Intermediate Project Report on ERA5-based Heatwave Prediction

Jonas Thalmeier, Ana Martinez

December 20, 2024

Abstract

This report describes the intermediate steps, data selection, preprocessing, and initial modeling attempts for a machine learning project aimed at predicting heatwaves in Spain based on ERA5 climate data. The chosen approach involved restricting spatial coverage to the Iberian Peninsula (Spain) and focusing on daily temperature observations at 00:00 and 12:00 UTC. Multiple atmospheric and surface variables were considered. The data were processed to identify hot days according to a percentile-based threshold and then aggregated into a rolling window structure suitable for supervised learning. Preliminary machine learning and deep learning models were tested, and a reduction in the percentile threshold was introduced to achieve a more balanced dataset before future adjustments to stricter criteria.

1 Introduction

Effective identification and prediction of heatwaves are critical for public health, agriculture, and energy management. To accomplish this, a dataset containing atmospheric measurements was downloaded from the Copernicus Climate Data Store (CDS)[1]. The dataset in use is ERA5, which provides hourly atmospheric reanalysis data with a high temporal and spatial resolution. As part of this project, steps were taken to extract spatially and temporally restricted subsets of ERA5 data and to apply percentile-based definitions of heatwaves.

This report details the intermediate steps undertaken: data acquisition, preprocessing, feature labeling, and initial approaches to building machine learning models for predicting future heat events. The approach followed best practices in data handling and model preparation, ensuring that the methodology can be refined and improved in subsequent phases.

2 Data Acquisition and Preprocessing

2.1 Dataset Selection and Constraints

The ERA5 dataset was accessed through the Climate Data Store after creating an account on the platform. Due to storage and download constraints, the spatiotemporal coverage

[1] `https://cds.climate.copernicus.eu/`

had to be limited. Therefore, only data covering the territory of Spain were selected. Furthermore, since full hourly coverage would be voluminous, only two daily time points (00:00 and 12:00 UTC) were chosen. The chosen variables included:

- 2m temperature (t2m)
- 2m dewpoint temperature (d2m)
- Mean sea level pressure (msl)
- Total cloud cover (tcc)
- Volumetric soil water layer 1 (swvl1)
- 10m u-component of wind (u10)
- 10m v-component of wind (v10)

Initially, the dataset encompassed all months of the year, with the intention of capturing a wide range of meteorological conditions. It was later observed that including all months, even those unlikely to produce heatwaves, was not the most efficient approach and that restricting to relevant months might have been more prudent.

2.2 Temporal and Spatial Restructuring

After downloading the data in GRIB format, standard Python libraries such as `xarray` and `pygrib` were employed to read and handle the data. The data were initially structured by time and location. To facilitate the learning task, measurements at 00:00 and 12:00 for the same day and location were merged so that they appeared as features side-by-side on a single row. This reformatting step simplified the input to the machine learning model, ensuring that each daily record included both early morning and midday conditions.

3 Definition of Hot Days and Heatwaves

The Spanish meteorological service defines a heatwave as a sequence of at least three consecutive days during which the maximum daily temperature at a given location exceeds the 95th percentile of the daily maximum temperature distribution for July and August at that location. To implement this criterion, the temperatures at 12:00 for July and August were extracted from the ERA5 dataset. The 95th percentile threshold was then computed for each grid point (latitude-longitude pair).

Subsequently, any day whose 12:00 temperature exceeded the computed local 95th percentile threshold was labeled as a “hot day.” Rolling computations were carried out to identify sequences of three or more consecutive hot days (i.e., potential heatwave events). These computations confirmed the occurrence of multi-day hot periods in certain locations and provided an initial assessment of the frequency and distribution of heatwave events.

4 Labeling and Feature Engineering

To train a predictive model, labels indicating whether a heatwave would occur in the near future were needed. For each 30-day period (rolling window of consecutive days), a binary label was assigned: 1 if at least 3 hot days would occur in the following 7 days, and 0 otherwise. This labeling strategy provided a supervised learning setup, enabling the prediction of imminent heatwave conditions based on recent weather patterns.

To facilitate machine learning, the data from all consecutive 30-day periods were stacked so that each record included the full sequence of daily measurements for those 30 days. At this stage, a challenge arose: only about 2.09% of the labels were positive (indicating a future heatwave), resulting in a highly imbalanced dataset. To address this issue, the threshold temperature was temporarily lowered (using a lower percentile) to increase the proportion of positive labels and thus create a more balanced training set. This balanced scenario simplifies initial model development. The threshold will be raised again at a later stage once models have been tuned to handle the more complex and realistic scenario.

5 Preliminary Modeling Approaches

Initial attempts to predict heatwaves involved testing several machine learning models:

1. A simple neural network (NN) with fully connected layers was trained on the transformed dataset.
2. A deeper neural network architecture, employing additional layers and dropout, was also tested.
3. A random forest classifier was applied for baseline comparison.

All models were trained using the preprocessed ERA5 features and the binary labels indicating future heatwave conditions. However, due to the extreme label imbalance and the intrinsic complexity of predicting rare events, preliminary results showed limited differentiation from a trivial classifier. Most models simply predicted the majority class, achieving high recall but low precision and consequently moderate F1 scores.

The code snippet below outlines an example of loading the data, performing initial transformations, and training a neural network model. For brevity, only select steps are shown:

```
import xarray as xr
import pygrib
import numpy as np
import torch
import torch.nn as nn
from torch.utils.data import DataLoader, TensorDataset

# Example: Loading ERA5 data and selecting Spain region,
# restricting to 00:00 and 12:00 UTC times, merging them,
# and computing hot day labels.

data = xr.open_dataset('./era5_spain.grib', engine="cfgrib")
data_summer = data.sel(time=data["time"].dt.month.isin([5,6,7,8]))
```

```
# Filter times to 0:00 and 12:00, merge them into a single daily record, etc

# Compute 95th percentile for July-Aug temperatures:
temp_july_aug = data_summer["t2m"].sel(day=data_summer["time"].dt.month.isin
percentile_95 = temp_july_aug.quantile(0.95, dim="day")

# Label hot days:
hot_days = (data_summer["t2m"].sel(time=data_summer["time"].dt.hour == 12) >
# Further steps: create rolling windows, assign binary labels for future 7-d

# Example NN model definition and training:
class HotDayPredictor(nn.Module):
    def __init__(self, input_size):
        super(HotDayPredictor, self).__init__()
        self.fc = nn.Sequential(
            nn.Linear(input_size, 128),
            nn.ReLU(),
            nn.Linear(128, 64),
            nn.ReLU(),
            nn.Linear(64,1),
            nn.Sigmoid()
        )
    def forward(self, x):
        return self.fc(x)

# After preparing X_train, y_train:
# model = HotDayPredictor(input_size)
# optimizer = torch.optim.Adam(model.parameters(), lr=0.001)
# loss_fn = nn.BCELoss()
# Training loop ...
```

5.1 Comparison of Configurations

The models were evaluated using two different configurations: the original setup with a 30-day input window and 3 hot days in 7 days, and a revised setup with a 60-day input window and 2 hot days in 5 days. For the Random Forest model, the performance metrics were:

- **Original Configuration:** Accuracy: 0.2391, Precision: 0.0429, Recall: 0.7940, F1 Score: 0.0813.
- **Revised Configuration:**
  - Threshold 0.3: Accuracy: 0.0637, Precision: 0.0637, Recall: 1.0000, F1 Score: 0.1197.
  - Threshold 0.5: Accuracy: 0.1346, Precision: 0.0655, Recall: 0.9492, F1 Score: 0.1226.

The revised configuration showed marginal improvement in F1 Score due to its higher recall. However, it introduced many false positives, as evidenced by the low accuracy. The 60-day input window likely provided additional historical context, improving sensitivity to potential heatwaves, but further tuning is needed to balance precision and recall.

6 Conclusions and Future Work

This intermediate report illustrates the complexity of defining and predicting heatwaves using ERA5 climate data. The steps described include:

- Acquiring and filtering ERA5 data to a specific region (Spain) and temporal pattern (00:00 and 12:00 UTC).
- Defining hot days based on the 95th percentile of July/August temperatures.
- Constructing a machine learning dataset with rolling windows and binary labels indicating imminent heatwaves.
- Addressing the severe class imbalance by lowering the threshold temporarily.
- Testing with other windows sizes (60-day input and 2 hot days in 5 days) which demonstrated higher recall but at the cost of significantly lower accuracy.

Preliminary modeling attempts, including neural networks and random forests, have shown the challenges inherent to the problem, particularly due to label imbalance. Future work will focus on:

- Creating an ERA5 dataset better tailored to the requirements and incorporating a greater number of data points
- Refining the feature engineering pipeline.
- Exploring advanced balancing techniques or more sophisticated modeling approaches (e.g., anomaly detection or tailored loss functions).
- Gradually restoring the percentile threshold to the originally intended 95th percentile and evaluating model performance under more stringent and realistic conditions.
- Exploring the possibility of predicting droughts as a future application

This initial exploration lays the groundwork for more targeted modeling and hyperparameter tuning. Further improvements and refinements are expected to improve predictive performance and produce a robust and reliable heatwave forecasting model.