Alpine Climate Data Challenge

Noise Enhanced Students
Alessandro Catalano, Gabriele Lo Cascio, Gabriele Lo Milia

02/03/2025

Introduction

Main Task

The main task is to develop a custom Climate Model for Val di Susa and Val de Maurienne. To better understand and explain our model, the results will be manged with interactive instruments and visual contents.

First Steps

We start by looking for the state-of-art approach to Climate Models. Since there lots of article about this topic, we decided to focus on:

- Climate Model;
- Climate Change;
- Evaluate a Climate Model;
- Global and Regional Model;
- Time Series;
- Reanalysis.

According to [1] we report:

"General circulation models (GCMs) are the foundation of weather and climate prediction. GCMs are physics-based simulators that combine a numerical solver for large-scale dynamics with tuned representations for small-scale processes such as cloud formation. Recently, machine-learning models trained on reanalysis data have achieved comparable or better skill than GCMs for deterministic weather forecasting."

According to [2] we report:

"Reanalysis data provide the most complete picture currently possible of past weather and climate. They are a blend of observations with past short-range weather forecasts rerun with modern weather forecasting models. They are globally complete and consistent in time and are sometimes referred to as 'maps without gaps'."

Our search continue and we find an article named "21st Century alpine climate change" [3]. This is our main reference, since the approach and the analysis are really weighty with the task.
So far we know:

- What is a Global Climate Model (GCM) and Regional Climate Model (RCM);
- Forecasting is done by combining both physics numerical simulation and machine learning models;
- Importance of features like temperature, precipitation, snow, sea temperature and anthropogenic factors.

We note thate GCM needs a large quantity of data and lots of simulations to better forecast and even if RCM needs less data, they still remains a lots. We read about ERA5 dataset and other Global or European datasets, but they are so big to download and have so many features to analyze.
So, base on [3], we decided to focus our analysis on a specific geographic area. In this way we reduce the amount of data and also create a more custom Climate Model for the valleys. Always base on the work of [3] we decide to work with data from 1980 to 2024.

Dataset Collection

Now we know what we want: a dataset having weather features based on what we read and not so big as ERA5. Our choice is to work with data from IOWA STATE UNIVERSITY (https://mesonet.agron.iastate.edu/request/download.phtml). These data are airport weather observations from around the world. First we look for the nations near the valleys and manually download all the stations for them. Then we aggregate all these stations together and apply the filter discussed above: geographic and time. Finally we download the data for each station that matchs our constrains.

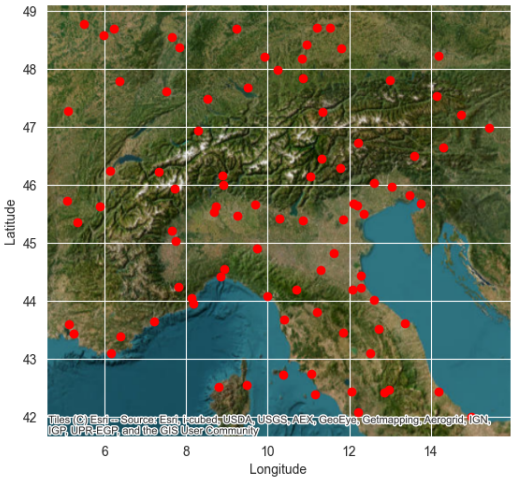

Figure 1 - Stations in our geographic area

Data Preparation

Filtering

First of all we convert all feature dimensions in S.I dimensions. Then we remove columns having all NaN values. Unfortunately we noticed there are some missing days in our datasets and in some of them these missing days become a lot. We decided to handle only datasets which less than 500 consecutive days missing. So we will work with 61% of our initial downloaded datasets. For these datasets we are going to impute missing values.

Imputation

For each dataset we are going to do imputation by following these steps:

1. Remove useless features (just for example a few of them: raw_metar, first_cloud_layer_cover, second_cloud_layer_cover, third_cloud_layer_cover);
2. Take a part only numeric columns and the station ID;
3. Aggregate by date and apply the mean (so for a day we consider mean of all hours values);
4. Since, as discussed before, there are some missing dates, we create a list of dates between 1981-01-01 to 2025-02-21;
5. Insert the missing dates in our aggregate dataframe and fill the row with NaN values for all the columns;
6. Impute missing values using KNNImputer.

Finally we ensemble all these single datasets in a big one.

Region Analysis

Base on what we read in [3], we decide to divide our stations into three groups:

- South (S);
- North-West (NW);
- North-East (NE).

In particular we see the regions in the paper and tryed to replicate them in our case. Furthermore we watch out about number of stations between NW and NE by following this approach:

- South Region if latitutde $<= 46.25$;
- North-West Region if longitude $<= 11$ and latitutde > 46 ;
- North-East Region if longitude > 11 and latitude > 46.

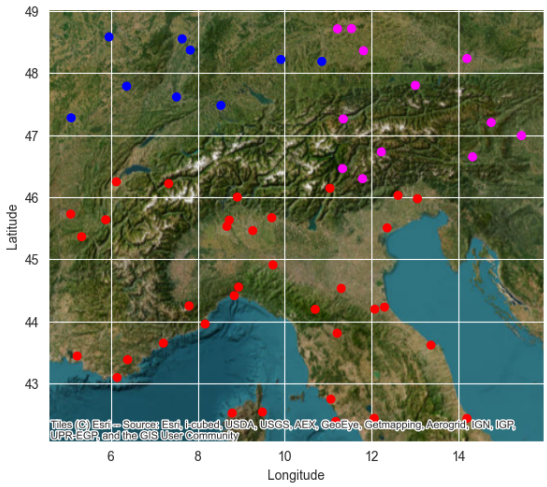

Figure 2 - Stations per region in our geographic area

Elevation Group

Always base on what we read in [3], we discover that regions at different elevations have different behaviours. So we analyze the cumulative distribution function of elevation for the region S, which is the region where our valleys are located. In particular we count how many values in elevation are $\leq$ each unique elevation value.

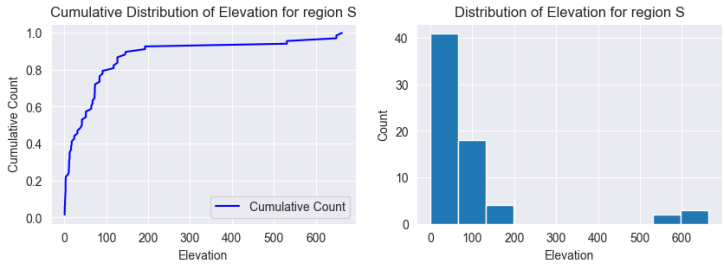

Figure 3 - Cumulative distribution for stations in South region

We decide to create a further division base on elevation. We create three groups, each one has the 33% of the stations.

We do the same for the North-West and North-East regions.

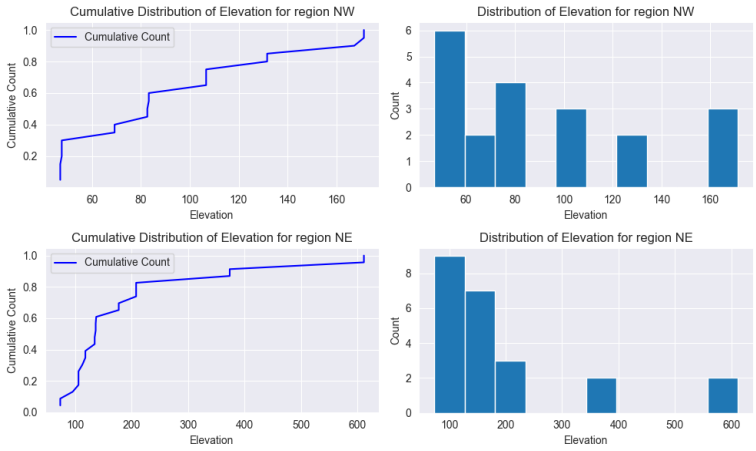

Figure 4 - Cumulative distribution for stations in North regions

Unfortunately here we have not so much data as before for three eventually categories. So, we choose to assign a label to region S for different elevation group. For the other regions we assing "H0" just for not leave it empty.
Furthermore we create another feature for the season. So far we have our clean dataset with our new features: elevation_group, region, season. These new features let us to reduce heterogeneity and discover new correlations.

Outliers Detection

Before going further we are going to considerate only some features, this because in the other features we noticed a massive precence of outliers maybe due to bad measuraments by instruments. We decided to apply winsorization, because is robust to outliers. We use a 90% winsorization where all data below the 5th percentile are set to the 5th percentile, and all data above the 95th percentile are set to the 95th percentile. Here there are the plots before and after the winsorization.

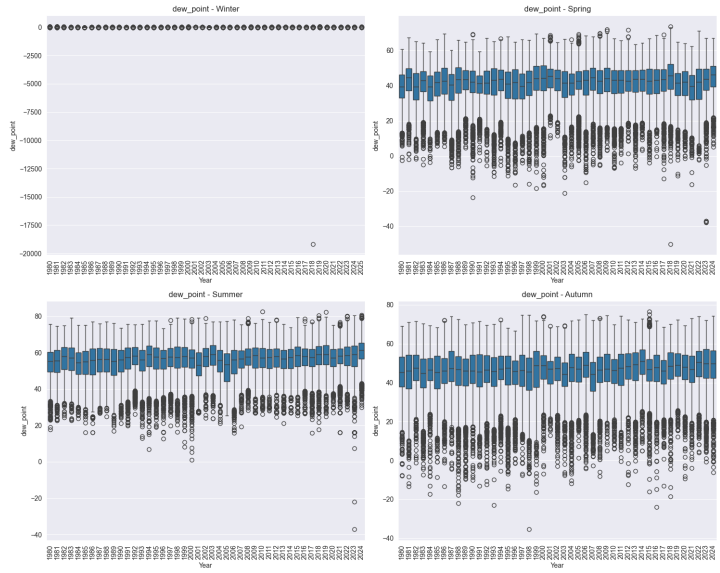

Figure 5 - Boxplots by Year and Season for Dew Point

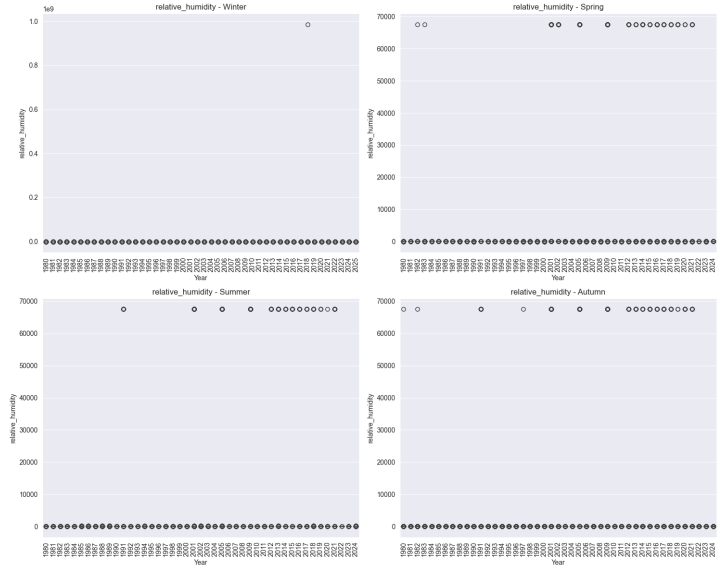

Figure 6 - Boxplots by Year and Season for Relative Humidity

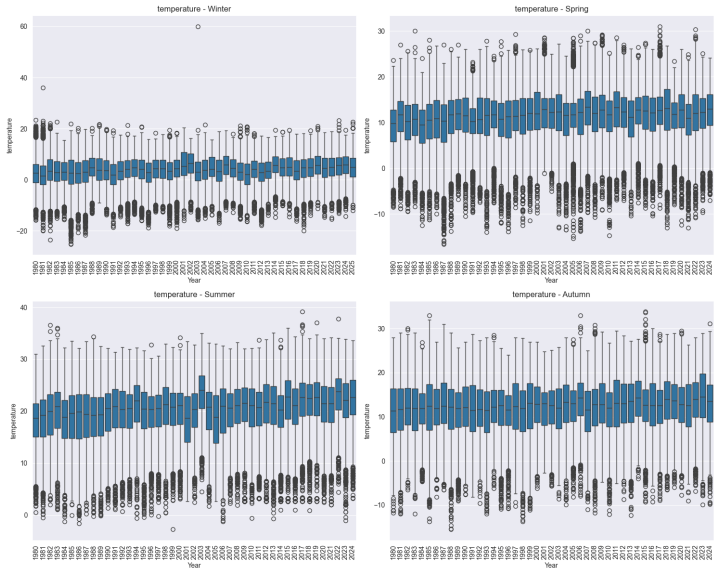

Figure 7 - Boxplots by Year and Season for Temperature

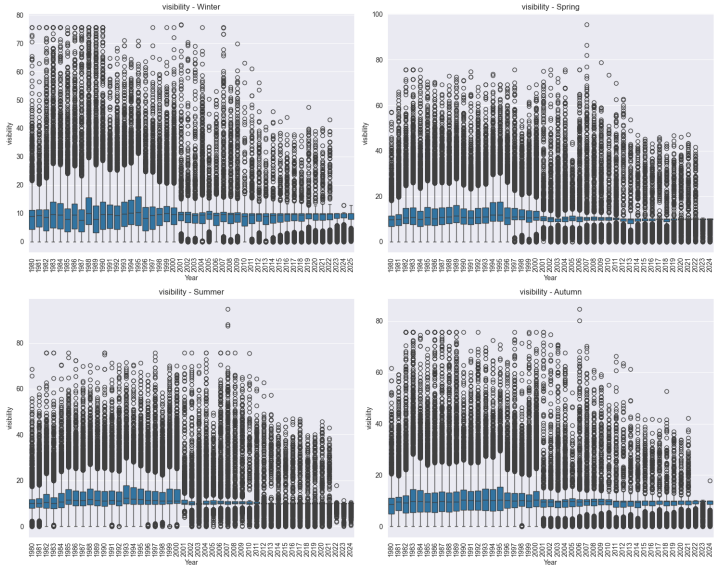

Figure 8 - Boxplots by Year and Season for Visibility

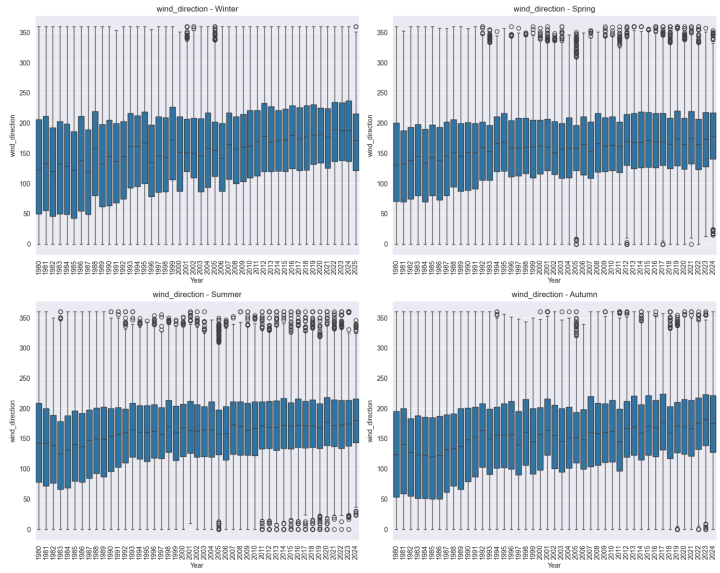

Figure 9 - Boxplots by Year and Season for Wind Direction

Here there are the plots after the winsorization.

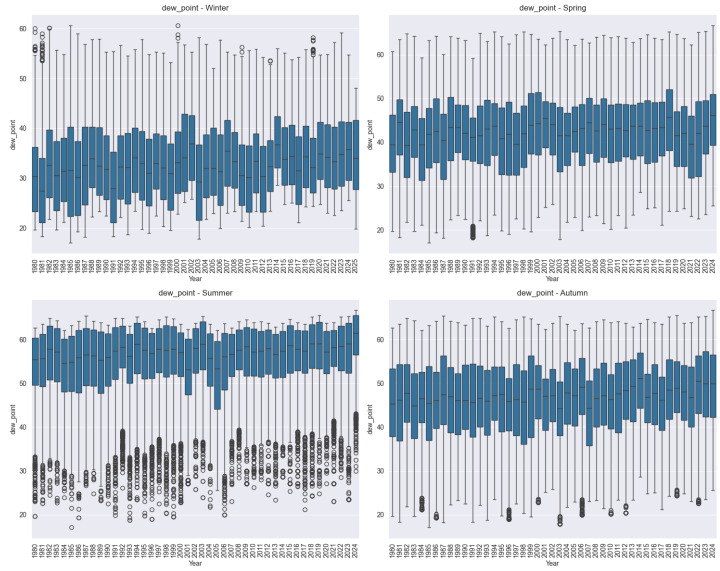

Figure 10 - Boxplots by Year and Season for Dew Point

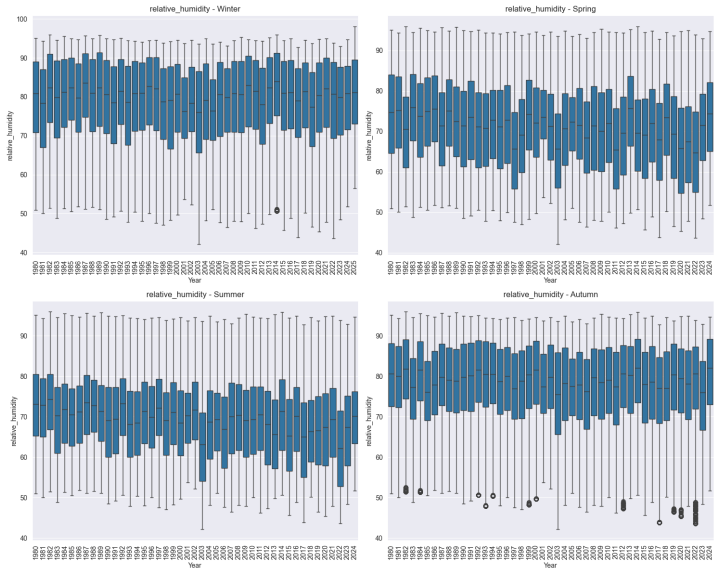

Figure 11 - Boxplots by Year and Season for Relative Humidity

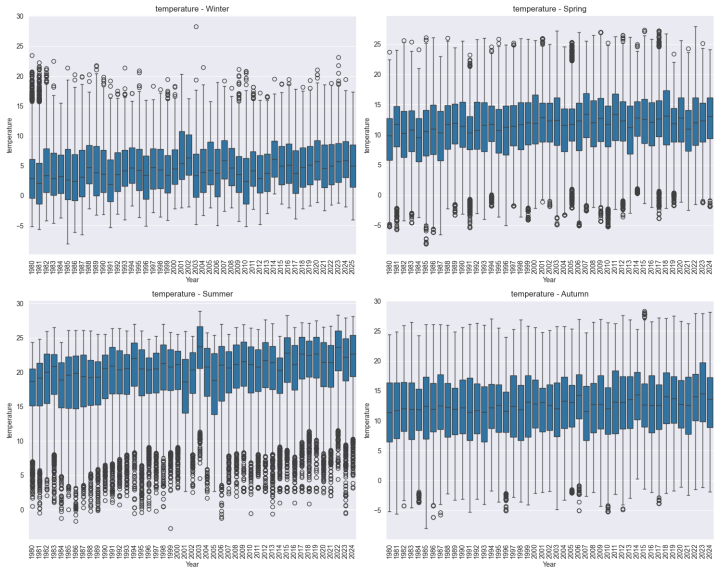

Figure 12 - Boxplots by Year and Season for Temperature

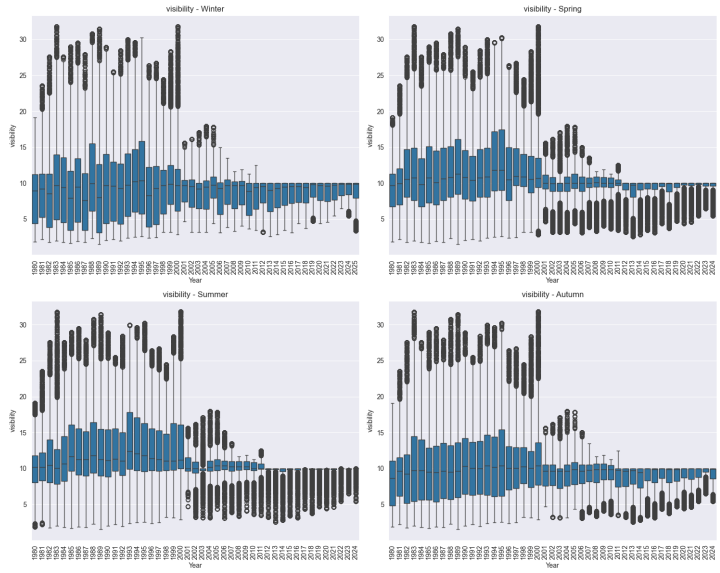

Figure 13 - Boxplots by Year and Season for Visibility

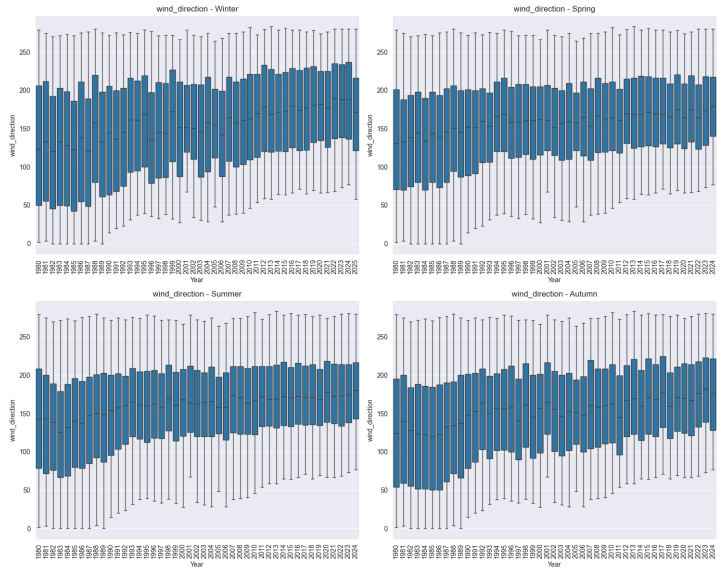

Figure 14 - Boxplots by Year and Season for Wind Direction

To have a better view we plot all these features together.

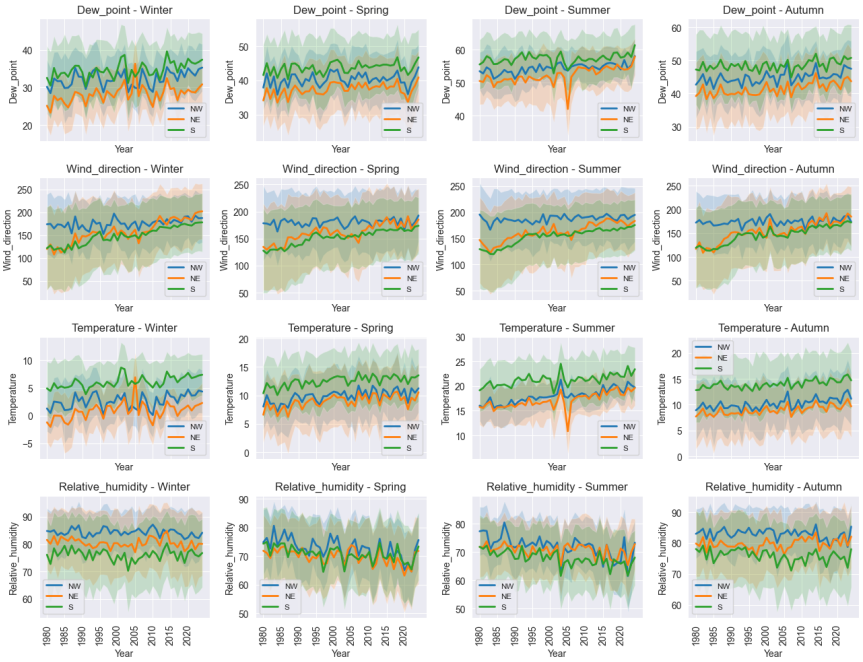

Figure 15 - Time series of all numeric relevant features

Data Analysis

Time Series Analysis

Since our valleys are in the South region, we want to analyze this region for first.

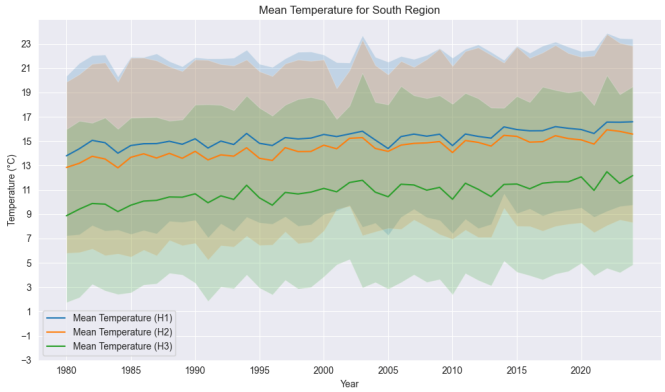

Figure 16 - Mean Temperature for South region by Year

Figure 17 - Mean Temperature for South region by Year and Season

We notice that our elevation groups might be used by a model to better predict the temperature given an elevation value. Furthermore, a little difference in elevation bring anyway a significant differences in the mean temperature, both seasonal and year.
Now we are going to see max and min temperature reached respectively in Summer and Winter.

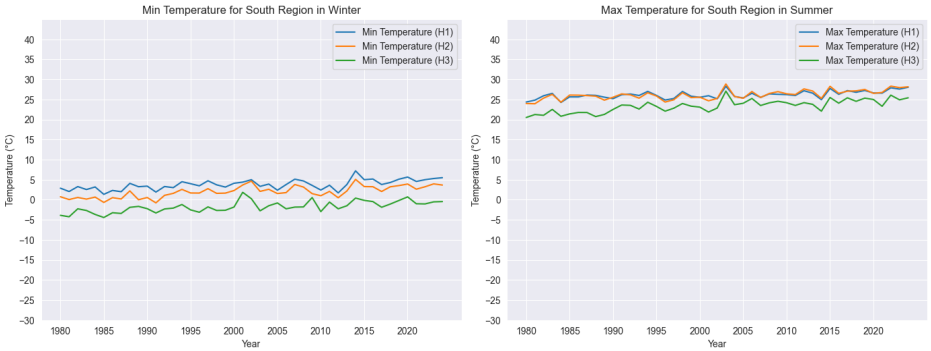

Figure 18 - Min and Max Temperature reached in Winter and Summer for South region by Year

We note not so much difference during Summer, so high temperatures are reached no matter the elevations. Instead in Winter we notice that stations with higher elevation reach lower temperatures. Now we are going to make the same plots also for the other regions.

Figure 19 - Mean Temperature reached for all regions by Year

Here we notice a positive trend in each plot. Moreover there is an evident difference for the three regions: South region reachs higher temperatures than the North regions.

Season and Trend Analysis

Temperature

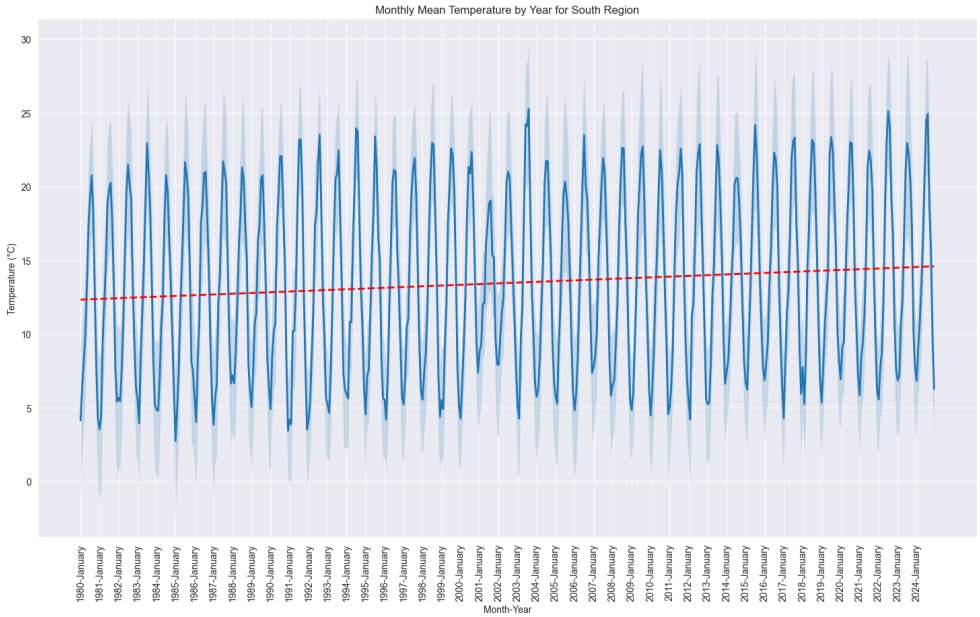

Figure 20 - Temperature trend and a linear fit in red

We see a little positive trend and a clear seasonality: every 6 months.

Dew Point

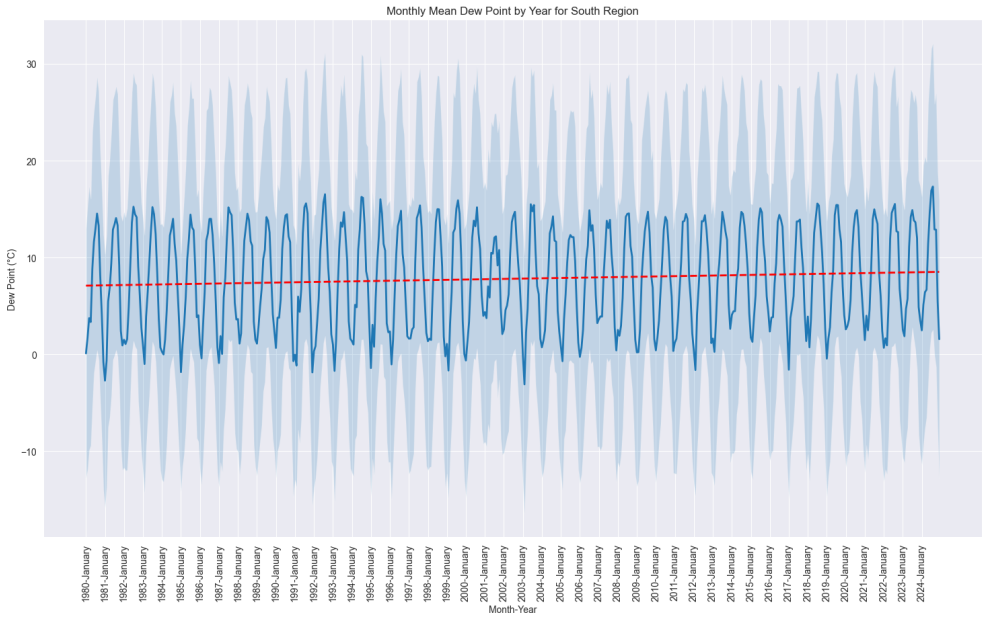

Figure 21 - Dew Point trend and a linear fit in red

We see a very little positive trend and a clear seasonality: every 6 months. Note that the range of values is good and not worrying.

Wind Direction

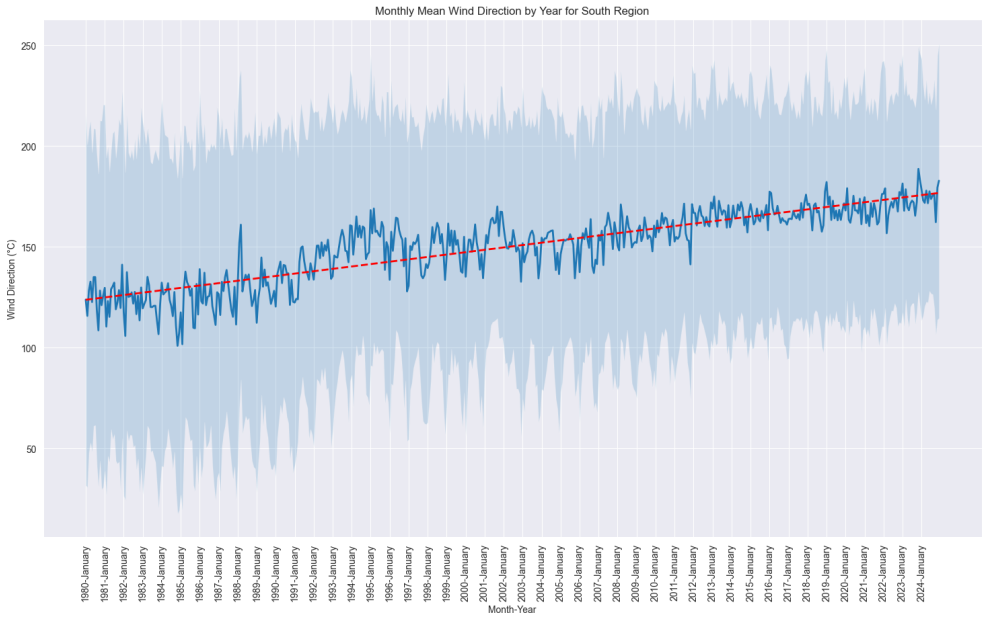

Figure 22 - Wind Direction trend and a linear fit in red

We see a positive trend but no seasonality. This plot clearly indicates the nature of wind: most coming from south so it brings more warm.

Relative Humidity

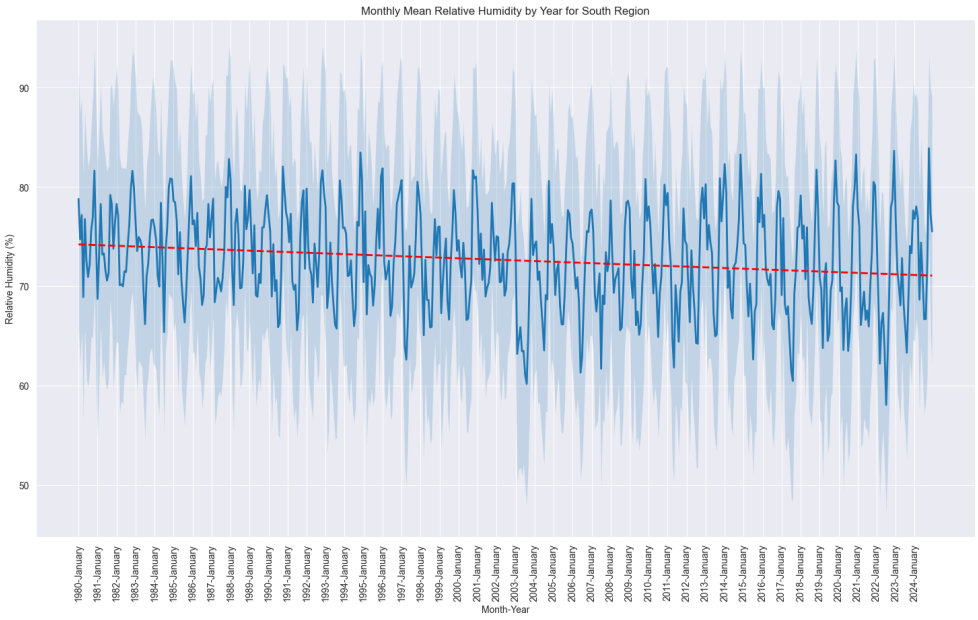

Figure 23 - Relative Humidity trend and a linear fit in red

We note a negative trend and no seasonality.

To make the point:

- Temerature is rising;
- Dew Point remains almost costant;
- Wind Direction is shifting from the south;
- Relative Humidity decrease.

This scenario indicates that nevertheless there is moisture in the air , the increasing temperature will make the air relative dryer. At the same time wind brings warmer and potentially more humid air. However the relative humidity drops indicates that incoming air is not adding enough moisture to compensate for the warming or dry air is mixing into the region. If this trend continues it could lead to warmer and drier conditions.

Model Development

Since we found some seasonality in the data, we decide to train and also predict mean values of temperature, dew point, relative humidity and wind direction for each month. Furthermore we focus on the South region, so our model is more custom for the task.

We focus our attention on these features: station, lon, lat, elevation, region, elevation_group, day, month, year, season. So we train four models with these features as input and a different target for each model: temperature, dew point, relative humidity and wind direction.

Data Pipeline

We follow these steps to preparate our dataframe:

- Select relevant features;
- Reorder the data by date and station, so we have all stations in 1980 then in 1981 and so on, all stacked in a single dataframe;
- Convert into category codes all the categorical features;
- Split train and test (we noticed that random selection for train and test datasets produce better result in our model)

Model Training

We decide to use the XGBRegressor from xgboost library. In particular the four models have these hyperparameters: n_estimators=1000, learning_rate=0.05, random_state=177, max_depth=16.
Since our features depends from stations and we have no stations in the valleys, we are going to make forecast for both using latitude and longitude.
So we have two groups of models:

1. The first four models trained on all the South region dataset and make predictions base on the station ID and coordinates;
2. The second four models trained all the South region dataset filtered with Elevation Group H3 (the highest elevation) and make predictions base on coordinates.

Here the residual plots for the first group.

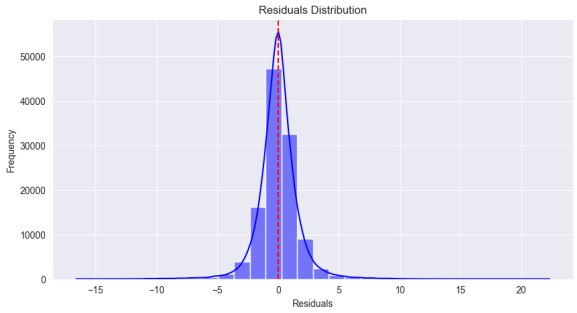

Figure 24 - Distributions of residual for Temperature in South Region

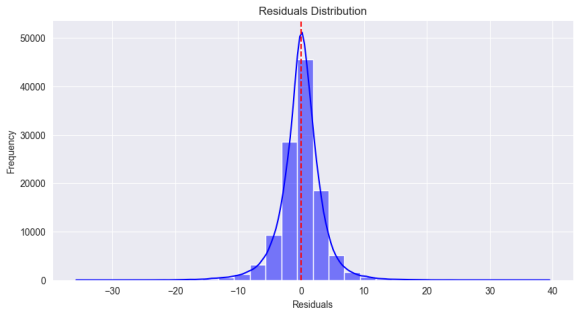

Figure 25 - Distributions of residual for Dew Point in South Region

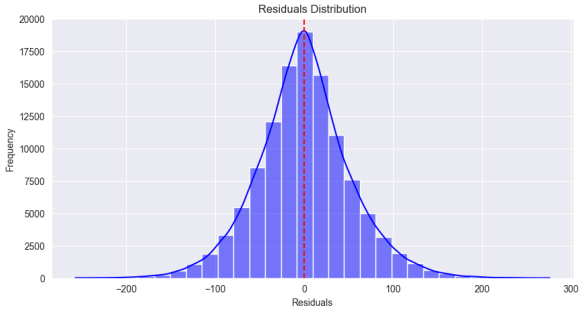

Figure 26 - Distributions of residual for Wind Direction in South Region

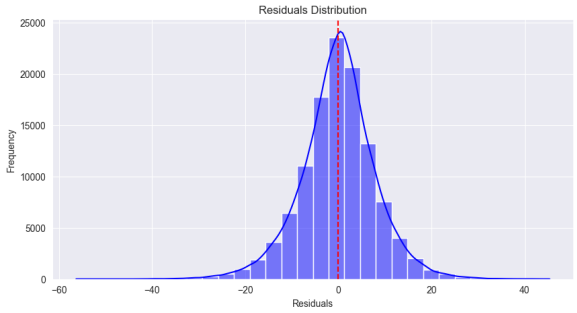

Figure 27 - Distributions of residual for Relative Humidity in South Region

Here the residual plots for the second group.

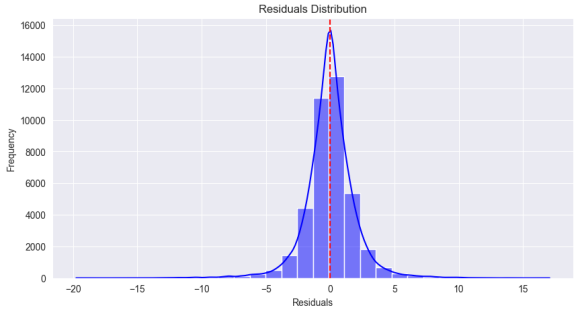

Figure 28 - Distributions of residual for Temperature in Val di Susa

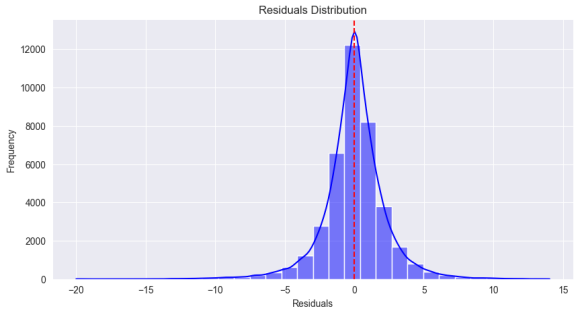

Figure 29 - Distributions of residual for Dew Point in Val di Susa

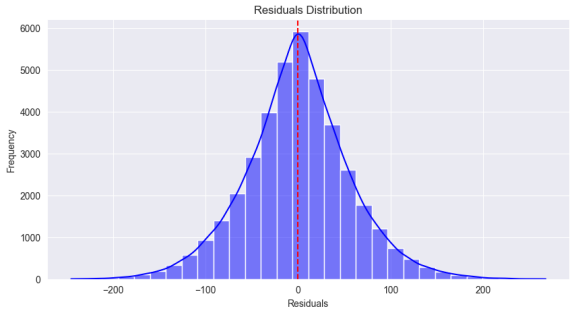

Figure 30 - Distributions of residual for Wind Direction in Val di Susa

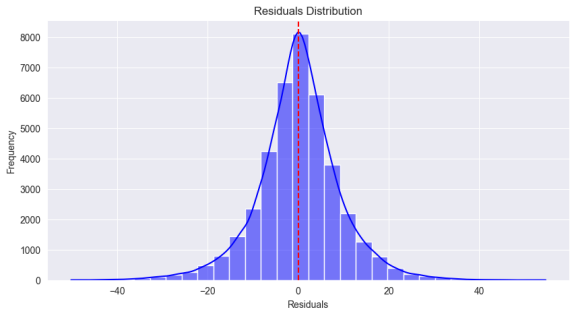

Figure 31 - Distributions of residual for Relative Humidity in Val di Susa

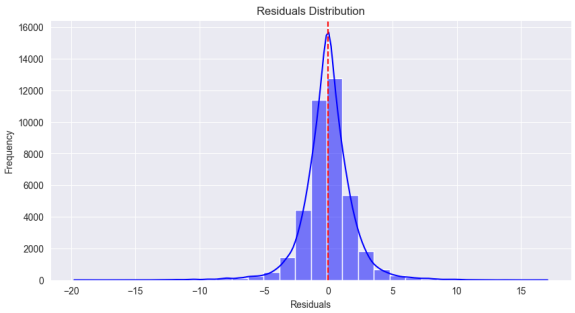

Figure 32 - Distributions of residual for Temperature in Val de Maurienne

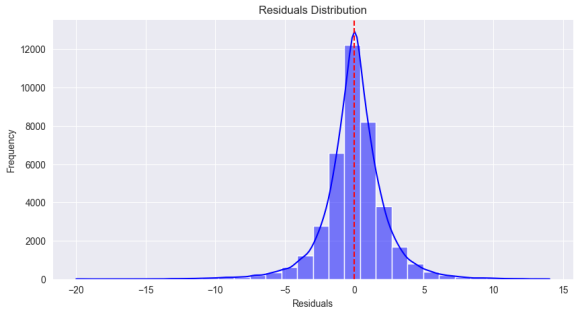

Figure 33 - Distributions of residual for Dew Point in Val de Maurienne

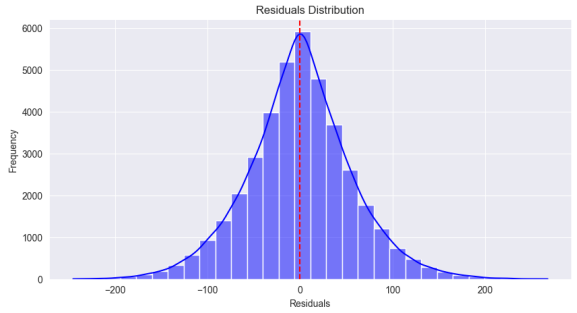

Figure 34 - Distributions of residual for Wind Direction in Val de Maurienne

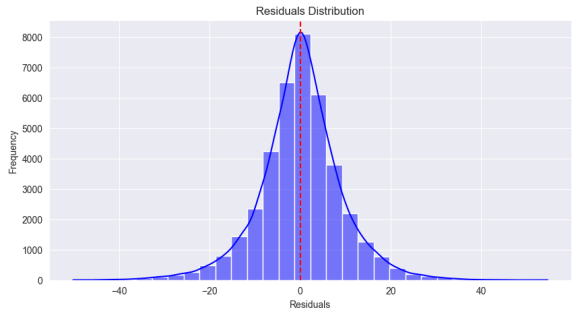

Figure 35 - Distributions of residual for Relative Humidity in Val de Maurienne

Model Forecasting

As discussed before we make month mean predictions, this is done inside a loop: for each month the model calculate the prediction and we store the result. Here the plots about forecasting for the first group.

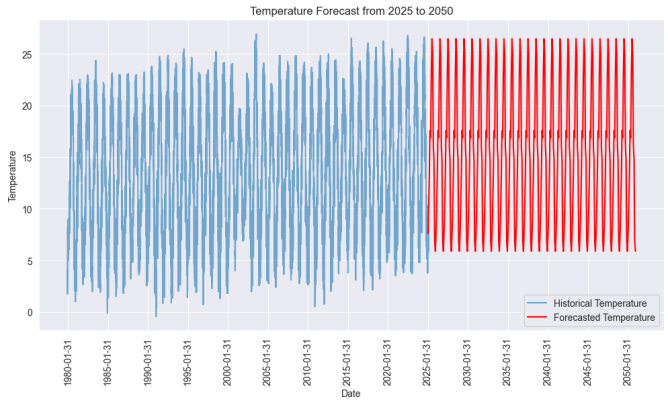

Figure 36 - Temperature forecast in South Region

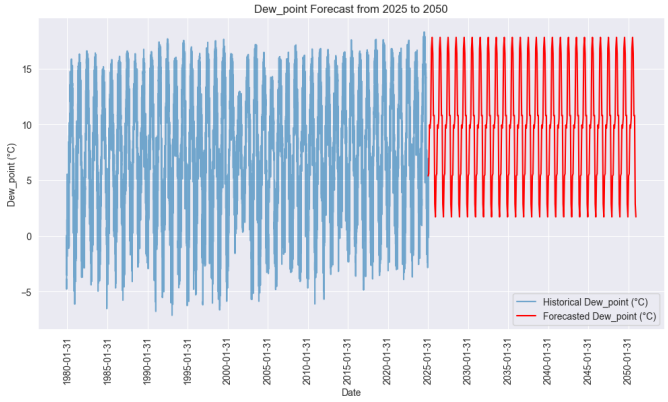

Figure 37 - Dew Point forecast in South Region

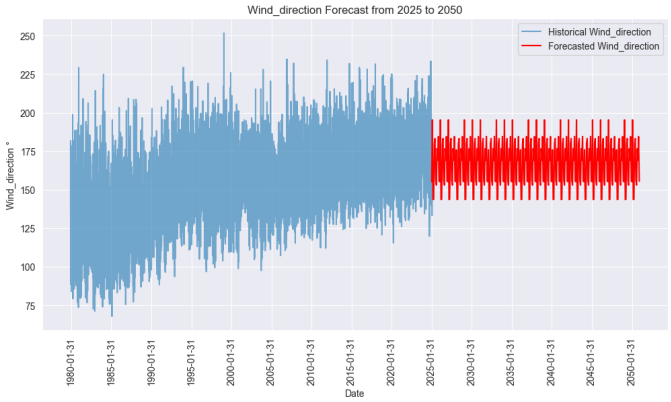

Figure 38 - Wind Direction forecast in South Region

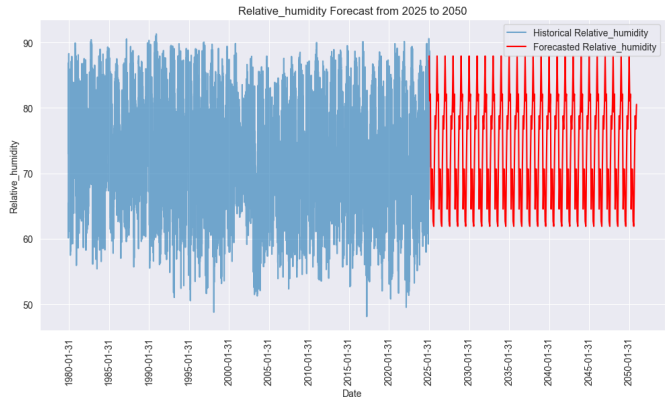

Figure 39 - Relative Humidity forecast in South Region

Here the plots about forecasting for the second group.

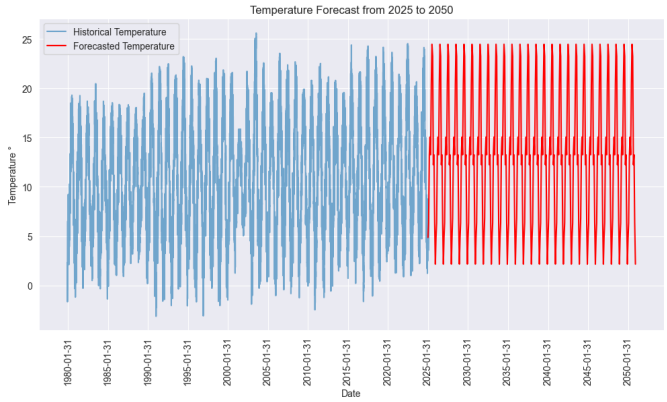

Figure 40 - Temperature forecast in Val di Susa

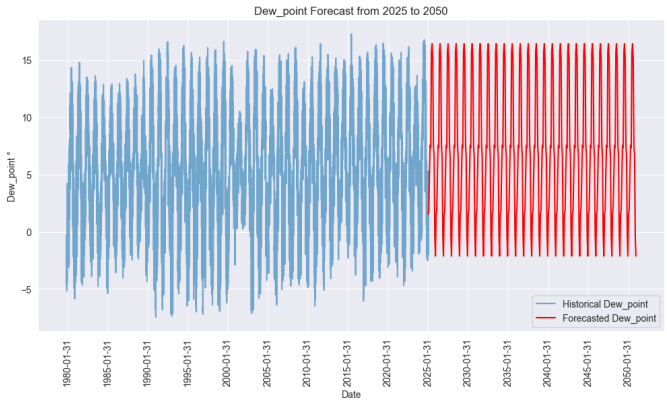

Figure 41 - Dew Point forecast in Val di Susa

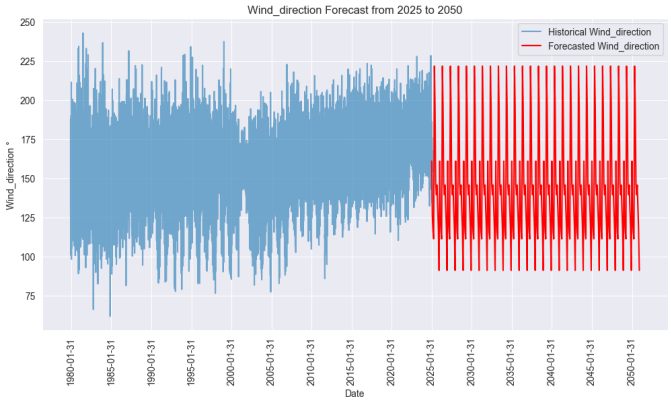

Figure 41 - Wind Direction forecast in Val di Susa

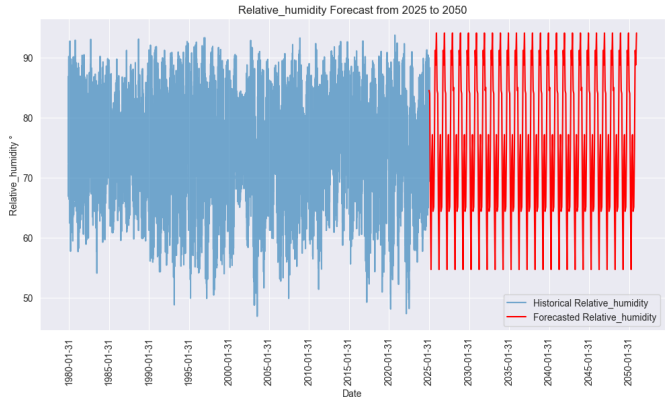

Figure 42 - Relative Humidity forecast in Val di Susa

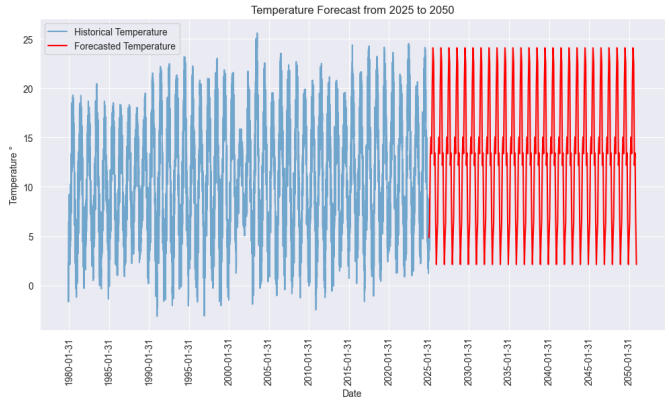

Figure 43 - Temperature forecast in Val de Maurienne

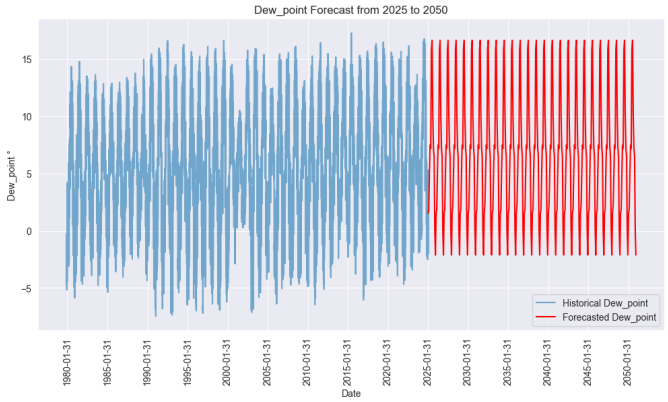

Figure 44 - Dew Point forecast in Val de Maurienne

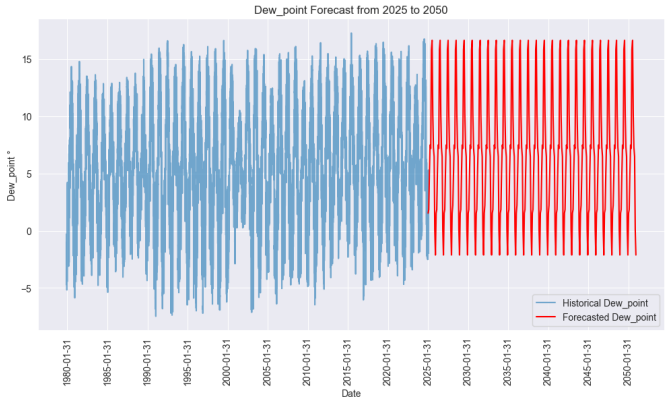

Figure 45 - Wind Direction forecast in Val de Maurienne

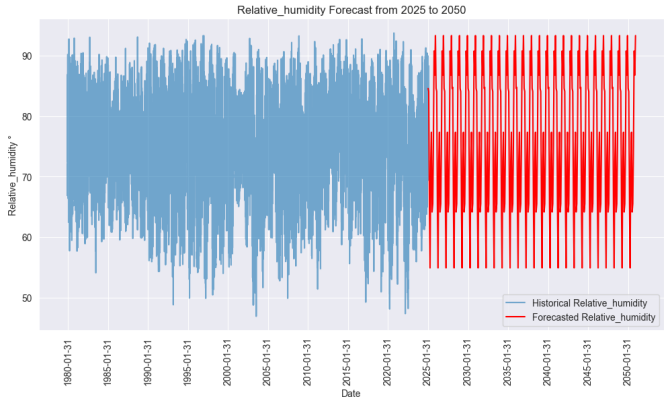

Figure 46 - Relative Humidity forecast in Val de Maurienne

Model Validation

Since we do not have a validation set, we decide to evaluate our predictions with the results obtainde in [3].

From the article we report:

"Here, the summer mean temperature increase by the end of the century (2100) surpasses $5°C$ for RCP 8.5. This stronger warming South of the Alps is part of a well-documented North–South gradient in the projected summer temperature increase across the European continent (Jacob et al. 2014) and probably related to the so-called Mediterranean amplification."
Our projections follow the trend reported in the paper, we are not in the same range of values because we use a little dataset but we are quite close.

Conclusion

Our models fitted the data and in most of the cases followed the seasonality pattern. Also with not so much data respect to the ERA5 we are able to make predictions for future values. Even if we do not have specific dataset for the valleys we are able to understand the trend by looking at the South region and we can train a model that understand the link between spatial features and target values.

One thing is sure: the climate is undergoing significant and concerning changes. It is imperative to seek solutions to prevent these projections from becoming reality.

References

[1] Kochkov, D., Yuval, J., Langmore, I. et al. Neural general circulation models for weather and climate. Nature 632, 1060–1066 (2024). https://doi.org/10.1038/s41586-024-07744-y
[2] https://www.ecmwf.int/en/about/media-centre/focus/2023/fact-sheet-reanalysis
[3] Kotlarski, S., Gobiet, A., Morin, S. et al. 21st Century alpine climate change. Clim Dyn 60, 65–86 (2023). https://doi.org/10.1007/s00382-022-06303-3