# Assignment TAI: Weather Forecasting game

Andrei Covaci

## 1 Introduction

Modern weather forecasting models have come far since the first simulations in the 50s. Numeric weather prediction (NWP) models can now do forecasts at 1km x 1km spatial resolution. Still, these simulations do not take into account urban areas and only deal with standardized "open green field, short grass" conditions. These standardized variables and conditions are optimal for large-scale predictions (forecasting over Europe). When we want to downscale this data further, it becomes crucial to consider the impact of the environment (such as urban heat island) on the meteorological variables. Calculating corrected versions for these variables using traditional methods tends to be rather difficult and computationally heavy. A potential solution is using AI methods to correct weather forecasts for specific environments.

## 2 Data

### 2.1 ECMWF forecast data

The first data set you will work with is the forecasts from ECMWF (European Centre for Medium-Range Weather Forecasts). Specifically, for a handful of meteorological variables, you will get the medium-ranged ensemble forecasts for March-June 2017 and 2018. These forecasts are the outcome of NWP and are made on a grid of 0.25° x 0.25° (latitude x longitude). This assignment will only consider the forecasts at the gird point closest to our target station (Synopic station at the RMI in Ukkel). You will, with this notice that for all the data you will get, the latitude and longitude will be constant. When making these types of forecasts, it is common to repeat the calculations of these forecasts a couple of times with slightly varying initial conditions. When you only use one forecast, you risk it being not representative of the actual weather. As you forecast further into the future, this error will grow. With this having an ensemble of forecasts will help make the predictions more certain. In this data sets, you will have 51 forecasts at any given moment (see variable *number*). The first of these forecasts ($number = 0$ is a control forecast that you can ignore. The other ones you will need for training your models.

There are two files available for you to explore: $ECMWF\_2017\_2018\_precip.csv$ and $ECMWF\_2017\_2018\_surface.csv$. The first contains data on variables related to precipitation. The primary variable from this file that will interest you is $tp$ (total cumulative precipitation since the last forecast lead time). The second file contains the forecasts for the meteorological variables close to the earth's surface. The main variables you will have to consider from this file are $t2m$ (the temperature at 2 meters height above the ground) $u10$ and $v10$ (wind components along the latitude and longitude directions). Think about how you can convert $u10$ and $v10$ to the total wind speed.

To further clarify the data set, we have printed here the head of the ECMWF_2017_2018_surface.csv with some of the more important columns.

- *time*

This variable represents the moment at which the forecast is started. All the forecasts are always initiated at the start of the day (midnight). For this data set, a forecast is initiated for every day between March and June. These forecasts predict the meteorological variables up to 5 days into the future.

- *step*

This variable tells us how far into the future the current forecast is predicting. Example: *step* = 1 day 6 hours and *time* = 3/3/2017 means that the meteorological variables in that row are forecasted values for 4/3/2017 at 6 AM (valid_time column) for a forecast that was initiated at 3/3/2017 0:00 AM.

- *number*

As mentioned, when making these forecasts, we repeat these several times with slightly different initial conditions. In particular, for this data set, every forecast is initiated 50 times for every day. This means that for each value in the column *valid_time*, there will be 50 rows in the data set containing forecast for the meteorological variables for that moment in time. The variable *number* denotes which of these forecasts you are looking at from the ensemble.

| | number | time | step | t2m | valid_time |
|---|---|---|---|---|---|
| 0 | 0 | 2017-03-01 | 0 days 00:00:00 | 277.47363 | 2017-03-01 00:00:00 |
| 1 | 0 | 2017-03-01 | 0 days 06:00:00 | 278.44824 | 2017-03-01 06:00:00 |
| 2 | 0 | 2017-03-01 | 0 days 12:00:00 | 281.63452 | 2017-03-01 12:00:00 |
| 3 | 0 | 2017-03-01 | 0 days 18:00:00 | 279.18726 | 2017-03-01 18:00:00 |
| 4 | 0 | 2017-03-01 | 1 days 00:00:00 | 280.44238 | 2017-03-02 00:00:00 |

## 2.2 Synoptic data set

The next data set that you will consider is the synoptic data set. *synop*_2017_*March*_*June.csv* and *synop*_2018_*March*_*June.csv*

This is the data measured by the two automatic weather stations at Ukkel. These two stations are positioned next to each other. In the data set, you will find two values (one for each station) at every timestamp. You only need the data from one of these stations, therefore you can throw away the data from the second station. You will use the data from this file as the target values for your machine-learning models.

## 2.3 Weekly forecast data

On a weekly basis, you will get an ensemble forecast (50 members). You will have a test file *test*_*wheather*_*forecasting*_*game.csv*. Every forecast you will receive on a weekly basis will look like this. You will receive a weekly forecast on Monday mornings and you will have to submit your calculated quantile values by Tuesday midnight.

# 3 Your tasks

Your main task is to train a machine learning model that takes the weather forecast as input and outputs values of the meteorological variables corrected for the station at Ukkel. With this when training your model, you should use the weather forecasts as input and the data from the synoptic station at Ukkel as your target. However, the output of your method should still be an ensemble of values (50 values) such that you can calculate quantiles from it. The way you do pre-processing and the design of the models is entirely up to you. We expect that you will implement the regression models that were brought up during the lectures of TAI. At the end of the semester, you will have to write a scientific report on your findings of the models. Here you are expected to explain and justify the choices you made for design, models, and preprocessing ... You will also discuss and interpret the results of your models.

Every week you will receive on Monday the forecast for that week, and you will have to submit for every timestep of the forecast starting at 6 hours to 120 hours (5 days) into the future the following five quantiles calculated from the output of your ML model: q0.025, q0.25, q0.5, q0.75 and q0.975. You will have to do this for temperature (t2m), windspeed (wind), and precipitation (precip) and submit the output in a .csv file by Tuesday midnight on canvas. (There will be an assignment opened where you can submit this file). The format of your file needs to be the following:

- name: the file's name needs to be the date when the forecast was initiated (no spaces in between) YearMonthDay+_ The name of your group with no spaces and every word should be capitalized + .csv. e.g. 20210801_MemphisMayFire.csv
- a column called forecast_date contains the day the forecast was initiated.
- a column called target which tells you for which meteorological variable you have calculated the quantiles in that row. The names of these variables need to be *wind* for wind speed, $t2m$ for temperature, *precip* for precipitation, and $rh$ for relative humidity.
- a column called horizon with the information on how far into the future your forecasted values (quantiles) are compared to the date of initiation. (similar to the column 'steps' in the 2017-2018 ECMWF files). The values should go from '6 hour' to '120 hour' with step of 6.
- 5 columns called q0.025, q0.25, q0.5, q0.75 and q0.975 respectively with these values for these quantiles

the file 20210801_MemphisMayFire.csv on Canvas serves as the template for what your submission needs to look like.

Optionally, you can also calculate the values and quantiles for the relative humidity (rh) net to the three obligatory target variables. However, for this variable, there are no input values in the 2017-2018 forecast data that you can use as input to train your model. In the weekly forecast data, this variable will also not be available, but it can be calculated from the dewpoint temperature (d2m) and the normal temperature (t2m) by the following formula:

$$RH = 100 * \frac{EXP((17.625 * d2m)/(243.04 + d2m))}{EXP((17.625 * t2m)/(243.04 + t2m))} \tag{1}$$

In the first week (Monday, 3/4/2023) of making the forecast, you will be asked to NOT use any ML methods yet and calculate the quantiles directly from the (raw) forecast data you receive that week. For this week, the calculation of quantiles for relative humidity is mandatory, but this will become optional for the upcoming weeks. For the rest of the semester, you will use your ML methods to correct the forecast and calculate the quantiles from the output. On the submission location, you will also be asked to send in a .txt file where you mention which ML model was used to obtain the results you submitted that week. If you attempted multiple models, mention this in the .txt file. However, for your quantile submission, you only need to send in the outcome of 1 of these models (mention which one was chosen for this in the .txt file).