HDSLEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

JÜLICH Forschungszentrum

Coupling regional air quality simulations of EURAD-IM with street canyon observations - a machine learning approach

Charlotte Neubacher[1], Philipp Franke[1], Alexander Heinlein[2], Alex Klawonn[3], Astrid Kiendler-Scharr[1], Anne Caroline Lange[1]

[1] Forschungszentrum Jülich, Institute of Energy and Climate Research - Troposphere (IEK-8), Jülich, Germany
[2] Delft University of Technology, Institute of Applied Mathematics, Delft, The Netherlands
[3] Universität zu Köln, Institute for Mathematics, Köln, Germany

State of the Art Atmospheric Chemistry Transport Model

The EURAD-IM (EURopean Air pollution Dispersion-Inverse Model) is a state of the art atmospheric chemistry transport model on regional scales. It simulates physical and chemical processes in the atmosphere to predict the dispersion of air pollutants. With its 4D-var data assimilation application, detailed analyses of the air quality can be conducted. These allow for the improvement of initial atmospheric states as well as emission source strength assessments [2]. EURAD-IM simulations can be nested to a spatial resolution of 1×1 km^2. However, this does not represent the inner city scale.

Integration of Informed ML Module towards a Hybrid Model

- Inner city scales of air quality and pollution (≤ 100 m) are mainly influenced by
  - heterogeneous anthropogenic emission sources (traffic, heating, etc.)
  - highly reactive atmospheric chemistry [1]
  - micro-meteorology
- **Aim:** Exploring street canyon observations with additional relevant environmental features (e. g. street architecture, traffic density or meteorology) by using informed machine learning (ML)
- **Tool:** Hybrid model: coupling ML module to EURAD-IM

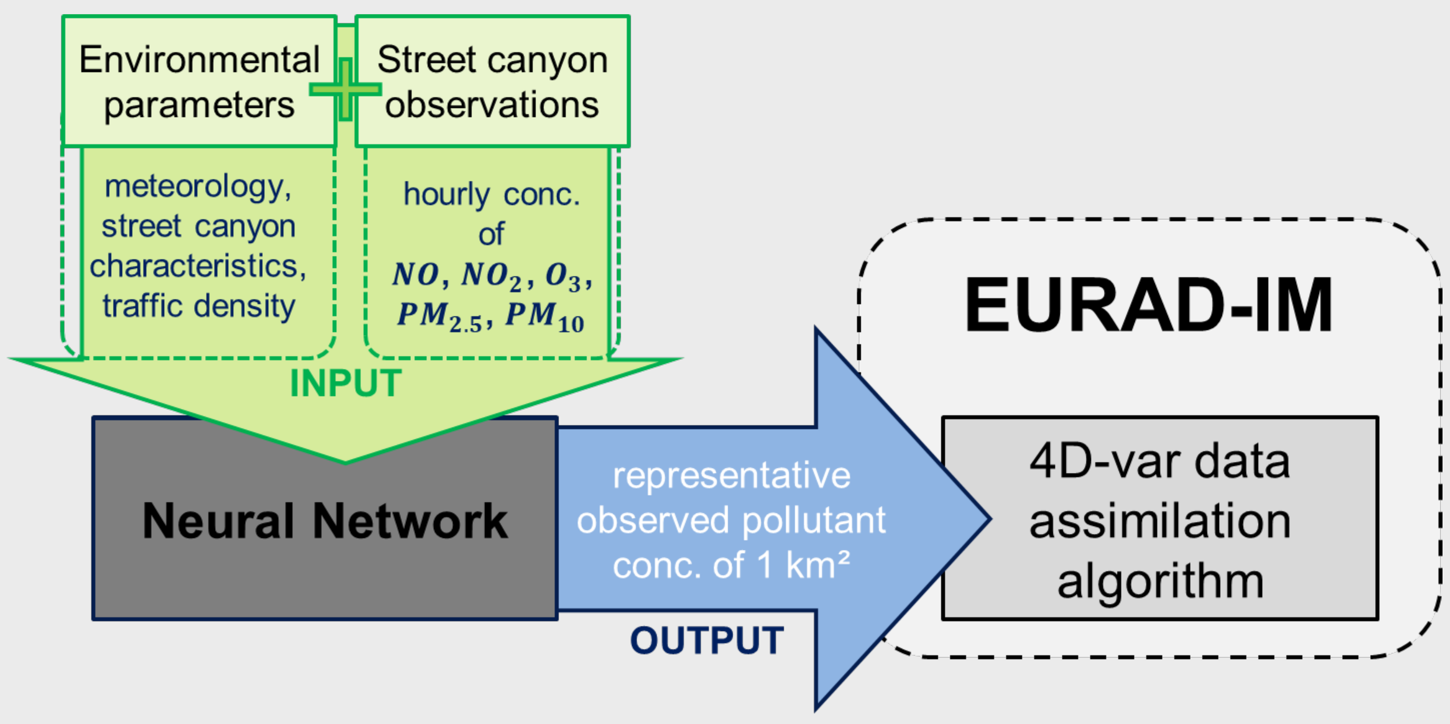


Figure: Scheme of Hybrid Model: ML module is embedded in 4D-var data assimilation system of EURAD-IM.

Test Environment towards the Targeted Hybrid Model

- EURAD-IM forecast in 1 km and 3 km resolution (Rhine-Ruhr area)
  - Combine 3 km grid cell with nine overlaying 1 km cells
  - Predicting 3 km concentrations with central 1 km values as artificial observations

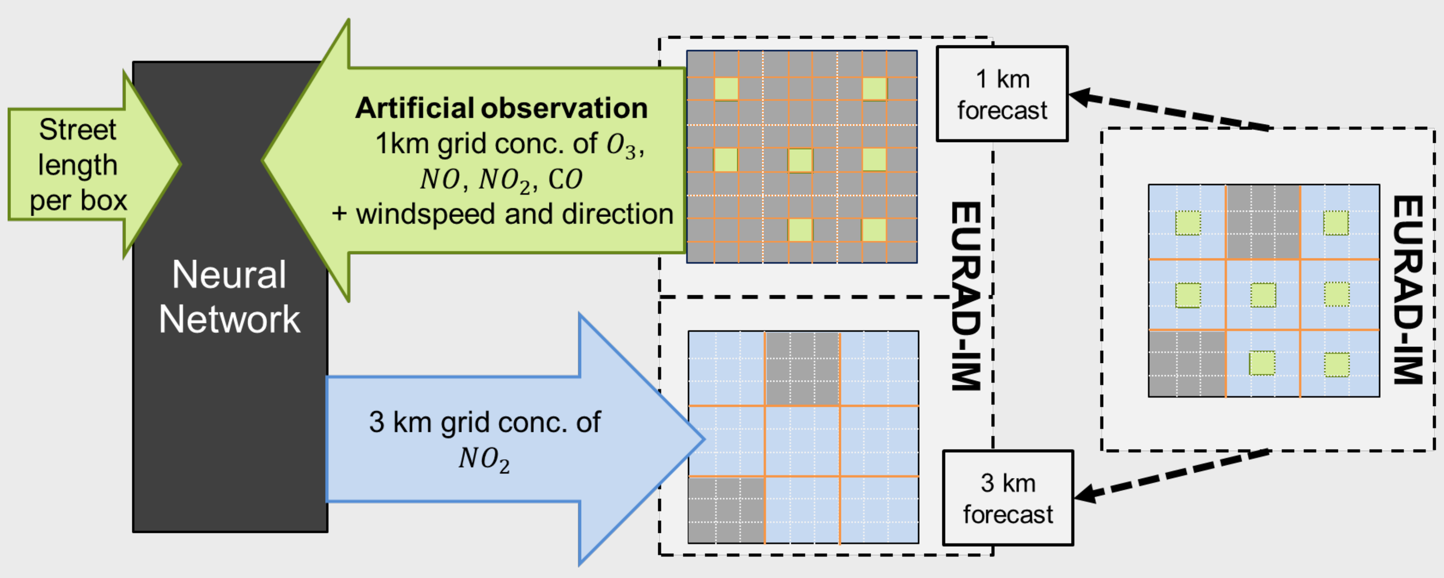
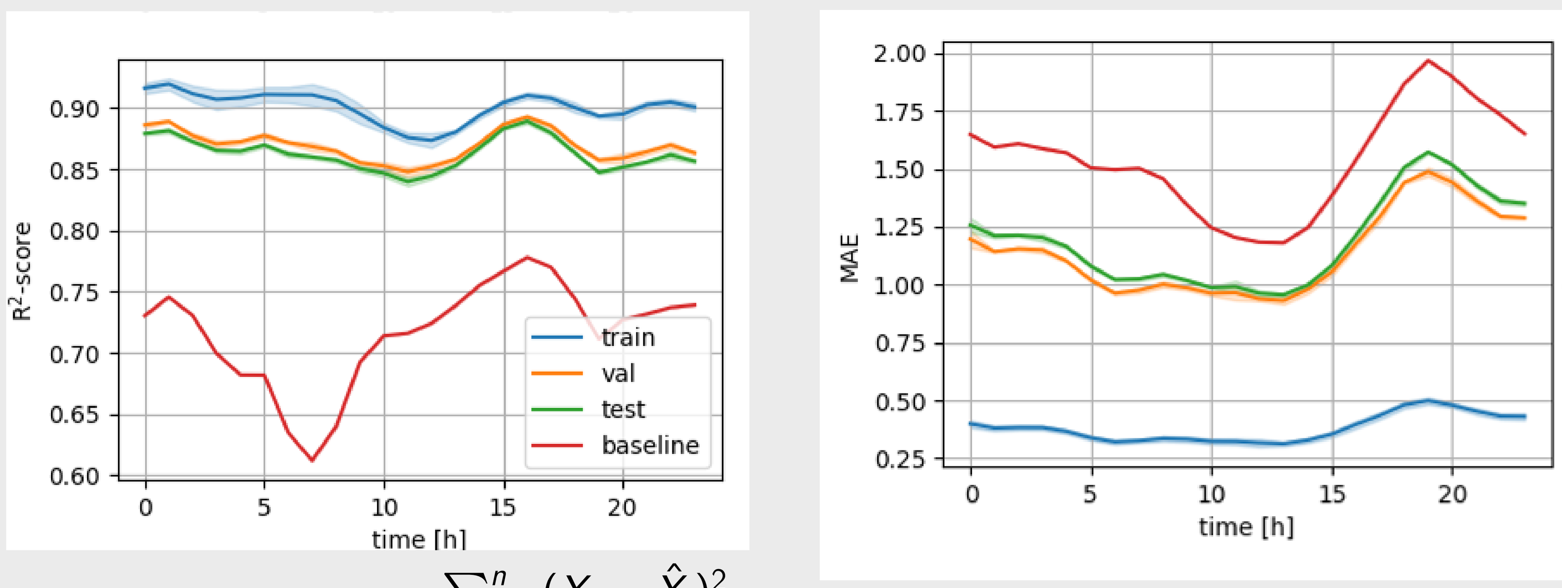


Figure: Scheme test environment

⇒ **Most relevant features** (found by feature importance analysis): NO, NO_2, O_3, CO, wind speed and street lengths

ML Model Errors

- For unseen data sets: R^2-score ≈ 0.87, MAE ≈ 1 ppb
- For ML model, R^2-score and MAE are better than for baseline model → but performance of training data set is much better than performance of test and validation data set especially for MAE



(a) $R^2(X, \hat{X}) = 1 - \frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ (b) $MAE(X, \hat{X}) = \frac{1}{N}\sum_{i=1}^{N}|X_i - \hat{X}_i|$

Figure: Average error metric: a) R^2-score, b) mean average error (MAE).

- Scatter plots show similar behavior for all subsets:
  - ML model prevents overprediction
  - weak performance above targeted concentration of 50 ppb
  - huge underprediction

  ⇒ data point distribution: dense ≤ 50 ppb, and sparse > 50 ppb

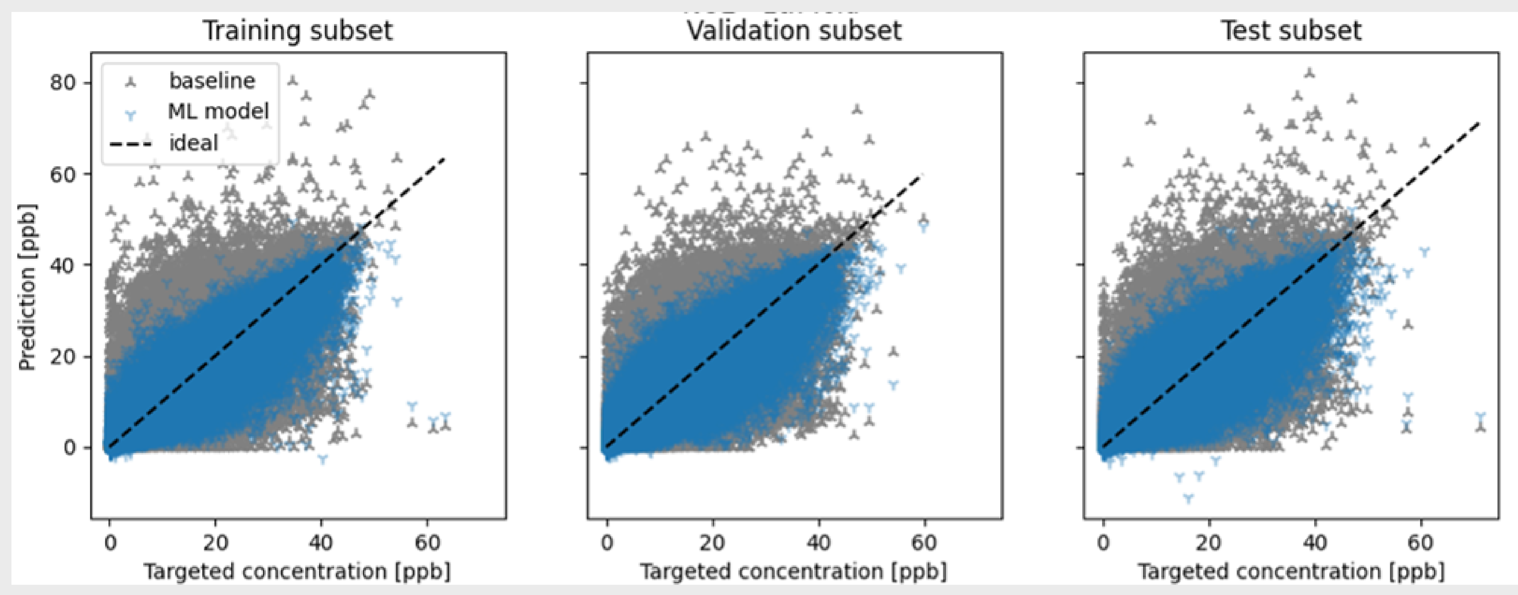


Figure: Scatter plots: predicted vs. targeted concentration, comparing ML model with baseline model for training, validation and test subset.

Baseline Model: artificial observation of 1 km concentration equal 3 km concentration

Conclusion

- Difficult data set: wide spread of values, not equally distributed → high concentrations of particular interest
- Average errors suggest overfit regime → no improvement through hyper parameter optimization

Outlook and Future Plans

- Transfer model from test environment to more complex street canyon observation
  - No labeled data
  - Detailed environmental parameters needed
- Acquire environmental parameters for observation sites
- GAN like structure to generate representative observed pollutant conc. of 1 km^2
  - Based on test environment model as pre-trained generator

References

[1] I Düring et al. "Update of the Romberg-approach and simplified NO/NO2 conversion model under consideration of direct NO2-emissions". In: *13th Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes* (Paris, France, June 1–4, 2010). 2010.

[2] H Elbern et al. "Emission rate and chemical state estimation by 4-dimensional variational inversion". In: *Atmospheric Chemistry and Physics Discussions* (2007).