Data Processing Document

Maforikan Amoussou

April 4th, 2025

1 Data Sources

GOES (Geostationary Operational Environmental Satellites) are a series of U.S. weather satellites operated by NOAA that provide continuous observations of the Western Hemisphere. The two satellites used in this project, GOES-16 (positioned over the eastern U.S.) and GOES-17 (covering the western U.S. and Pacific), are part of the GOES-R series. These satellites use the GOES-R Advanced Baseline Imager (ABI) and the Geostationary Lightning Mapper (GLM) to collect observational weather data. The ABI provides weather data such as cloud top height, CAPE, rain fall estimates, and more with a spatial resolution of 0.5 to 2 km and a temporal resolution of 5 to 10 minutes. The GLM detects lightning flash extent density across the Americas at 8–10 km resolution with a minutely time resolution.

ERA5, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), is a high-resolution global climate reanalysis dataset covering data from 1950 to present. It combines model output with observations using data assimilation to produce consistent historical datasets. ERA5 provides atmospheric variables on both single levels (e.g., surface) and pressure levels (e.g., 450 hPa). It offers hourly data with a spatial resolution of roughly 31 km ($0.25^\circ \times 0.25^\circ$ grid). ERA5 is widely used in climate and weather research due to its accuracy, completeness, and temporal depth.

The data acquisition pipeline uses GLM cleaned data from the University of Maryland and ERA5 data from the Climate Data Store. (Note that another possibly useful source of GLM flash extent density data is the NASA EarthData repository.)

2 Data Acquisition and Processing

The data_retreiver_2.py file is a pipeline for retrieving, processing, and regridding lightning data from the GOES-16 satellite, specifically the Flash Extent Density (FED) product. It uses concurrent file downloads and multiprocessing to download a full day's worth of minutely satellite files. Each hourly chunk of FED data is downloaded in 5-minute intervals and aggregated into hourly values using parallel processing. The data, initially in GOES satellite projection, is transformed to geographic coordinates (latitude and longitude) using a projection algorithm enhanced with Numba's Just In time calculations for speed.

To enable integration with other datasets such as ERA5, the FED data is interpolated onto a target grid using spatial nearest-neighbor interpolation. Nearest neighbor is best because it is the most efficient for this type of analysis, and the least computationally expensive. Interpolation is needed to make meaningful comparisons between ERA5 and GLM data because ERA5 is a lower resolution dataset both in time and space. The code also includes an ERA5 downloader based on the CDS API requests for both pressure-level and single-level variables, which are merged for compatibility with the lightning dataset. The final goal of this workflow is to produce a time-aligned, spatially matched dataset that can be used for machine learning, statistical analysis, or visualization. Overall, the code is highly optimized for batch processing and large-scale geospatial data integration.

The key functions of this pipeline are retrieve_goes_glmf() and parallel_interp(). The remaining functions are helpers or specialized for other datasets (ABI/ERA5). Example usage of the key functions is available in the repository.

3 Data Access

To access and manipulate processed data, familiarity with Xarray, Pandas, NumPy, and Dask is needed. For large Datasets (3+GB) efficient manipulation requires Dask chunking.

`https://github.com/amousso3/LightningResearch2024/tree/main`
`https://lightningdev.umd.edu/feng_data_sharing/`