Project Techniques of AI

2022-2023

Project - Techniques of AI

- Option 1: Breast cancer data
- Option 2: Weather forecast game
- Practical details

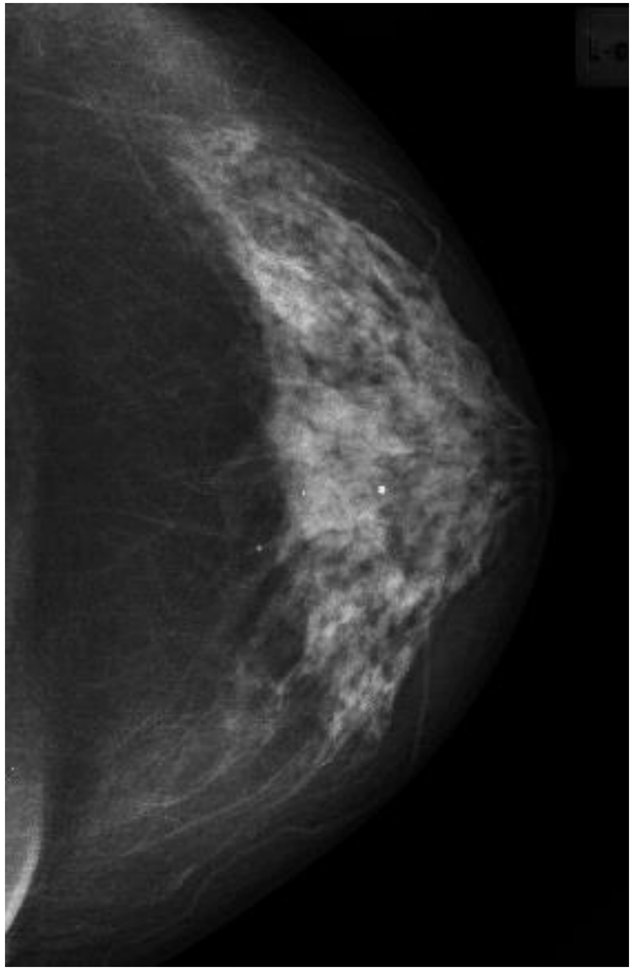
Option 1: breast cancer data

- Detecting breast cancer from images.
- Binary classification problem on 3500 instances, 150 attributes.
- But there are some clear challenges …

Some background on breast cancer CAD

1mm

- State of the art is mammography
  - 2 X-RAY images
  - Due to anatomy of the breast, tumours are hard to see
  - Unless you are lucky or they are quite big
- Micro-calcifications:
  - Are tiny white calcium deposits in the breast
  - Are very easy to see on mammography
  - Micro-calcifications appear as a natural process of ageing

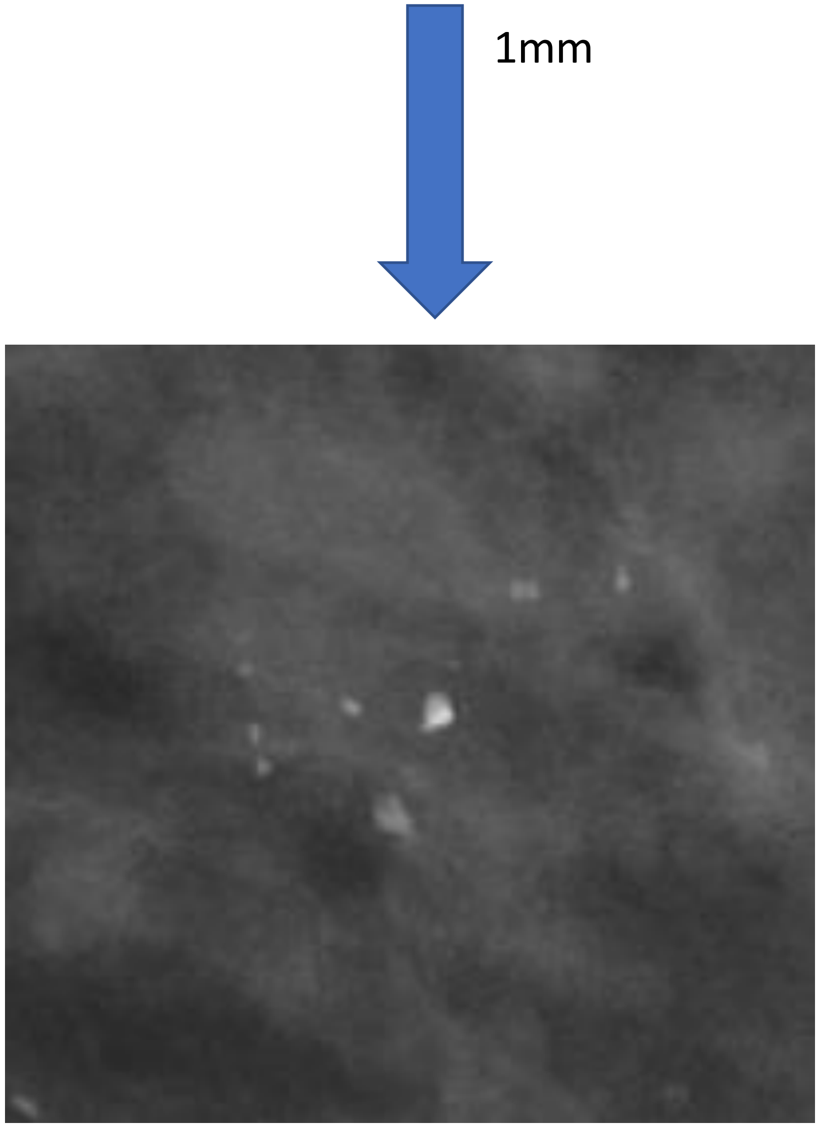


Some more background

- The presence of certain groups of micros is indicative for breast cancer.

- So, there can be a tumour without (visible) micros and the other way around,
  but there is a correlation.

- Radiologists look for clusters of micros on the mammography

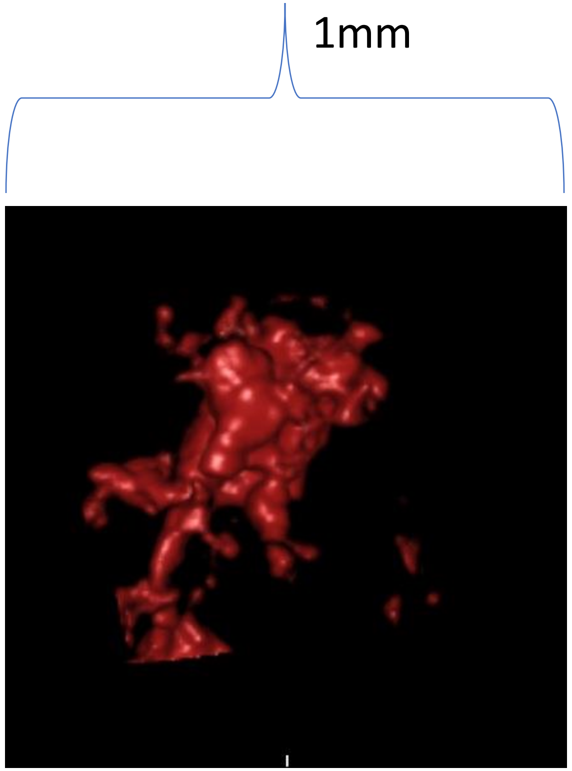


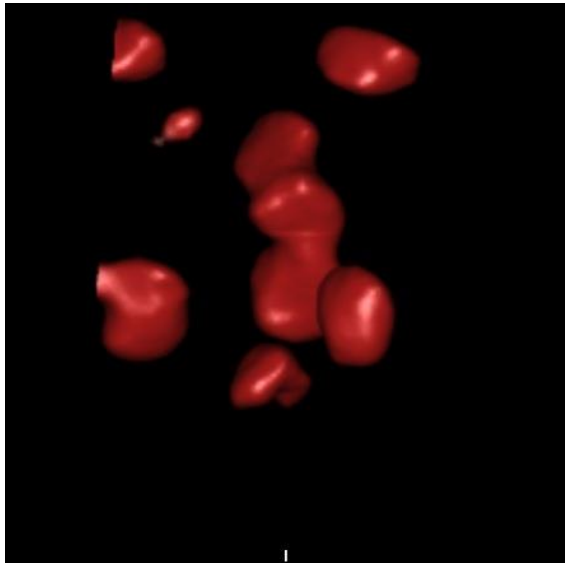
Research hypothesis

- There is even a link between individual micros and cancer: “Shape and texture properties of individual micros allow to predict malignancy”.

- “malignant micro”: micro-calcification present in the neighbourhood of a tumour.
- “benign micro”: micro-calcification not in the neighbourhood of a tumour.

- Problem is having 3D high resolution images of micros





Your challenge

- Based on ~50 properties computed on the ~3500 micros, classify them as benign or malignant.
- **Problem: Every subject in the dataset presents multiple micros and we only know for sure whether the subject has breast cancer or not.**
- **So, 50 benign cases and 50 malignant cases result in 3700 micros in total.**

- Task 1: how well can you classify individual micros assuming all micros per subject have the same label?
- Task 2: how well can you classify whether a subject has cancer based on your classification of the multiple micros per subject?
- We want you to work on both task 1 AND task 2

- You should inform us on the approach you plan to take for both tasks before you start.

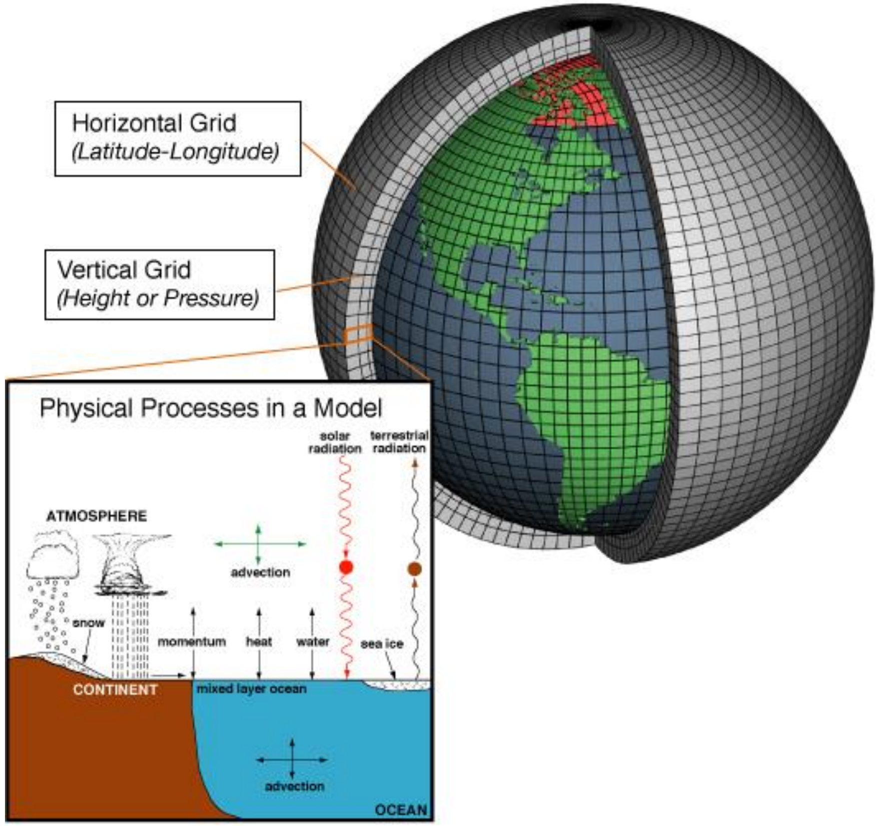
Option 2: weather forecast game

- Goal: **improve local weather forecasts**
- Training data:
  - Past ensemble predictions for 4 weather variables (temperature, relative humidity, precipitation & wind speed) ECMWF (European Centre for Medium-range Weather Forecasting)
  - Observations from a weather station
- Target the 4 variables at a station in the city (VLINDER station in Brussels)
- Instead of exact values, we want quantiles

  => multiple options to get these.

Background: what is NWP?

- Discretize the atmospheric variables (temperature, humidity, wind, pressure) into a grid
- Numerically integrate forward the physical laws that govern the atmospheric evolution
- The result is a weather forecast, either global or for a specific region.



NWP models have errors

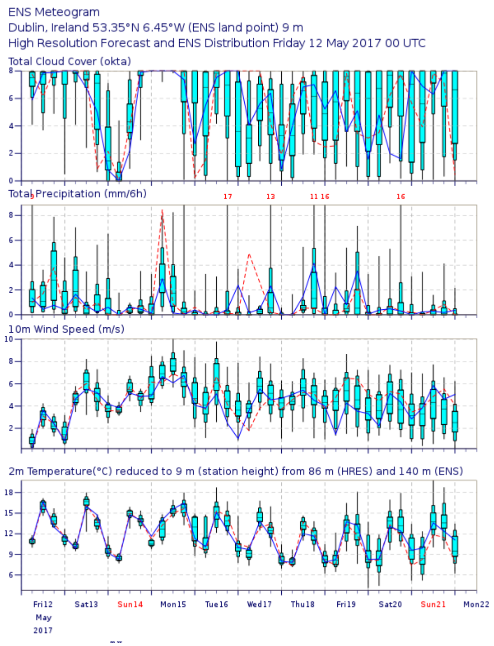
Different types of errors within the NWP models:

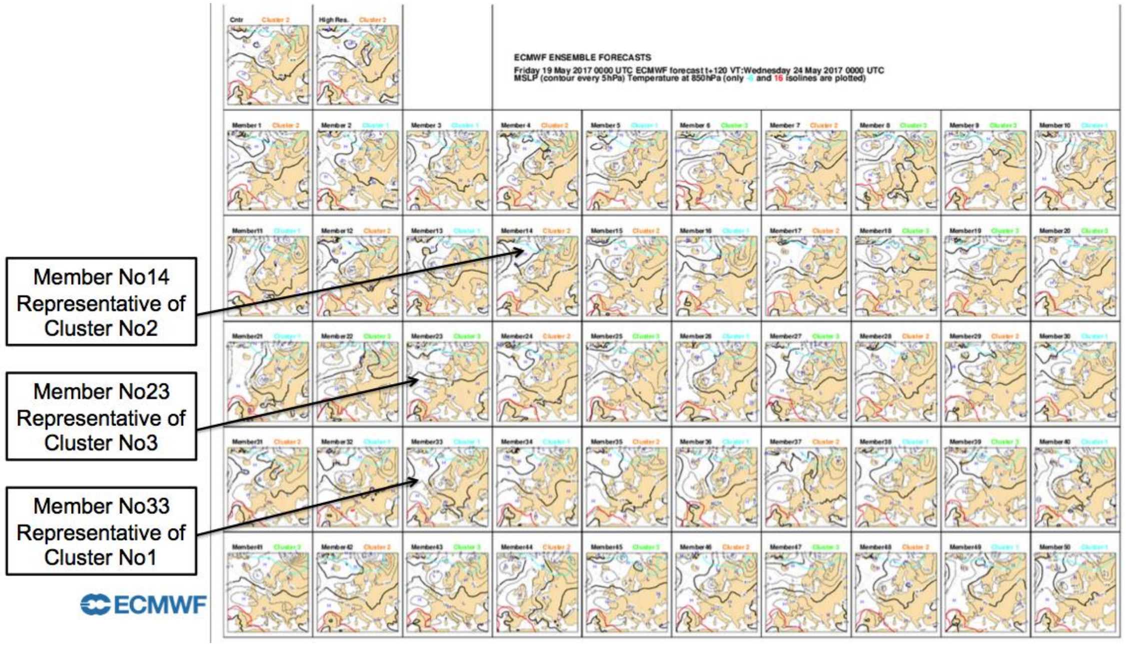
- Initial condition errors
  - sparse measurement: no 100% correct actual initial state of the atmosphere
  - chaotic system means these errors will grow, predictability is limited to 1-2 weeks ("butterfly effect")
- Model errors, related to
  - Numerical errors, discretization of the equations
  - Complex processes that are represented in a simplified way (turbulence, radiation...)
  - Coarse resolution: terrain is not represented accurately

Post-processing: statistical methods to (partially) correct for these (systematic) errors: ML is interesting here!

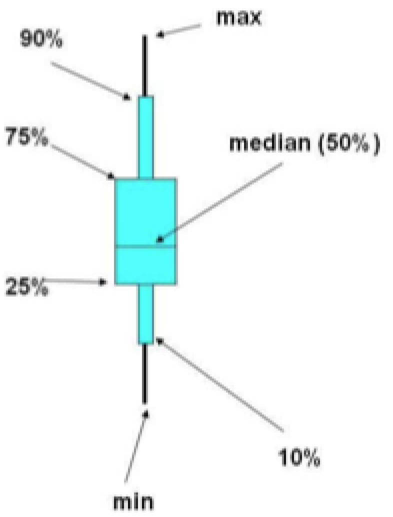
Ensemble forecasts to estimate uncertainty

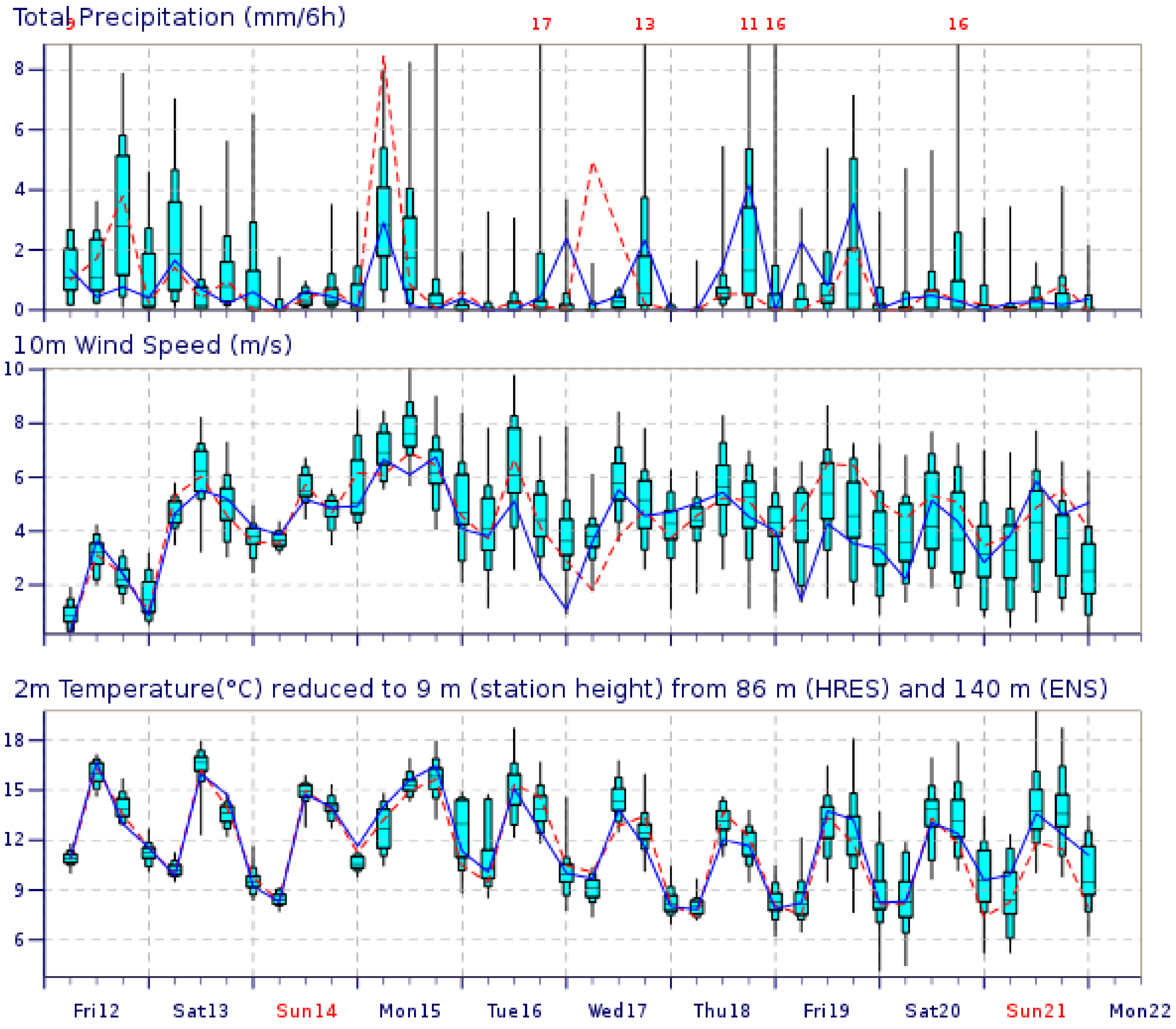
- Because of chaos, forecasts are always uncertain.
- By calculating multiple future scenarios, we generate an "ensemble forecast" that gives us an idea of the uncertainty of the forecast.





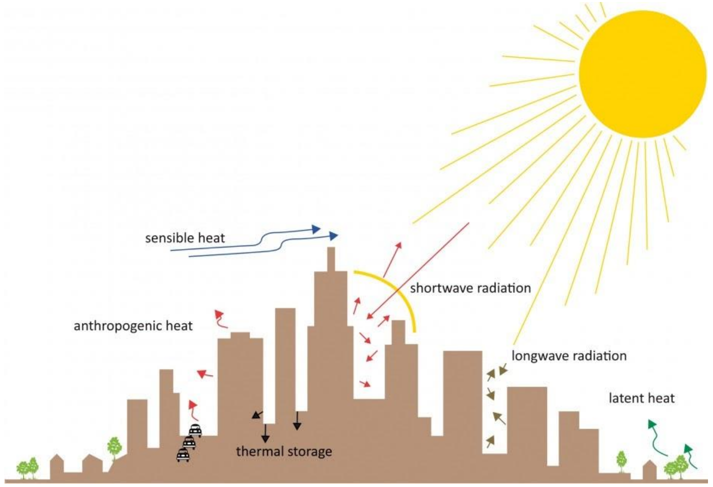
- Forecasts can be visualized as maps or as **meteograms** (for one location)
- Uncertainty info shown as **quantiles** (box-and-whisker plots)





Local influence of the urban environment

- Weather forecasts target 'open green field with short grass' conditions
- An urban environment impacts the local weather (e.g. urban heat island, buildings block the wind, …)
- Physical simulation of the hyperlocal influence of the urban environment=> tends to be computationally expensive
- Research @ ETRO: Use crowdsourced data for ML emulation of the equations (see assistant Andrei Covaci's research)
    - Learning the impact of the environment on the meteorological features without having to solve equations
    - Less computationally costly (?)



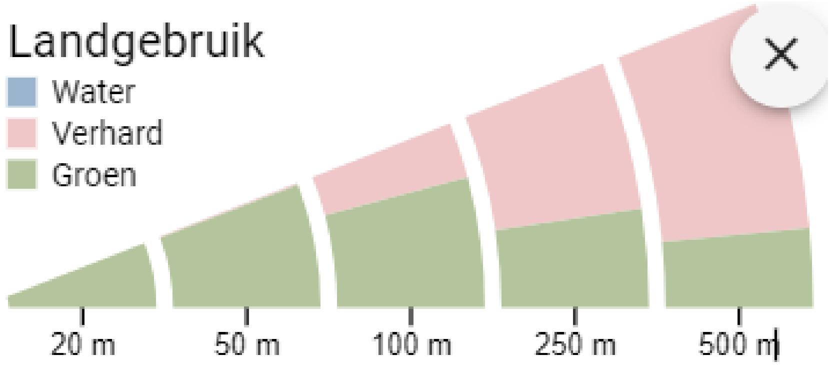
The VLINDER network

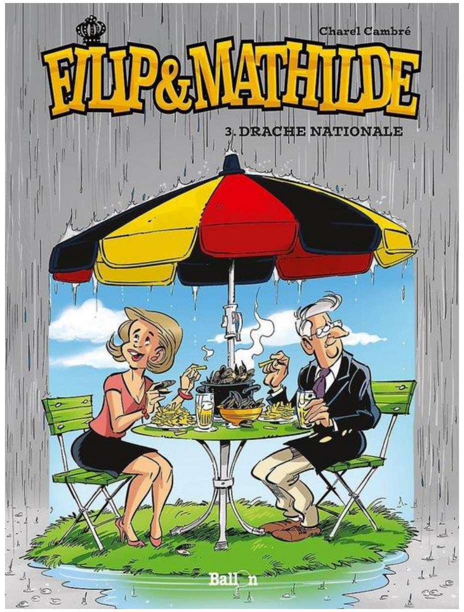
- Weather stations network in nonstandard environments, run by UGent

Observations: VLINDER station

- Target: VLINDER station 19
- Location: Royal Palace, Brussels
- Particular influence of the environment on the variables (raised temperature during nights, less wind)

Vlinder 19





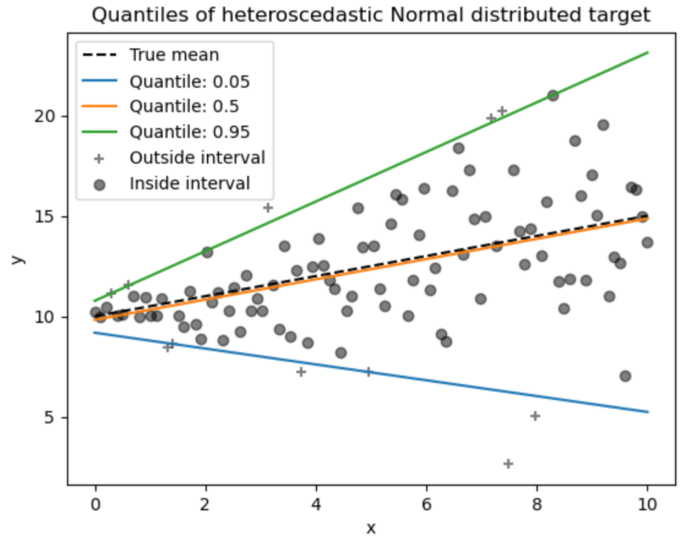
Your challenge

- Train a machine learning model to make corrections on the numerical weather prediction ensemble for the "royal" weather station, Vlinder 19.
- Every week, you upload your corrected forecasts for the upcoming week in advance at 6-hour intervals
- Your predictions will be posted on http://ancovaci.shinyapps.io/ptfsc_viz (anonymized by choosing a musical artist/band as your group name)
- **Challenge: we don't want just a regression model, but quantiles. Multiple ways to do the setup for this kind of experiment.**
- We propose three ways to get you started, but feel free to propose and try as many things as you want => explain why and how!!!

Method 1: quantile regression models

- Use models that have as output quantiles

- E.g. QuantileRegressor instead of LinearRegression from scikit-learn

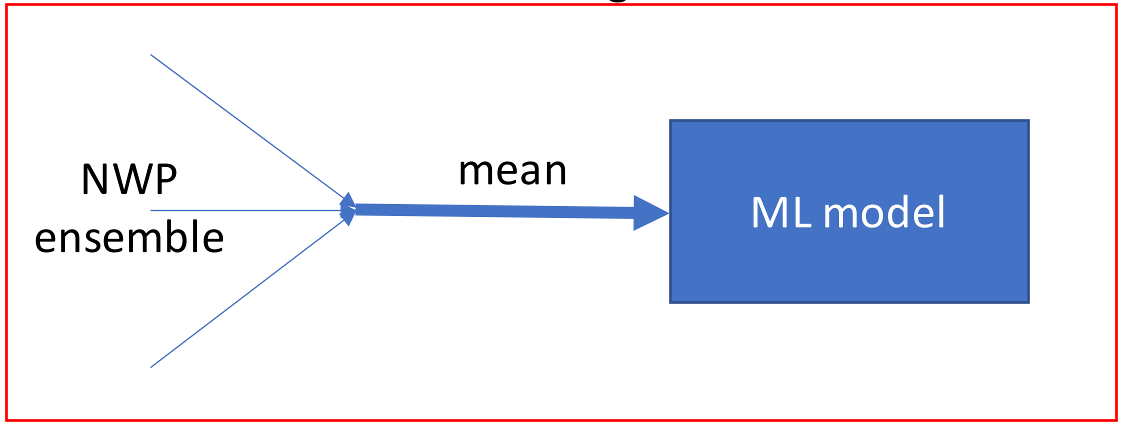


Method 2: mean as input

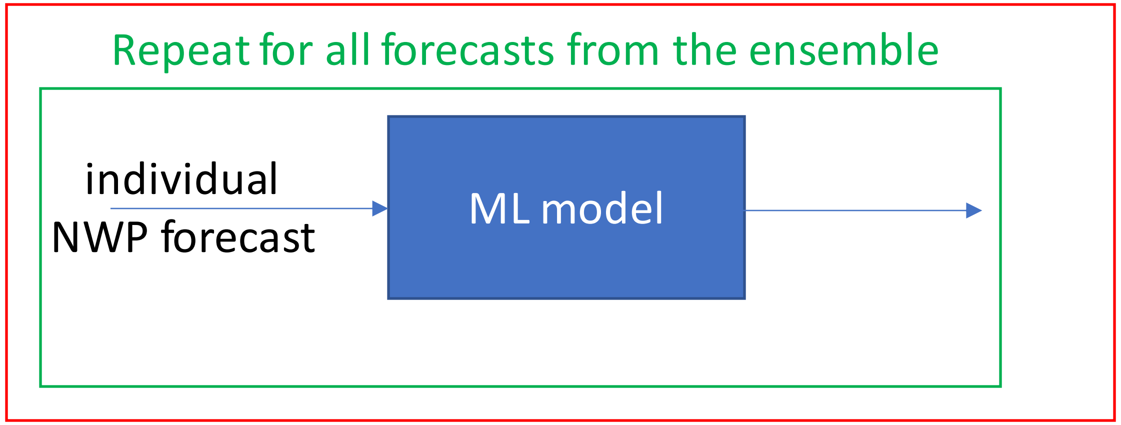
- Train one regression model on the mean of the NWP data

- For evaluation, use each individual prediction instead of the mean from the NWP, multiple output values at every time step => spread => calculate quantiles

Training



evaluation

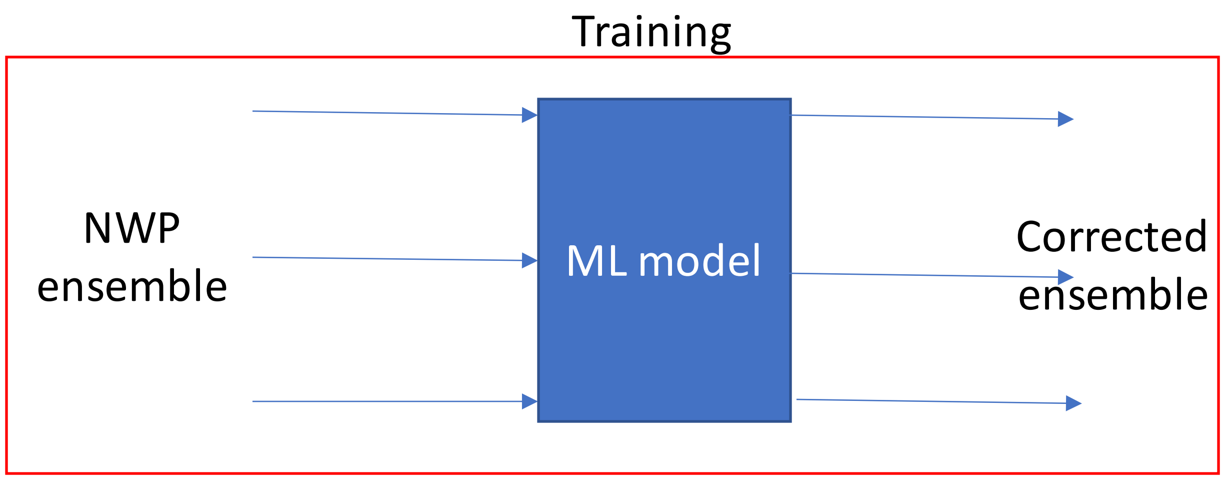


Method 3: full ensemble post-processing

- Train a model that produces an ensemble

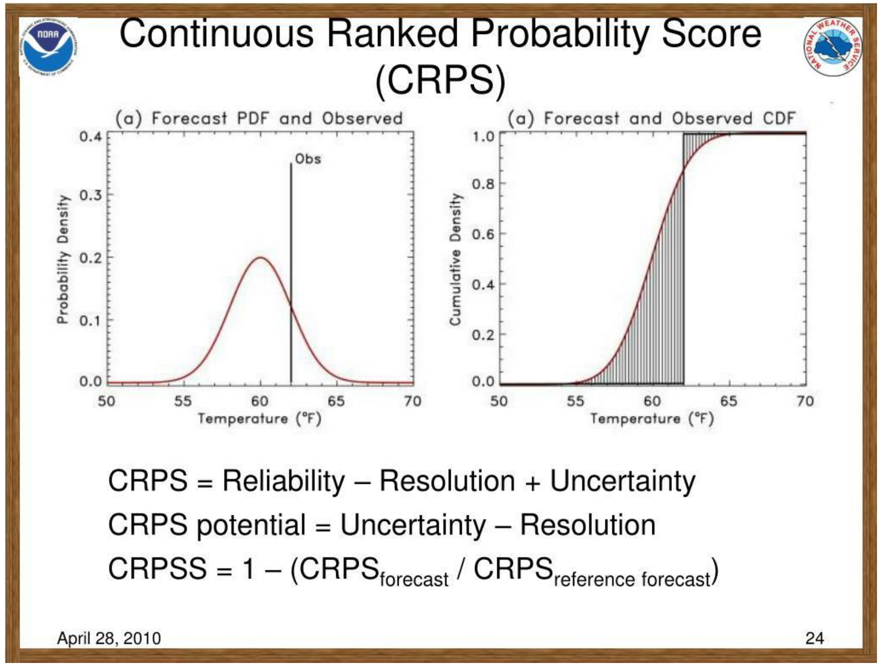
- Multiple ensemble members => multiple output values

=> quantiles can be calculated



Scores

- In addition to traditional scores such as root mean square error, you can evaluate probabilistic scores
- For example: CRPS (continuous ranked probability score)



Remarks

- Think carefully about the option to select.
- Each option requires critical thinking, analysis, understanding and insight.
- It is OK if the results are not that good, the implementation of the models and ‘experiment design’ are the most important parts
- Alone or in teams of two. Think about how to make 1+1=3.

Important dates and information

- Project is **50% of total score** !!! (other 50%: closed book exam)

- Selection of topic (through a Canvas quiz): **March 13**
- Midterm evaluation for breast cancer project: **April 17**
- For the weather forecasting game: weekly submission starting **March 27** (always on Monday)
- Submission of project report on Canvas: **May 29**
- Note on project report: follow the template that will be provided on Canvas.
- **Oral defense**: in June (one of 2 days, you will be able to book your timeslot)

- Contact us in case of doubts through dedicated Discussion session in Canvas.