ClimateAi Data Science Challenge: Seasonal Forecasting of Summertime Precipitation in the Midwest US

Time investment expected: ~3-8 hours

Please do not hesitate to reach out to me at david.f@climate.ai or WhatsApp at +1-607-351-9339 if you have any questions or concerns.

At ClimateAi, we aim to push the boundaries of seasonal climate forecasting by integrating ML methodologies with expert domain knowledge. We value innovation, analytical rigor, and thoughtful communication. In this challenge, we are interested not just in the solution you propose, but also in understanding your reasoning and decision-making process.

The Challenge

The primary objective is to propose an approach (and, if time allows, produce preliminary results) for forecasting summertime (June, July, and August) precipitation in the Midwest US – a region and time of year critical for agriculture. The goal is to explore the use of a statistical, ML, and/or AI model that leverages the spatio-temporal predictor fields provided. The approach you outline / attempt to implement is very open-ended, though we are providing a set of oceanic and atmospheric predictors (described below). While the scope is open-ended, we are particularly interested in understanding the rationale behind your modeling decisions – even if some decisions are made primarily "for convenience and due to time constraints."

Data Provided

You will have access to two NetCDF (.nc files) here.

1. **Predictors File:**
    - Contains several relevant monthly oceanic and atmospheric and variables, including:
        - sst: Sea surface temperature
        - thickness: Tropospheric thickness between 850 and 500 hPa
        - pottemp: 45m ocean potential temperature
        - pr_wtr: Precipitable water
        - lftx4: Atmospheric lifted index
        - chi: upper troposphere velocity potential
        - tcdc: total cloud cover
        - shum: 2m specific humidity

2. **Predicted Variable File:**

    - Contains monthly precipitation data for the Midwest US, gridded across the region.

Expectations for Deliverables

Given the very limited time available, we are primarily interested in your approach and thought process rather than your results. Your submission should include:

1. **Exploratory Data Analysis (EDA):**
    - Briefly explore the provided data to uncover key insights that inform how you would approach constructing a seasonal forecasting model.
2. **Forecasting Approach & Rationale:**
    - Describe how you decide to approach the problem:
        - Will you forecast precipitation as a gridded field across the Midwest or average it over the region?
        - Are you forecasting all summer months together, or each month separately?
        - How many lead times (e.g., one month ahead, two months ahead) will you consider?
    - Please briefly explain your reasoning, even if most choices are driven by convenience or time constraints.
3. **Modeling Strategy & Preliminary Results:**
    - **Summarize the model(s) that you would attempt to implement (or do implement, if time permits) and provide a brief rationale for why you made this decision.**
    - **If time allows, provide preliminary results; otherwise, outline your intended modeling approach and how you would validate it.**
4. **Presentation of Findings:**
    - Prepare a brief (~15 minute) summary – in the form of an informal presentation – for our group discussion that covers:
        - Any key insights from your EDA, including plots to illustrate.
        - Your forecasting approach and rationale.
        - (If obtained) any results of your approach.
        - Assumptions, caveats, and areas for further development if given more time. (e.g., What other EDA might be useful to inform the modeling approach? Are there other predictors that you think could help?)

Evaluation Criteria

You will be evaluated (via submission & discussion) primarily along the following dimensions.

- **Analytical Rigor:** The quality of your exploratory analysis and how well you justify your modeling decisions.

- **Communication:** Your ability to clearly articulate your approach, reasoning, and any results.
- **Strategic Thinking:** Your approach to balancing exploratory work, conceptualization, and model fitting within the time constraint.
- **Technical Implementation:** The quality and reproducibility of any code or workflows you provide.

Submission Details

- Please submit your materials for our discussion – i.e., slides, code, etc, to Dave at david.f@climate.ai at least 1 hour prior to our scheduled discussion time.
- Our discussion will include at least three of Dave, Carlos Hoyos, Diego Alfaro, and Arik Tashie. (Carlos, Diego, and Arik are the senior-most members of ClimateAi's DS department.)

Appendix:
In case it is of interest, here is some additional information about the datasets.

Regarding the time, latitude, and longitude dimensions:
Time: While the reference date is different, all fields span Jan 1982 through Dec 2023 (you will see this when you unpack and convert to dates from days past the reference date.)

Latitudes: For both datasets, the units are degrees and north is positive and south is negative (i.e., (90) is 90 deg North, 0 is the equator, and (-90) is 90 deg South).

Longitudes: For both datasets, the units are degrees and east (of the Greenwich meridian) is positive and west (of the Greenwich meridian) is negative (i.e., Seattle, WA, is at a longitude of about -122.25)

The full names, source, and units of the predictor datasets are as follows:

sst: Sea surface temperature

- Source: NOAA/NCEI 1/4 Degree Daily Optimum Interpolation Sea Surface Temperature (OISST) Analysis, Version 2.1
- Units: degC

thickness: Tropospheric thickness between 850 and 500 hPa

- Source: NCEP-NCAR Reanalysis 1 Monthly Mean of Thickness
- Units: m

pottemp: 45m ocean potential temperature

- Source: NCEP GODAS Ocean Potential Temperature 45m
- Units: degK

pr_wtr: Precipitable water

- Source: NCEP-NCAR Reanalysis 1 Monthly Mean of Precipitable Water Content
- Units: kg/m^2

lftx4: Atmospheric lifted index

- Source: NCEP-NCAR Reanalysis 1 Monthly Mean Surface Lifted Index
- Units: degK

chi: upper troposphere velocity potential

- Source: NCEP-NCAR Reanalysis 1
- Units: m*m/s

tcdc: total cloud cover

- Source: NCEP-NCAR Reanalysis 1 Monthly Mean of Total Cloud Cover
- Units: %

shum: 2m specific humidity

- Source: NCEP-NCAR Reanalysis 1 Monthly Mean of Specific Humidity
- Units: grams/kg