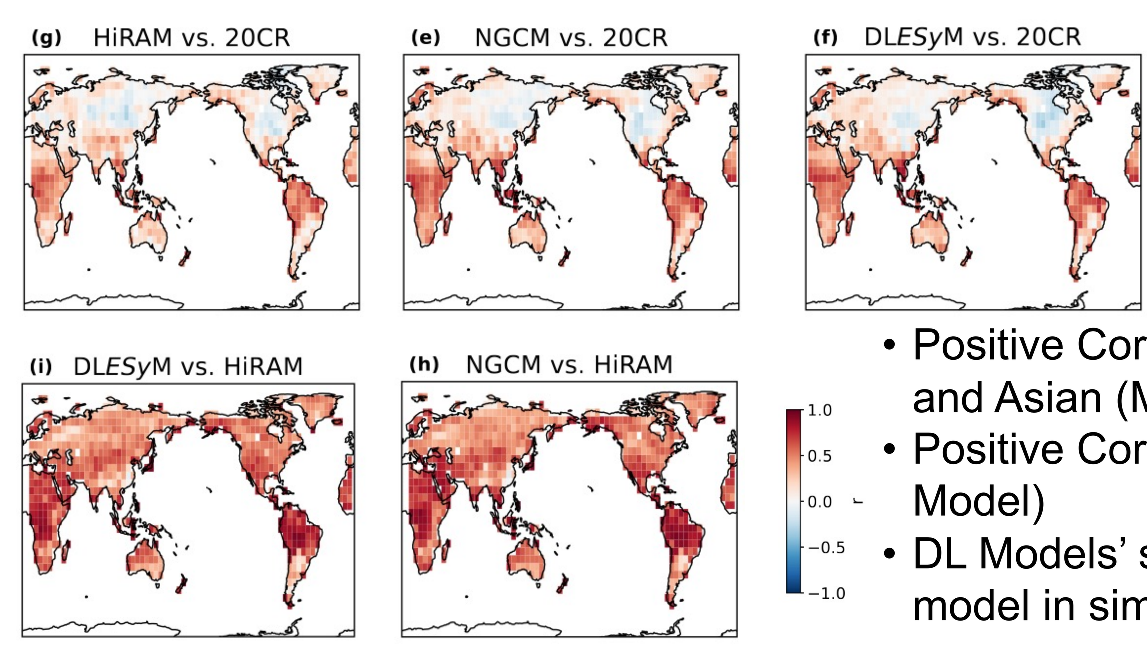


2. Global Mean Heatwave Frequencies



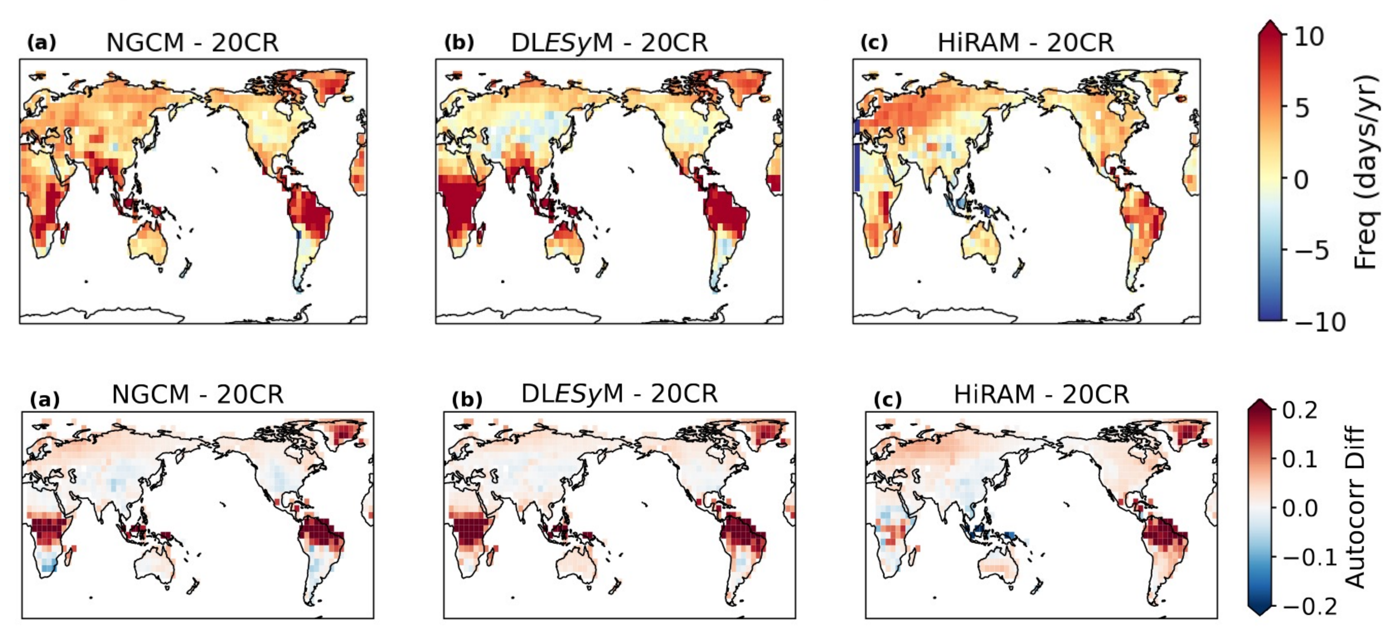
- High correlation for global mean frequencies.
- Models overestimate heatwave frequencies.
- Cold waves show similar behavior (not shown).

3. Spatial Correlation (1900 ~ 1960)



- Positive Correlation except North American and Asian (Model vs. 20CRv3).
- Positive Correlation everywhere. (Model vs. Model)
- DL Models' skill is comparable to a physical model in simulating heatwave and cold waves

4. Why the Models overestimate the frequencies relative to 20CRv3?



- Positive autocorrelation differences (vs. 20CR) align with frequency overestimation regions
- High bis in autocorrelation→ greater temperature persistence.
- Increased persistence raises the chance of exceeding the 3-day threshold for heatwave or cold wave events.

5. Main Takeaways for heatwave and cold waves

- How well can the frequency of heatwaves and cold waves be simulated by prescribing only observed SST and SIC?

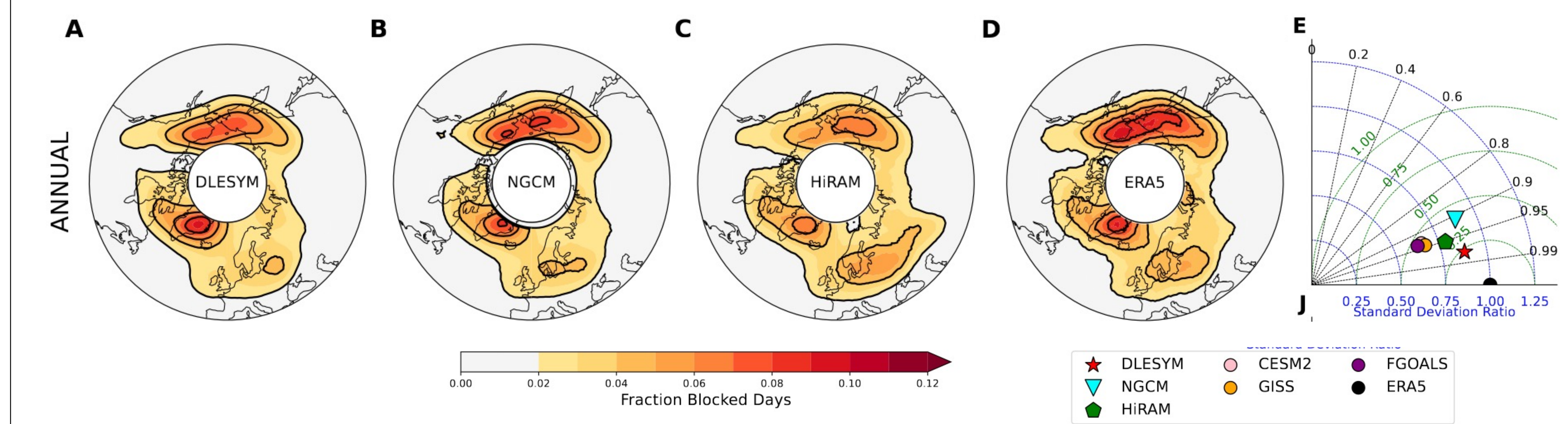
High correlation for global mean frequency; Positive correlation except North Asian and America (Reasons and discussion can be seen in the paper).

- How well do DL-based GCMs reproduce heatwave and cold wave frequencies during the out-of-sample period 1900–1960 when compared with reanalysis datasets and a traditional physical GCM?

DL-based GCMs successfully simulate out-of-sample (1900–1960) heat and cold wave frequencies with skill comparable to a physical model

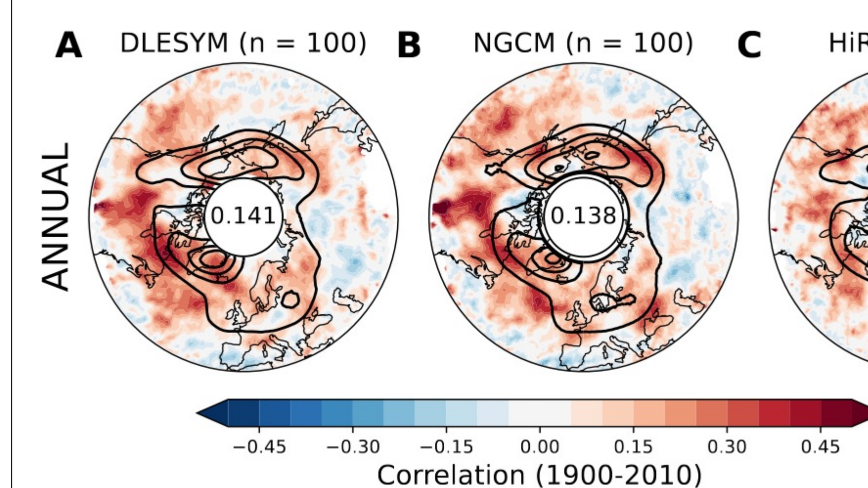
Atmospheric Blocking

1. Blocking Climatology (1980~2010)– (DL) models are more consistent with ERA5



2. Spatial correlations (vs. 20CR, 1900~2010)

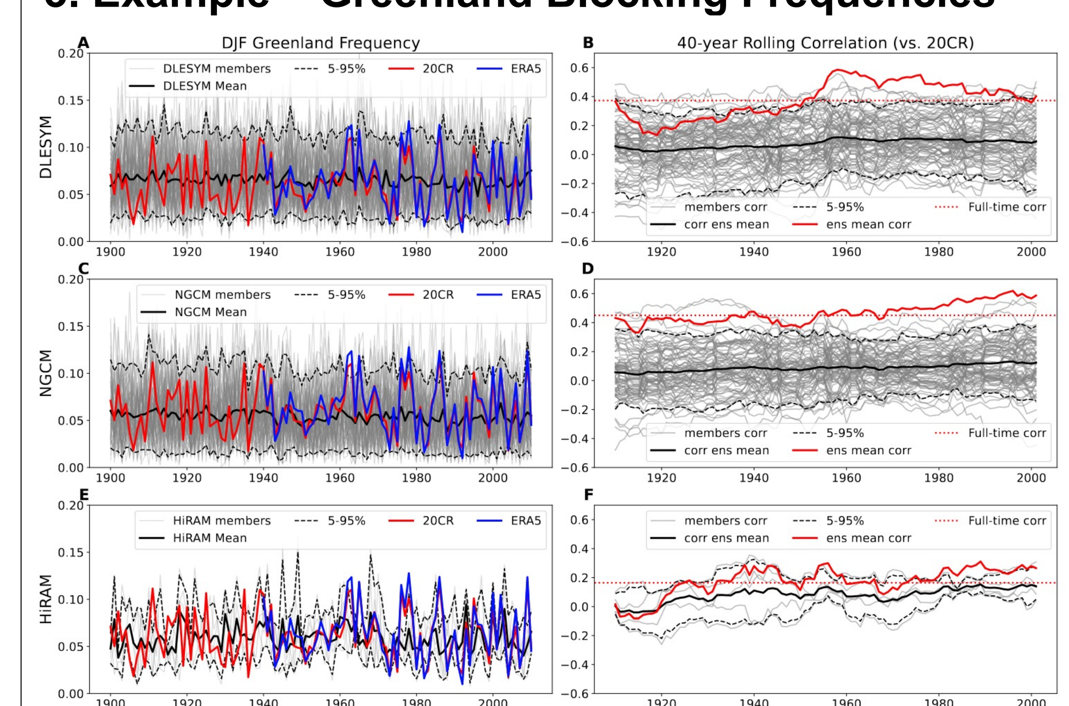
Central circle : NH-averaged correlation weighted by blocking frequency and area. Contour: Blocking Frequency Climatology.



- With larger ensemble, DL models outperform HiRAM — correlations higher by ~two orders of magnitude when compared the correlation with 20CRv3.
- Out-of-sample period (1900-1960) shows similar skills.

- Large ensemble simulations filter out internal atmospheric variability, thereby isolating the SST-forced signal, leading to high correlations (see Greenland Example)

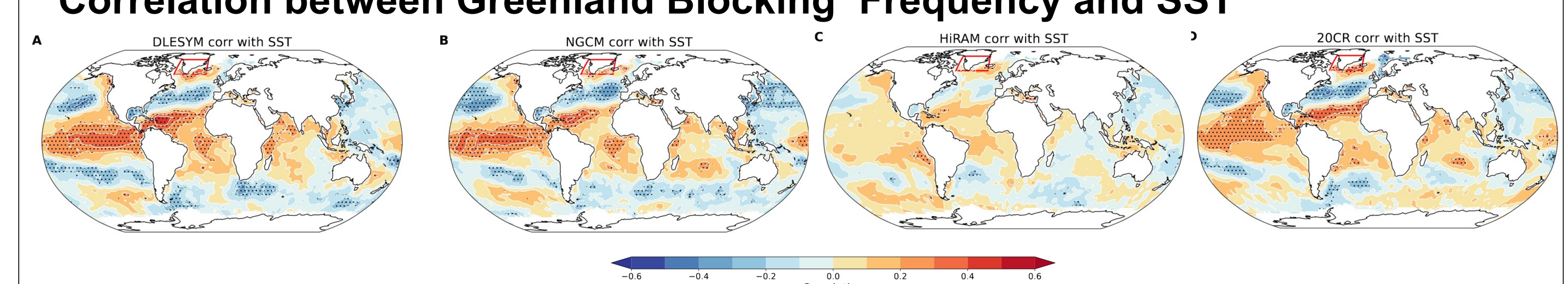
3. Example – Greenland Blocking Frequencies



Correlation (5-member HiRAM vs. reanalysis) = 0.14
Correlation (100-member DL AMIP vs. reanalysis) = 0.45

Large-ensemble (n = 100) DL models (DLESYM and NGCM) successfully capture the observed range of Greenland blocking variability, with ensemble means showing moderate correlations (~0.45) with reanalysis.

Correlation between Greenland Blocking Frequency and SST



DL models and reanalysis show a clear **North Atlantic tripole** and **El Niño–like** SST pattern linked to enhanced Greenland blocking, while HiRAM fails to capture these signals.

4. Main Takeaways for atmospheric blocking

- Deep learning (DL) models' climatology is more consistent with ERA5.
- Large ensemble simulations show SST-forced blocking frequency variabilities, correlated with 20CR.
- Large ensemble simulations filter out internal atmospheric variability, thereby isolating the SST-forced signal
- For atmospheric blockings, signal from the SST is smaller than the noise from atmospheric internal noise.