Wild Card Project: Solar Radiation Prediction
Data Preprocessing

Farshad Ghorbanishovaneh

University of Wyoming

Introduction

In this project, we aim to predict solar radiation using the GOES 16 dataset[1] and the National Solar Radiation Database (NSRDB[2].) By leveraging these comprehensive datasets, the project seeks to explore the potential of various machine learning algorithms in accurately forecasting solar radiation levels. The current phase of the project involves preparing and structuring the data, which will serve as a foundation for the predictive models to be developed in subsequent stages.

Utilizing solar radiation data from these sources, our goal is to enhance the accuracy and efficiency of solar radiation predictions. This endeavor not only contributes to the field of meteorology but also supports applications in solar energy management and agricultural planning. As we progress, we will compare the performance of different algorithms to identify the most effective approach for solar radiation prediction. This initial preparation phase is critical as it sets the groundwork for our analytical and predictive tasks ahead.

The project deals with time-series data, focusing on short-term predictions. The NSRDB dataset is designated as our target data due to its relevance and comprehensive coverage. Due to the large size of the datasets, direct inclusion in our repositories is not feasible; however, both datasets are publicly available for access. We plan to use the Temporal Fusion Transformer (TFT) for train our model. Additionally, we aim to employ automated machine learning techniques, including Hyperparameter Optimization (HPO) and pipeline optimization, to refine and enhance our predictive modeling capabilities which it will be in the next phase of the project.

Datasets

GOES 16 Dataset

The Geostationary Operational Environmental Satellite 16 (GOES 16), also known as GOES-East, is part of NOAA's latest series of geostationary weather satellites and primarily serves the Continental United States (CONUS). It provides continuous imagery and atmospheric measurements of Earth's Western Hemisphere. The GOES 16 dataset is particularly valuable for meteorological research due to its high spatial and temporal resolution. Offering near-real-time data capture, GOES 16 provides images at intervals as frequent as every 5 minutes for certain products.

[1] `https://noaa-goes16.s3.amazonaws.com/index.html`

[2] `https://data.openei.org/s3_viewer?bucket=nrel-pds-nsrdb&prefix=conus%2F`

This capability allows for detailed tracking of dynamic weather patterns and solar irradiance changes. The spatial resolution varies by the type of sensor and the data product, with the Advanced Baseline Imager (ABI) sensor offering a resolution of 2 km for various channels. The dataset includes types of data such as visible and infrared imagery, solar ultraviolet measurements, and derived products like cloud and moisture imagery. These features make it highly useful for monitoring and predicting severe weather events, tracking atmospheric and cloud dynamics, and studying solar radiation, particularly in understanding how cloud cover and atmospheric conditions affect solar irradiance at the Earth's surface.

National Solar Radiation Database (NSRDB)

The NSRDB is a free tool provided by the U.S. Department of Energy and operated by the National Renewable Energy Laboratory (NREL). It offers hourly solar radiation and meteorological data for the United States and a growing list of international locations. The dataset includes historical data typically spanning over 30 years, providing hourly solar radiation measurements with a spatial resolution available for a grid of 2 km x 2 km. This allows detailed geographical analysis of solar conditions. It includes measurements such as Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DIF). Additional meteorological parameters like temperature, dew point, and wind speed are also available, making it ideal for applications in solar energy system design and validation, climate research and weather forecasting, and agricultural planning and building energy management.

Both datasets are integral to the project as they provide the necessary data inputs for building and validating models that predict solar radiation. The rich detail and breadth of the data enable comprehensive analysis and modeling efforts, which are crucial for achieving accurate solar radiation predictions.

Expremental Setup

The experimental setup for the Solar Radiation Prediction Project begins with data preparation, which is crucial for ensuring the quality and consistency of inputs into the predictive models. The project utilizes several Python libraries such as Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for plotting, and others like h5py and xarray for handling specific dataset formats (HDF5 and netCDF). The setup process involves initializing a Python virtual environment and installing dependencies from a requirements.txt file to ensure reproducibility.

Directories are established to organize the large data files efficiently, including specific paths for the NSRDB data and GOES 16 data, and their respective processed outputs. Scripts automate the copying of data from shared disks to local directories and resample datasets to align them with the analysis' temporal resolution. Data from NSRDB and GOES 16 is filtered based on specific criteria such as dates and geographical locations, including temporal filtering to narrow down the dataset to specific dates for targeted analysis and spatial filtering by geographical criteria, such as within certain longitude and latitude bounds, to focus on areas like Wyoming.

Relevant metadata is extracted and saved, which includes information like country, state, county, latitude, longitude, elevation, and time zone. The feature engineering phase involves creating meaningful variables from raw data that can be used to train the predictive models. This includes converting string time indices to datetime objects to facilitate time-based analysis and combining solar radiation data (GHI) with location metadata to produce comprehensive datasets ready for analysis.

Visual analysis is conducted to understand the data distribution and quality, including creating scatter plots of solar irradiance against geographical coordinates to visualize how solar irradiance distributes across different regions. Processed data is saved in structured formats (like CSV) for use in modeling, and after processing, the data is saved in various formats, which might include separate files for different states or dates, ensuring easy accessibility for model training. Initial exploratory data analysis is performed to check the adequacy of the data processing steps and set the stage for more complex analytical tasks, including loading and displaying data to examine the contents and structure of the processed data, ensuring that everything is in order before moving into the modeling phase.

Challenges

The Solar Radiation Prediction Project, while promising, faces several significant challenges related to the nature and format of the data used. These challenges impact various stages of the project, from initial data handling to the final modeling phase. Data Size and Complexity

One of the primary challenges stems from the sheer volume of the datasets involved. Both the GOES 16 and NSRDB datasets are extremely large, encompassing high-resolution data captured over extended periods. Managing such vast amounts of data requires substantial computational resources and efficient data processing algorithms to ensure timely and effective analysis.

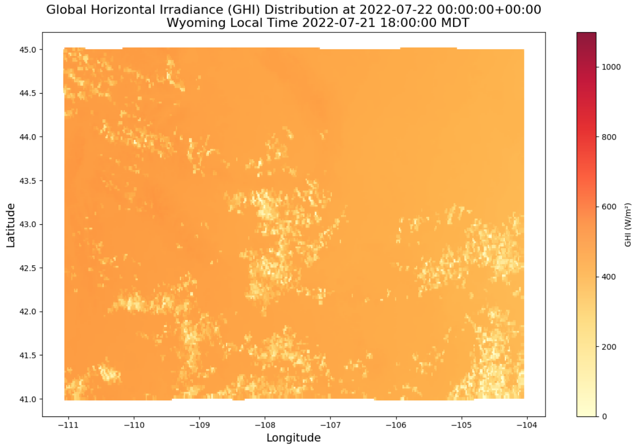

Figure 1: *NSRDB data, this figure shows the structure of the data from the NSRDB dataset specifically for the state of Wyoming.*

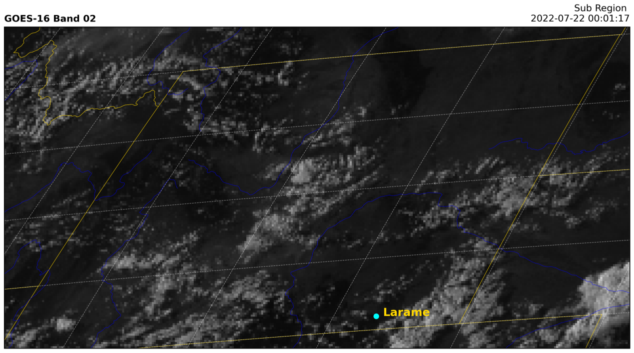

Figure 2: *GOES 16 data, this figure shows the structure of the data from the GOES 16 dataset for Band 2 (Red Visible) for a specific time point.*

The structure of the data from the two primary sources, NSRDB and GOES 16, differs significantly, adding complexity to data integration and preprocessing. The NSRDB dataset is relatively flat (Figure 1) and includes geographical coordinates, facilitating direct geographical analyses and mapping. Each record is easily accessible and corresponds to a specific geographic location and time point, making it straightforward for merging based on geographical and temporal

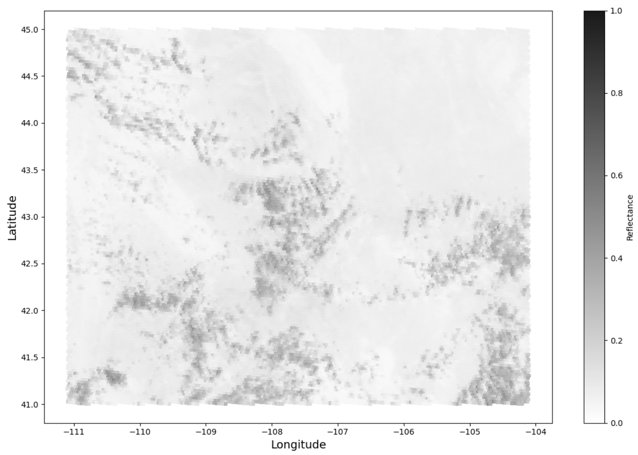

Figure 3: *GOES 16 data, this figure shows the structure of the data from the GOES 16 dataset for Band 2 (Red Visible) for a specific time point after flattening the data.*

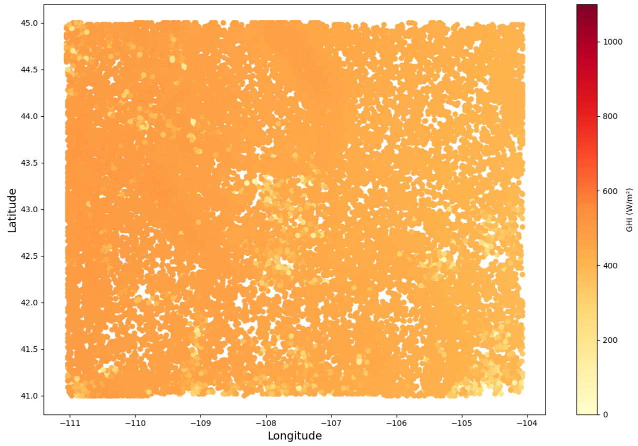

Figure 4: *Downsampling, this figure shows the process of downsampling the NSRDB data to match the temporal resolution of the GOES16 data and used K-means clustering to group the data points.*

keys. In contrast, The GOES 16 dataset inherently employs a geostationary projection, presenting data in a grid format based on the satellite's fixed position relative to the Earth (Figure 2) and includes data points in a projected coordinate system (x and y coordinates related to the satellite's perspective), which does not directly correspond to geographic coordinates. This necessitates an additional transformation step to convert the x and y coordinates into latitude and longitude. The conversion not only requires additional computational steps but also introduces potential sources of error and complexity in aligning this data with the NSRDB data (Figure 4).

Ensuring consistency across the datasets during data preprocessing includes aligning data points from different sources in time and space, resampling data to match resolutions, and normalizing features to ensure they are on comparable scales (Figure 1). Each of these steps must be handled carefully to avoid introducing biases or errors that could compromise the accuracy of the predictive models.

These challenges require thoughtful solutions and careful handling to ensure that the project can successfully predict solar radiation using the integrated data from the NSRDB and GOES 16 datasets.