Data Dictionary & Overview

The WiDS Datathon 2023 focuses on a prediction task involving forecasting sub-seasonal temperatures (temperatures over a two-week period, in our case) within the United States. We are using a pre-prepared dataset consisting of weather and climate information for a number of US locations, for a number of start dates for the two-week observation, as well as the forecasted temperature and precipitation from a number of weather forecast models (*we will reveal the source of our dataset after the competition closes*). Each row in the data corresponds to a single location and a single start date for the two-week period. **Your task is to predict the arithmetic mean of the maximum and minimum temperature over the next 14 days, for each location and start date.**

You are provided with two datasets:

1. `train_data.csv`: the training dataset, where `contest-tmp2m-14d__tmp2m`, the arithmetic mean of the max and min observed temperature over the next 14 days for each location and start date, is provided
2. `test_data.csv`: the test dataset, where we withhold the true value of `contest-tmp2m-14d__tmp2m` for each row.

To participate in the Datathon, you will submit a solution file containing the predicted values of `contest-tmp2m-14d__tmp2m` for each row in the test dataset. The predicted values you submit will be compared against the observed values for the test dataset and this will determine your standing on the Leaderboard during the competition as well as your final standing when the competition closes.

You are also provided with an example of a solution file prepared for submission.

Note: During the competition the leaderboard is calculated with approximately 50% of the test data. After the competition closes, the final standings will be computed based on the other 50%. As such, *the final leaderboard standings may be different than those during the competition*.

External Data Usage

The datathon task can be tackled ***successfully*** without the use of external data. In fact, the degree to which we have anonymized the data would make joining additional data to the competition data difficult. However, participants who wish to do so may use additional external data for the purpose of **building predictive models**.

Data Dictionary

The WiDS 2023 Datathon is using a subset of a pre-prepared dataset in which the variables were gathered from the following datasets (*source of the WiDS Datathon dataset will be revealed after the competition closes*):

- **Temperature**: Daily maximum and minimum temperature measurements at 2 meters from 1979 onwards were obtained from NOAA’s Climate Prediction Center (CPC) Global Gridded Temperature dataset and converted to Celsius. The official contest target temperature variable is `tmp2m = tmax+tmin / 2`. `ftp://ftp.cpc.ncep.noaa.gov/precip/PEOPLE/wd52ws/global_temp/`

- **Global precipitation**: Daily precipitation data from 1979 onward were obtained from NOAA’s CPC Gauge-Based Analysis of Global Daily Precipitation [42] and converted to mm. `ftp://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/RT/`

- **U.S. precipitation**: Daily U.S. precipitation data in mm were collected from the CPC Unified Gauge-Based Analysis of Daily Precipitation over CONUS. Measurements were replaced with sums over the ensuing two-week period. `https://www.esrl.noaa.gov/psd/thredds/catalog/Datasets/cpc_us_precip/catalog.html`

- **Sea surface temperature and sea ice concentration**: NOAA’s Optimum Interpolation Sea Surface Temperature (SST) dataset provides SST and sea ice concentration data, daily from 1981 to the present. `ftp://ftp.cdc.noaa.gov/Projects/Datasets/noaa.oisst.v2.highres/`

- **Multivariate ENSO index (MEI)**: Bimonthly MEI values (MEI) from 1949 to the present, were obtained from NOAA/Earth System Research Laboratory. The MEI is a scalar summary of six variables (sea-level pressure, zonal and meridional surface wind components, SST, surface air temperature, and sky cloudiness) associated with El Niño/Southern Oscillation (ENSO), an ocean-atmosphere coupled climate mode. `https://www.esrl.noaa.gov/psd/enso/mei/`

- **Madden-Julian oscillation (MJO)**: Daily MJO values since 1974 are provided by the Australian Government Bureau of Meteorology. MJO is a metric of tropical convection on daily to weekly timescales and can have a significant impact on the United States sub-seasonal climate. Measurements of phase and amplitude on the target date were extracted over the two-week period. `http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt`

- **Relative humidity, sea level pressure, and precipitable water for the entire atmosphere**: NOAA’s National Center for Environmental Prediction (NCEP)/National

Center for Atmospheric Research Reanalysis dataset contains daily relative humidity (rhum) near the surface (sigma level 0.995) from 1948 to the present and daily pressure at the surface (pres) from 1979 to the present. `ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis/surface/`

- **Geopotential height, zonal wind, and longitudinal wind:** To capture polar vortex variability, obtained daily mean geopotential height were obtained at 10mb from the NCEP Reanalysis dataset. `ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/pressure/`

- **North American Multi-Model Ensemble (NMME):** The North American Multi-Model Ensemble (NMME) is a collection of physics-based forecast models from various modeling centers in North America. Forecasts issued monthly from the `Cansips`, `CanCM3`, `CanCM4`, `CCSM3`, `CCSM4`, `GFDL-CM2.1-aer04`, `GFDL-CM2.5`, `FLOR-A06` and `FLOR-B01`, `NASA-GMAO-062012`, and `NCEP-CFSv2` models were downloaded from the IRI/LDEO Climate Data Library. Each forecast contains monthly mean predictions from 0.5 to 8.5 months ahead. `https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/`

- **Pressure and potential evaporation:** `ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis/surface_gauss/`

- **Elevation:** `http://research.jisao.washington.edu/data_sets/elevation/elev.1-deg.nc`

- **Köppen-Geiger climate classifications:** `http://koeppen-geiger.vu-wien.ac.at/present.htm`

Variable naming

Each variable name, `prefix__suffix`, consists of two parts (separated by a double underscore) that inform you of the meaning of the variable. The prefix indicates from which of the above-listed file the variable was derived (e.g. Madden-Julian oscillation, pressure, and potential evaporation from NOAA's `surface_gauss` etc), the suffix indicates the specific type of information that was extracted from the file.

Variable prefixes

- `contest-slp-14d`: file containing sea level pressure (slp)
- `nmme0-tmp2m-34w`: file containing most recent monthly NMME model forecasts for `tmp2m` (`cancm30`, `cancm40`, `ccsm30`, `ccsm40`, `cfsv20`, `gfdlflora0`, `gfdlflorb0`, `gfdl0`, `nasa0`, `nmme0mean`) and average forecast across those models (`nmme0mean`)
- `contest-pres-sfc-gauss-14d`: pressure
- `mjo1d`: MJO phase and amplitude
- `contest-pevpr-sfc-gauss-14d`: potential evaporation
- `contest-wind-h850-14d`: geopotential height at 850 millibars
    - `contest-wind-h500-14d`: geopotential height at 500 millibars
- `contest-wind-h100-14d`: geopotential height at 100 millibars
- `contest-wind-h10-14d`: geopotential height at 10 millibars
- `contest-wind-vwnd-925-14d`: longitudinal wind at 925 millibars
    - `contest-wind-vwnd-250-14d`: longitudinal wind at 250 millibars
- `contest-wind-uwnd-250-14d`: zonal wind at 250 millibars
- `contest-wind-uwnd-925-14d`: zonal wind at 925 millibars
- `contest-rhum-sig995-14d`: relative humidity
    - `contest-prwtr-eatm-14d`: precipitable water for entire atmosphere
- `nmme-prate-34w`: weeks 3-4 weighted average of monthly NMME model forecasts for precipitation
    - `nmme-prate-56w`: weeks 5-6 weighted average of monthly NMME model forecasts for precipitation
- `nmme0-prate-56w`: weeks 5-6 weighted average of most recent monthly NMME model forecasts for precipitation
- `nmme0-prate-34w`: weeks 3-4 weighted average of most recent monthly NMME model forecasts for precipitation
- `nmme-tmp2m-34w`: weeks 3-4 weighted average of most recent monthly NMME model forecasts for target label, `contest-tmp2m-14d__tmp2m`
- `nmme-tmp2m-56w`: weeks 5-6 weighted average of monthly NMME model forecasts for target label, `contest-tmp2m-14d__tmp2m`
- `mei`: MEI (`mei`), MEI rank (`rank`), and Niño Index Phase (`nip`)

- `elevation`: elevation
- `contest-precip-14d`: measured precipitation
- `climateregions`: Köppen-Geigerclimateclassifications

Variables without prefix

Some variables do not have a prefix. Instead, each variable name in its entirety indicates the information the variable captures.

- `lat`: latitude of location (anonymized)
- `lon`: longitude of location (anonymized)
- `startdate`: startdate of the 14 day period
- `sst`: sea surface temperature
- `icec`: sea ice concentration
- `cancm30`, `cancm40`, `ccsm30`, `ccsm40`, `cfsv20`, `gfdlflora0`, `gfdlflorb0`, `gfdl0`, `nasa0`, `nmme0mean`: most recent forecasts from weather models

Target

- `contest-tmp2m-14d__tmp2m`: the arithmetic mean of the max and min observed temperature over the next 14 days for each location and start date, computed as `(measured max temperature + measured mini temperature) / 2`