Atmospheric Aerosol Modeling

Daniel Holmberg
Ella Rauth
Mikko Markkinen
Aaro Suominen

In collaboration with

INAR
Institute for Atmospheric and Earth System Research
UNIVERSITY OF HELSINKI

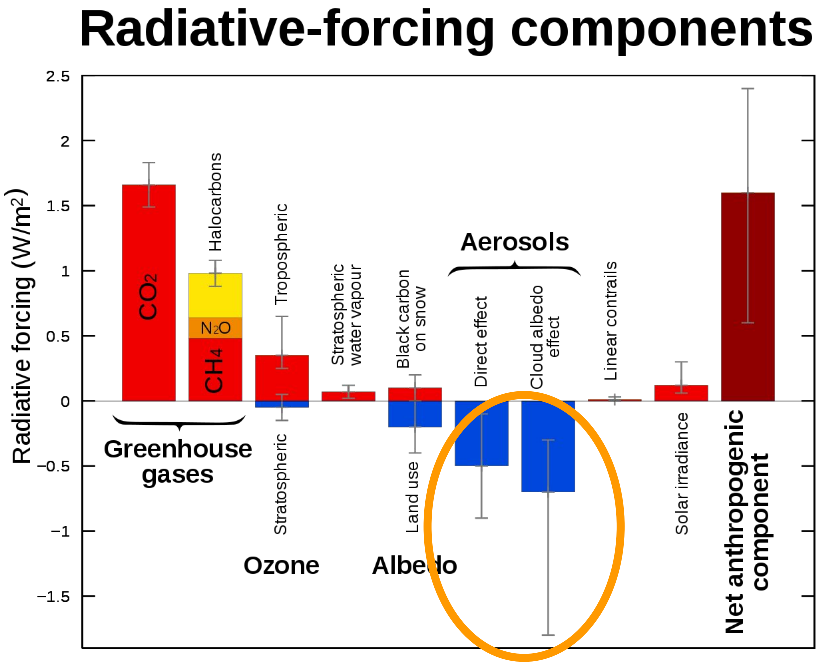
Instructions

The oral presentation consists only of the pitch that motivates the problem and the tool, and demonstration of the tool (in a manner suitable for the particular tool). This would be roughly 10 minutes in duration

Pitch

Need

- Significant uncertainty regarding the effect of aerosols on global climate
  ⇨ Strength of cooling effect unclear
- Unable to measure aerosol concentrations from satellites
  ⇨ Rely on scarce availability of field data

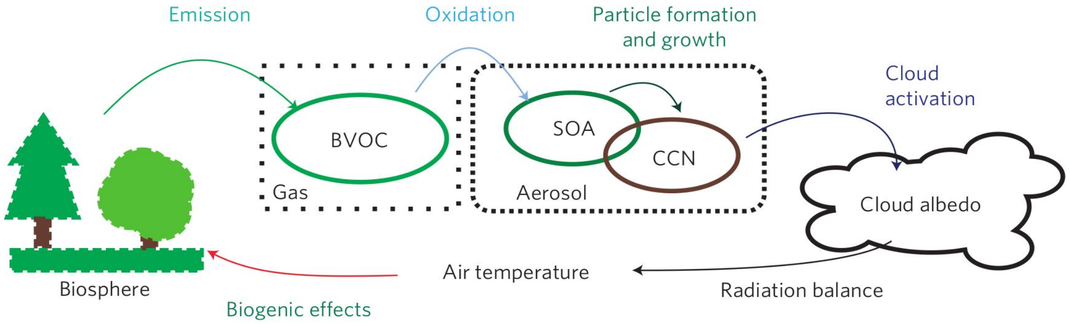


Source: Wikimedia Commons, IPCC report

Need

- Client studies cloud formation from aerosols, particularly cloud condensation nuclei (**CCN**)
- Number concentrations of particles with dry diameters larger than 100nm (**N100**) can be used as a proxy of **CCN** number concentrations

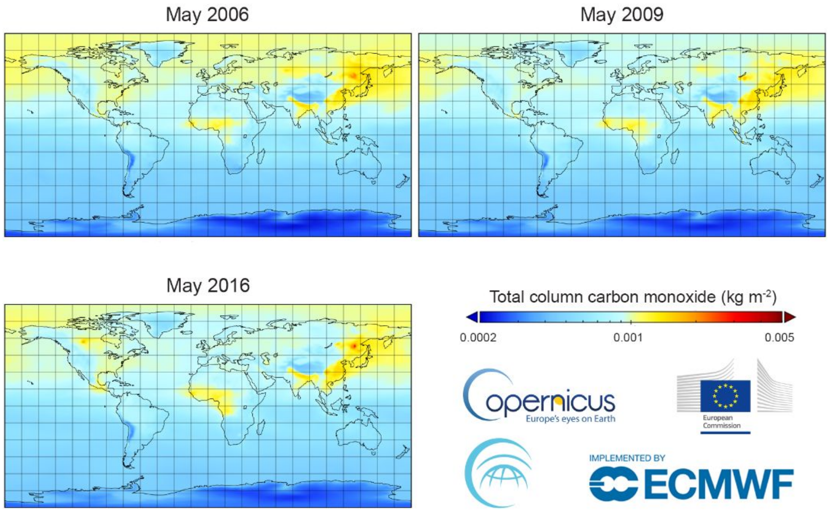
Task: build a model that predicts N100 concentrations



Source: Paasonen, P., Asmi, A., Petäjä, T. *et al.* Warming-induced increase in aerosol number concentration likely to moderate climate change. *Nature Geosci* 6, 438–442 (2013). https://doi.org/10.1038/ngeo1800

Approach

- We model N100 levels using ECMWF CAMS reanalysis data:
  - **Carbon Monoxide** (tracer for anthropogenic aerosol emissions)
  - **Temperature** (tracer for biogenic aerosol formation)
- Create and compare different models



Source: copernicus.eu

Benefit

- Ability to approximate N100 levels using CAMS reanalysis data only
  - CAMS data is free and available for the entire planet at high temporal resolutions
  - Directly measuring N100 concentrations is very expensive, difficult, and location specific
- More detailed aerosol data might improve climate model accuracy

Source: Wikipedia

Model

Data

Train set: CAMS reanalysis ECMWF (satellite)

- Carbon Monoxide CO
- Temperature T
- Nitrogen oxide NO
- Nitrogen dioxide NO_2
- Sulphur dioxide SO_2
- Terpenes $C_{10}H_{16}$
- Isoprene C_5H_8

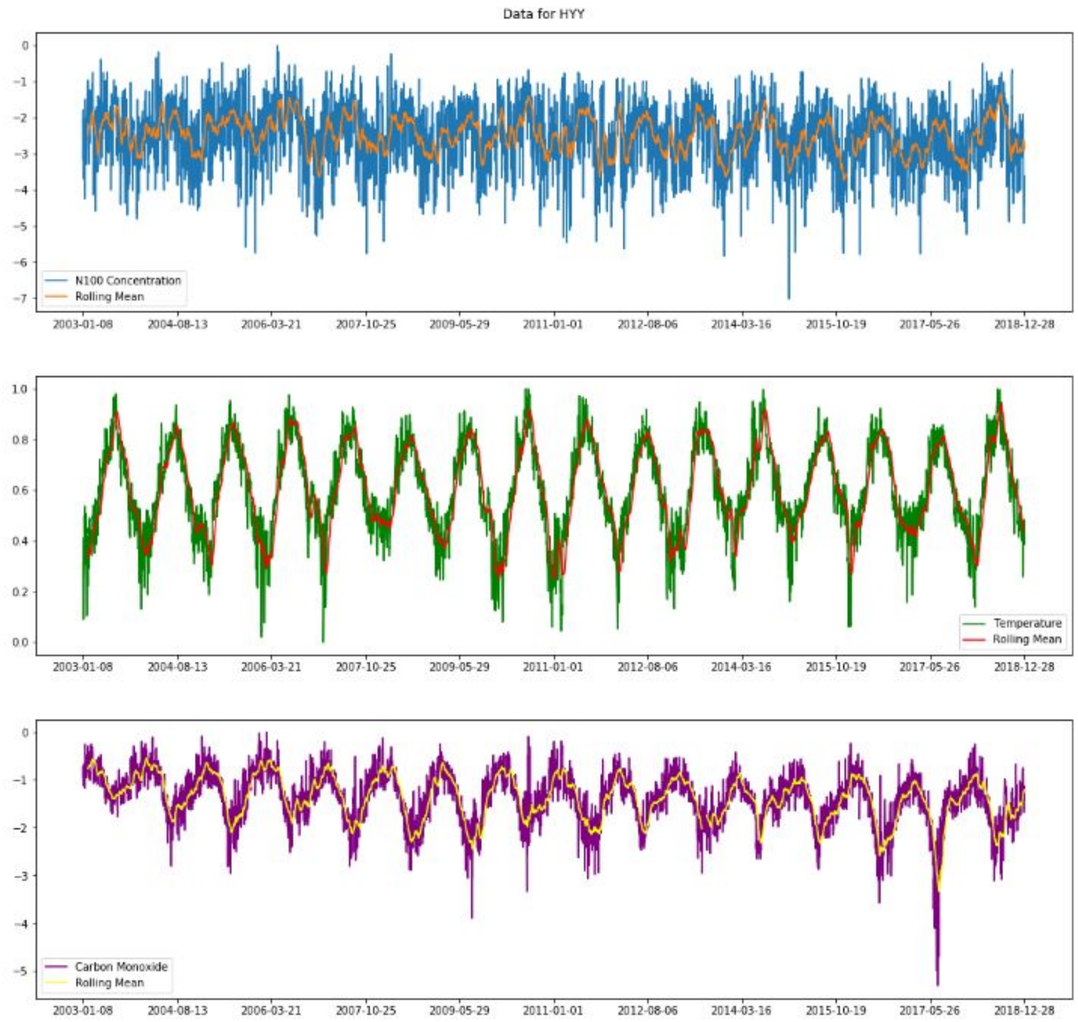
Target: in situ by INAR (22 sites spread across the globe)

- N100

Variables

- **6 inputs**
  - **Temperature**
    - Min-max scaled to [0.0, 1.0]
    - Previous week average (pwa) of min-max scaled temperature
  - **Carbon monoxide concentration**
    - Log-transformed (original data has strong positive skew)
    - Pwa of log-transformed carbon monoxide concentration
  - **Date**
    - Sine of decile* of the year (better performance than days, weeks, months or seasons)
    - Cosine of decile* of the year (together sine and cosine of decile create a “circle” of deciles)
- **1 output**
  - **N100 concentration**
    - log-transformed (original data has strong positive skew)

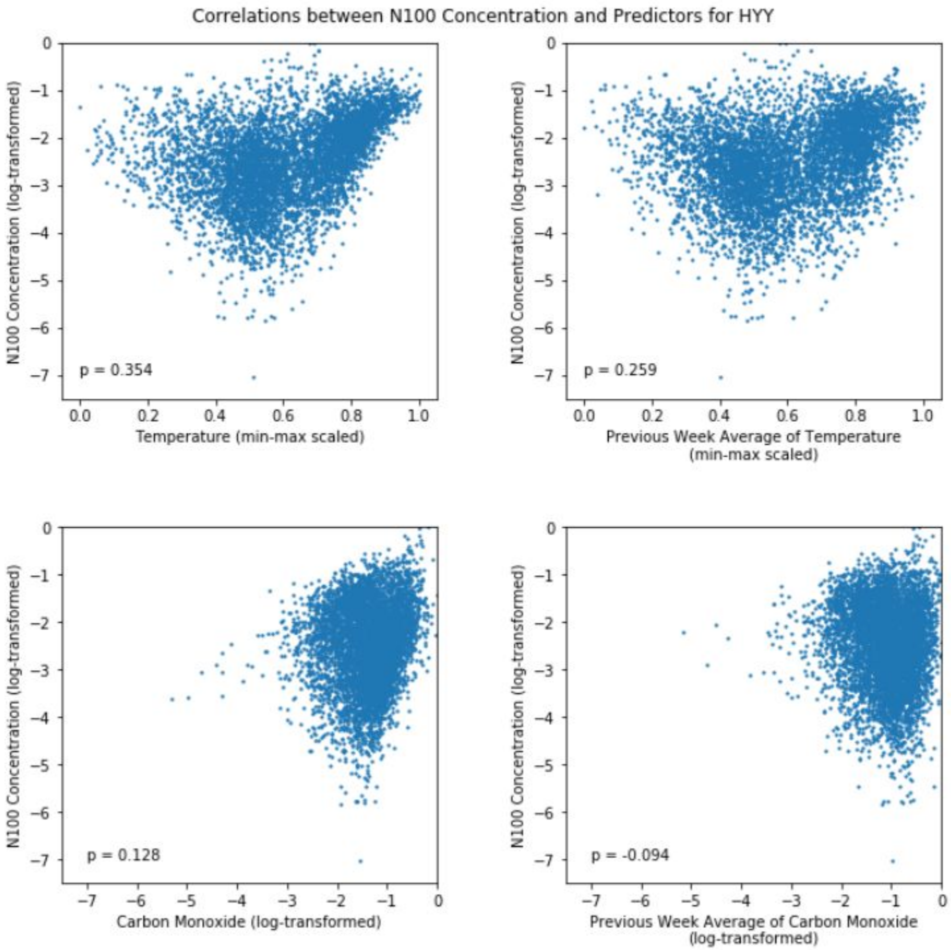
*Note: deciles are from here on referred to as seasons



Data for Hyytiälä, Finland

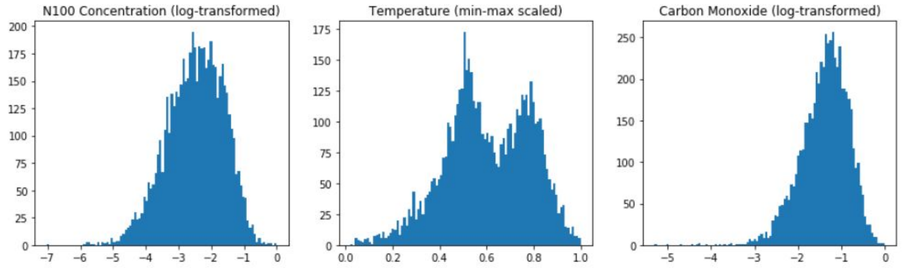
- Top: N100 concentration (log-transformed)
- Middle: min-max scaled temperature
- Bottom: carbon monoxide concentration (log-transformed)

The full dataset contains data from 22 sites around the world. Hyytiälä is the site with the longest record and clearest signal.



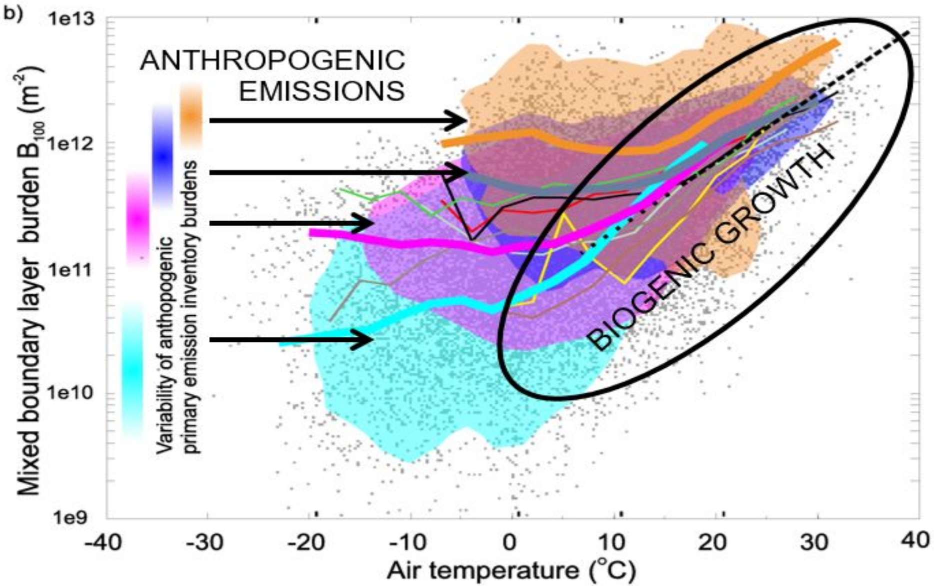
Correlations between some of the input variables and the N100 concentration

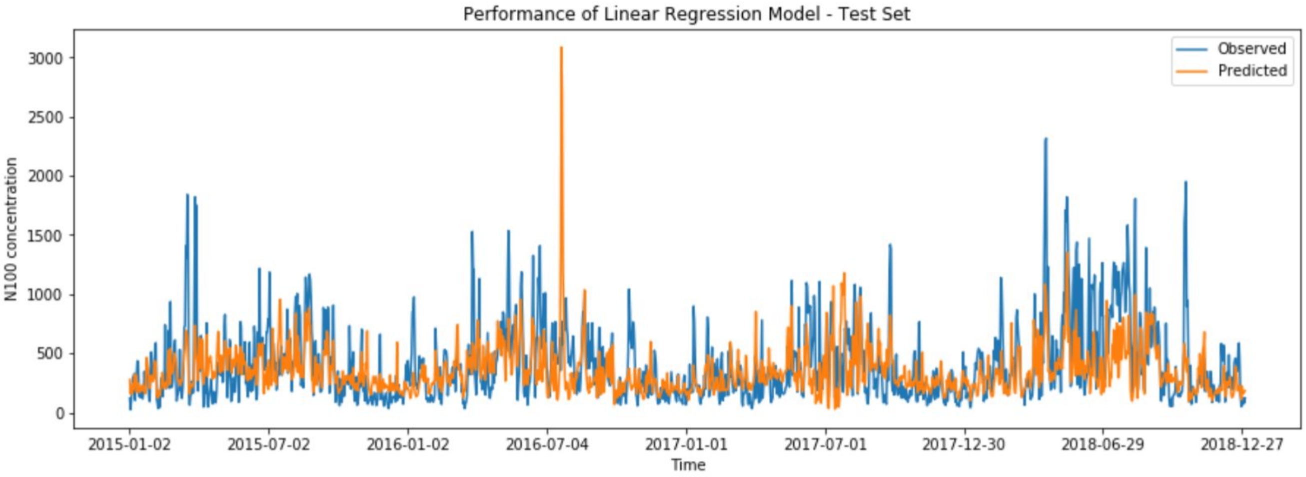
- Strongest correlation between temperature and N100 concentration
- Temperature vs. N100 plots show two distinguishable centers (also visible in histogram of temperature values)
- CO data shows only one center

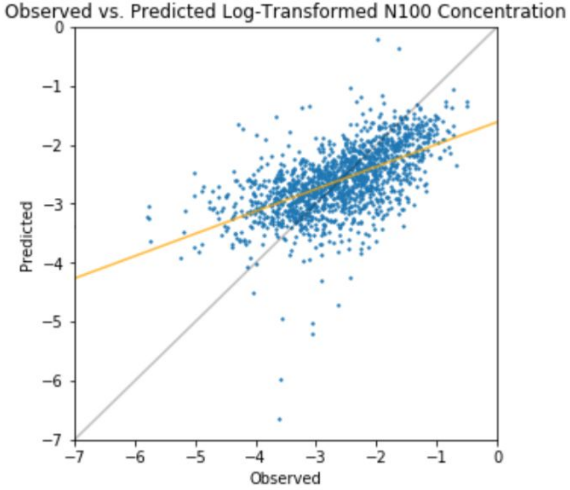


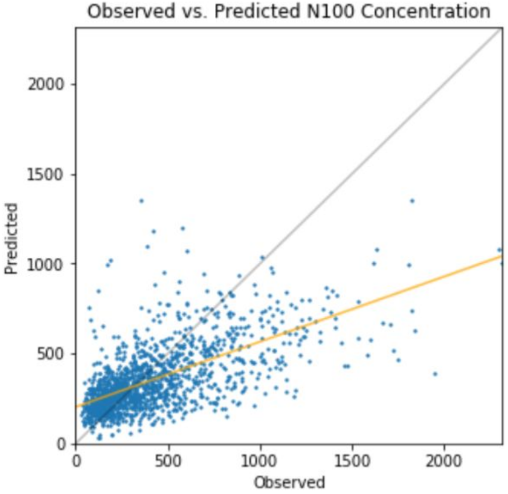
Modeling

- Anthropogenic emissions keep the aerosol level stabile, when temperature is low
- When temperature rises, biogenic growth takes place
- We are modeling these properties of aerosol levels









Linear Regression Model

Performance on test set

- **R2 score:** 0.292
- **RMSE:** 271.779

Correlation between observed and predicted N100 concentration

- **log-transf.** 0.584
- **actual:** 0.571

Equation:

log(N100) =

2.275 * min_max(T)

- 1.111 * min_max(T_pwa)

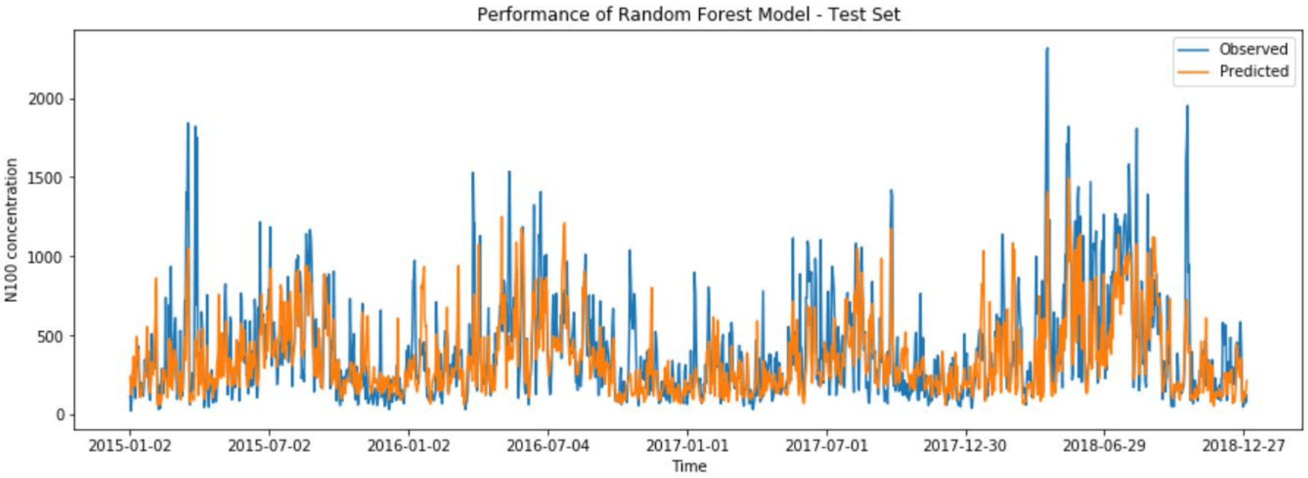
+1.456 * log(CO)

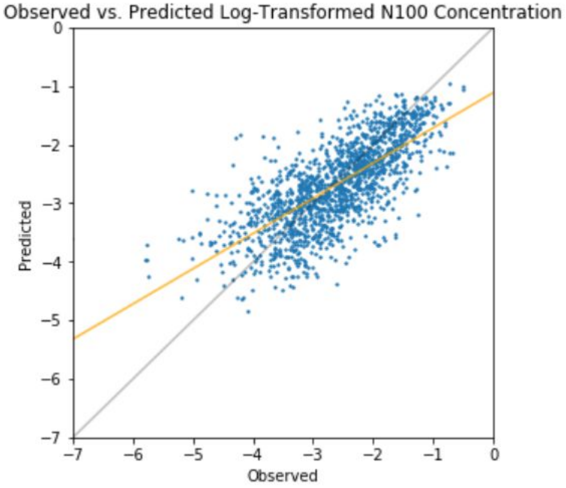
- 0.703 * log(CO_pwa)

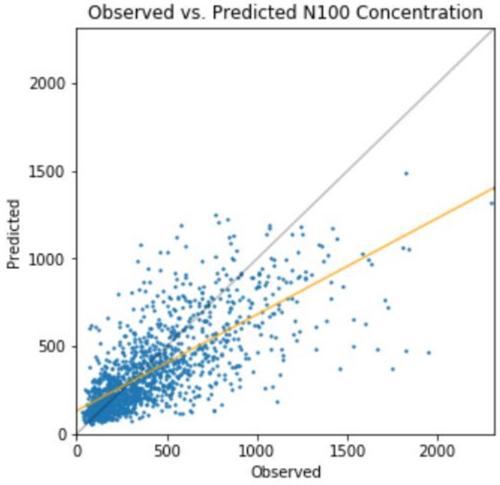
+0.139 * sin(season)

- 0.493 * cos(season)

- 2.0







Random Forest Model

Performance on test set

- **R2 score:** 0.514
- **RMSE:** 225.210

Correlation between observed and predicted N100 concentration

- **log-transf.:** 0.722
- **actual:** 0.734

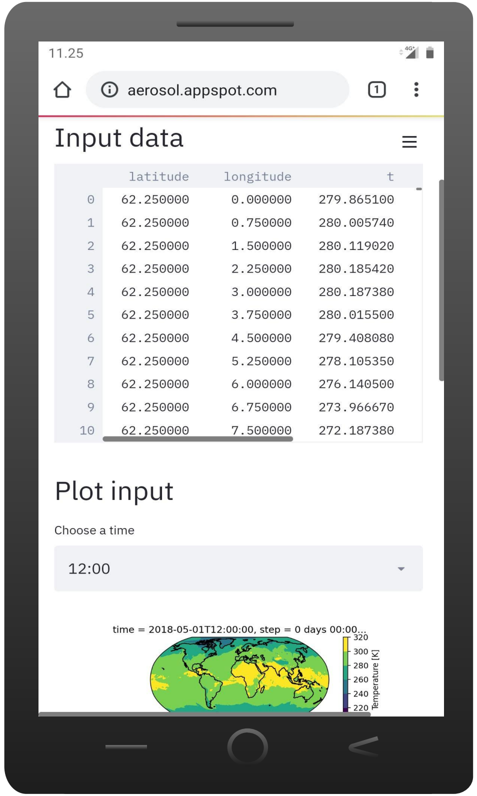
Model added for comparison. Gives much better results but is not interpretable (black-box algorithm).

Results

- Predictions of Linear Regression Model mostly follow the observed values well
  - However, predictions are not as good as of more advanced models like random forest
- Main points to improve
  - High N100 concentration values are often underestimated
  - Biggest errors occur in the summer months
- Other ideas for improving the current model
  - Exploit two peaks in temperature data through e.g. training two models on subsets of the data and combining them later
    - When the temperature is high, CO should be almost irrelevant
  - Removing outliers in the data before min-max scaling
  - Adding precipitation or boundary layer height data to the model

Proof of Concept

@ aerosol.herokuapp.com



Future plans/ideas

- Improve the Linear Regression Model
- Try out other interpretable models
  - e.g. Bayesian regression (using STAN)
- Changes to data
  - Increase the time resolution
  - Add new predictors (e.g. boundary layer height or precipitation).
- Finish the web-app