# ENVR Data Challenge 2021

This data challenge considers a rich collection of historical and future climate information from three different sources: 1) a global climate model, 2) a reanalysis data product, and 3) a dynamically downscaled product. Each of these different sources include daily values of maximum temperature, minimum temperature, and precipitation on different spatial scales over a swath of North America covering the continental United States of America (CONUS), much of southern Canada and northern Mexico. The future climate for both the global climate model and the downscaled product are based on RCP8.5, one of the four Representative Concentration Pathways (RCPs) describing 21st century greenhouse gas emissions and atmospheric concentrations, air pollutant emissions, and land use which are a part of the Intergovernmental Panel on Climate Change's Fifth Assessment Report (https://www.ipcc.ch/). RCP8.5 is considered a pathway with very high greenhouse gas emissions and is often referred to as "business as usual".

## The Datasets

These datasets are available at (be advised that these are very large files, over 230 GB):
https://www.dropbox.com/sh/f667nr0pe2nqubv/AADa3xVwIaegiqyPp6Q0kUyBa?dl=0

CESM-LENS

The CESM-LENS is a publicly-available set of climate simulations whose initial purpose was to examine internal climate variability and climate change. This ensemble is based on a 1-degree version of the Community Earth System Model version 1 (CESM1) with CAM2.5 as the atmospheric component. Daily values of maximum temperature, minimum temperature, and precipitation are included for each member of the 40-member ensemble. The time period covers 1920 to 2100, with the historical forcing up to 2005 and RCP8.5 after. Each ensemble member is subject to the same forcing but has a slightly different initial atmospheric state. The original global extent has been subsetted to a region covering CONUS and extending into Canada and Mexico. For more information, see the CESM-LENS website at

http://www.cesm.ucar.edu/projects/community-projects/LENS/

or the research paper:

https://journals.ametsoc.org/doi/full/10.1175/BAMS-D-13-00255.1

ERA-Interim

The ERA-Interim is a global atmospheric reanalysis dataset. Reanalysis is an approach to produce spatially and temporally gridded datasets via data assimilation for climate monitoring and analysis. ERA-Interim uses the 2006 release of the IFS (Cy31r2) and incorporates between 10^6 and 10^7 observations per day on average, mostly from satellite observations. Daily values of maximum temperature, minimum temperature, and precipitation from 1979 to 2017 are included. The spatial resolution is roughly 80km. As with the CESM-LENS, the original global extent has been subsetted to a region covering CONUS and extending into Canada and Mexico. For more information, see the ERA-Interim website at

https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim

or the research paper

https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.828

Liu et al. Dataset

The Liu et al. Dataset includes two 13-year Weather Research and Forecasting (WRF) model simulations at 4km over much of North America. WRF is a next-generation mesoscale numerical weather prediction system that is used both for atmospheric research and operational weather forecasting. For the historical simulation, the time period is from October 1, 2000 through September 30, 2013, and boundary conditions based on ERA-Interim were used. The second simulation could be thought of as perturbation or climate sensitivity experiment. This simulation uses the same time period as the historical simulation, although the boundary conditions in this case are from ERA-Interim plus a climate perturbation that represents the 95-year CMIP5 multi-model ensemble-mean change signal under the RCP8.5 emission scenario. As in the other two datasets, daily maximum temperature, minimum temperature, and precipitation are provided. For more information, see

Liu, C., Ikeda, K., Rasmussen, R. et al. 2017: Continental-Scale, Convection-Permitting Modeling of the Current and Future Climate of North America. Clim. Dynamics, 49: 71. Doi: 10.1007/s00382-016-3327-9

## Tools for working with Netcdf files

The datafiles here are all in netcdf. Netcdf is a self-describing, machine-independent format for representing scientific data and is ideal for the gridded datasets common in the geosciences. There are numerous resources for reading and visualizing netcdf files. See the website at Unidata: https://www.unidata.ucar.edu/software/netcdf/.

- R: There are several packages to read and write netcdf files in R: ncdf4, tidync, raster, etc.
    - https://cran.r-project.org/web/packages/ncdf4/index.html
    - https://cran.r-project.org/web/packages/tidync/index.html
    - https://cran.r-project.org/web/packages/raster/index.html
- Python: Based on pandas, the xarray package is designed to work with netcdf files and is a part of the Pangeo community.
    - http://xarray.pydata.org/en/stable/
    - https://pangeo.io/
- Others: The packages ncview, panopoly, etc. can be used to visualize the data in netcdf files and are useful for quick exploratory looks at the metadata and the data in netcdf files.
    - http://meteora.ucsd.edu/~pierce/ncview_home_page.html
    - https://www.giss.nasa.gov/tools/panoply/

## Potential Research Topics

These topics are intended only as a guide. You are not limited to these particular questions, and, in some cases, there might be a reason to combine different elements from these topics as well as additional datasets. Some general things to consider include choice of variable (i.e., temperature or precipitation separately or both simultaneously), extremes, or even location (i.e., central plains versus mountain west, etc.). And, of course, uncertainty estimates are always an important aspect.

- As with any climate model experiment, a key goal is an assessment of the magnitude of climate change and the uncertainty associated with that assessment.
- Along with assessments of climate change is the impact that such changes could have in a variety of sectors including energy, pollution, health, ecology, etc. Coupled with additional outside information, this dataset could be useful to assess climate change impacts. For example, natural resource managers and biologists are interested in understanding correlations among historic and predicted climate with biological measurements in order to inform potential management actions (see a sample list of additional datasets below).
- While global climate models are useful tools for studying the Earth's climate, they are often run on larger spatial scales that lead to limited utility for many impacts studies that require more local information. Downscaling refers to different techniques that relate local or regional spatial scale climate variables to larger modeled or observed data. These techniques are often grouped into two categories: dynamical and statistical. Dynamical downscaling utilizes climate models, while statistical downscaling utilizes statistical or machine learning methods. With datasets on different spatial scales, the dataset could be used to facilitate the development of new downscaling methods or comparing the performance of a suite of methods.
- Analog methods are also used for climate downscaling. The idea is relatively simple. The large-scale spatial pattern from a GCM is compared to elements in a catalog of historical observations and the most similar element in the catalog is chosen as the analog. The simultaneous observed local weather for the analog is then associated with the large-scale pattern from the GCM. There are many facets of this problem to examine, including how to measure similarity and the choice of most similar element or elements.
- Climate models are important tools for studying how the Earth's climate responds to different forcings. An alternative to deterministic models is so-called weather generators that use statistical or generative models to produce stochastic realizations of realistic weather streams. This dataset could be used to train different algorithms for weather generators and explore their utility for assessments of climate change.

## Additional Datasets

Some additional datasets that may be of interest.

GHCN (daily meteorology observations): https://www.ncdc.noaa.gov/ghcn-daily-description

NLCD (land use) - https://www.usgs.gov/centers/eros/science/national-land-cover-database?qt-science_center_objects=0#qt-science_center_objects

https://irma.nps.gov/Portal/ NPS Database

https://ecos.fws.gov/ServCat/ USFWS Database

https://www.usanpn.org/usa-national-phenology-network go to tab "Data" Natl Phenology Network of observational data

https://www.northernwater.org/WaterQuality/WaterQualityData.aspx

https://www.waterqualitydata.us

https://data.cnra.ca.gov

https://iridl.ldeo.columbia.edu/index.html?Set-Language=en (many different climate and impacts related data sets)

https://catalog.data.gov/dataset/waterfowl-breeding-population-survey (breeding population survey; long-term survey of migratory waterfowl)

https://www.pwrc.usgs.gov/bbs/ (breeding bird survey, long-term bird monitoring survey)

https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (public use U.S. Census data)

https://www.fia.fs.fed.us/ (forest inventory analysis; US forest census)