Artificial intelligence for air quality

Martin Schultz, Felix Kleinert, Lukas Leufen, Clara Betancourt, Sabine Schröder, Bing Gong, Scarlet Stadtler, Michael Langguth and Amirpasha Mozaffari
Jülich Supercomputing Centre, Research Centre Jülich, Germany

Artificial intelligence (AI) is experiencing a wave of enthusiasm since ground-breaking results have been published on cognitive problems such as image and speech recognition, automated language translation, robotics, and strategic games.

Adobe Stock © NordCry

This has become possible because of recent advances in massive data processing capabilities ('big data') and the development of deep learning (DL) neural network architectures, which 'learn' millions of parameters. Even though machine learning, in general, has been in use for many years, the uptake of DL in environmental science has been slow, and IntelliAQ is one of the first projects in the atmospheric sciences to fully embrace the potential of modern big data processing and DL. IntelliAQ aims at shifting the analysis of global air pollutant observations to a new level and will provide a basis for the future development of innovative air quality services with a robust scientific underpinning.

Since the start of the IntelliAQ project in 2018, the development of machine learning solutions in all areas of environmental sciences has exploded. Long-term doubts that deep neural networks might not be able to faithfully represent the complex, multi-dimensional and multi-scale patterns embedded in environmental data have been replaced with a cautious optimism inspired by recent breakthroughs, especially in the area of weather forecasting (e.g. Scher & Messori, 2019; Weyn, Durran & Caruana, 2020). Numerous studies are exploring state-of-the-art machine learning techniques for analysing weather and climate data (e.g. Callaghan *et al.*, 2021) or developing new approaches to weather forecasting (e.g. Gong *et al.*, 2021, submitted). In a recent position paper, my team and I tried to connect the state-of-the-art in meteorological forecasting with the state-of-the-art machine learning and asked the question, "Can deep learning beat numerical weather prediction?" (Schultz *et al.*, 2021).

Air pollution is in many ways closely related to weather. From a data science perspective, both applications deal with spatiotemporal patterns and non-Gaussian data distributions, and the geometries and formats of meteorological and air pollution datasets are quite similar. Air quality is determined by several factors, including air pollutant emissions, chemical transformations, transport processes and weather. To analyse and understand air quality data and assess changes in air pollution levels, all of these factors must be taken into account. Air pollutant concentrations exhibit complex, time-dependent spatial patterns. Therefore, complex DL architectures and comprehensive datasets are needed when we want to use AI for the analysis of air quality and build air quality predictions based on modern machine learning. Furthermore, it is important to evaluate the machine learning results with proper statistical metrics. Meteorologists have developed a large arsenal of suitable metrics that differ from the standard evaluations applied in classical machine learning applications such as language or image processing.

The IntelliAQ project has positioned itself at the forefront of deep learning for the analysis of air quality information on the global scale. Specifically, IntelliAQ has three main objectives:

1. to develop novel spatial and temporal interpolation methods using deep neural networks in order to expand the coverage of historic and recent data while preserving fine-scale structures down to the street level
2. to develop an innovative air quality forecasting concept based on deep learning
3. to explore the use of deep neural networks to assess the quality of air pollution data and establish new, robust techniques for automated outlier detection and data screening.

Our research has primarily addressed tropospheric ozone, which is the second most important air pollutant with adverse impacts on human health (WHO, 2021), vegetation (e.g. Unger *et al.*, 2020) and climate (IPCC, 2021). A central asset of the project is the world's largest collection of ground-level ozone data in the database of the Tropospheric Ozone Assessment Report hosted at the Jülich Supercomputing Centre, the home of IntelliAQ (Schultz *et al.*, 2017). To account for the complexity of the multi-scale spatial and temporal interactions of tropospheric ozone, we have structured our research according to three main conceptual lines (Figure 1). In the following, I will summarise what we achieved in each of these areas. To conclude, I will then provide a glimpse into our plans to bring these three lines together and build one large, coherent analysis and forecasting system for tropospheric ozone during the remaining 21 months of the project.

Figure 1: Conceptual representation of the three methodological approaches pursued in IntelliAQ. Top: time series forecasting, centre: spatial interpolation (mapping), and bottom: the application of advanced video prediction models to capture spatial and temporal patterns.

Time series analysis

In light of the scattered geographic locations of air quality measurement sites and the available multi-variate long-term observation series from the TOAR database, the analysis of time series data, and specifically the attempt to forecast ozone concentrations with neural networks, appeared a natural first choice when the IntelliAQ project began. We explored different machine learning concepts and have investigated several strategies for the preparation of input data to test the limits of a purely data-driven approach for this task.

Kleinert *et al.* (2021) used an inception block architecture and trained a single neural network with ten years of daily observations from around 300 German air quality monitoring sites to forecast daily maximum eight-hour average ozone concentrations over four days into the future. The model made use of information from nine variables (three chemical and six meteorological parameters) and showed very good generalisation and reasonable forecast skill. It clearly outperformed a simple ordinary least square regression model, a persistence forecast and a climatological forecast. At the time of publication, this was the largest neural network that has been trained with air quality data. However, some deficits of the neural network became apparent, namely the strong reduction of forecast quality after two days and the over-emphasis of ozone concentrations near average values. The latter implied that peak air pollution would rarely be predicted correctly.

To improve on the achievable forecast length, Leufen, Kleinert and Schultz (2021, submitted) explored the use of time filters and hourly observations as inputs, which yielded a substantial improvement (Figure 2). Kleinert *et al.* (manuscript in preparation) looked at ways to incorporate upstream information from sites that have seen the same air mass earlier. They developed a wind sector approach. This was first tested on gridded data from a chemistry transport model to avoid extra complications when dealing with irregularly spaced and missing data values. The sector approach also led to improved forecasts over a longer period. Finally, a master thesis tested the use of oversampling approaches to better capture extreme values in the ozone prediction. This had mixed success because the better hit rate of air quality threshold exceedances was paid with increased bias and false alarms.

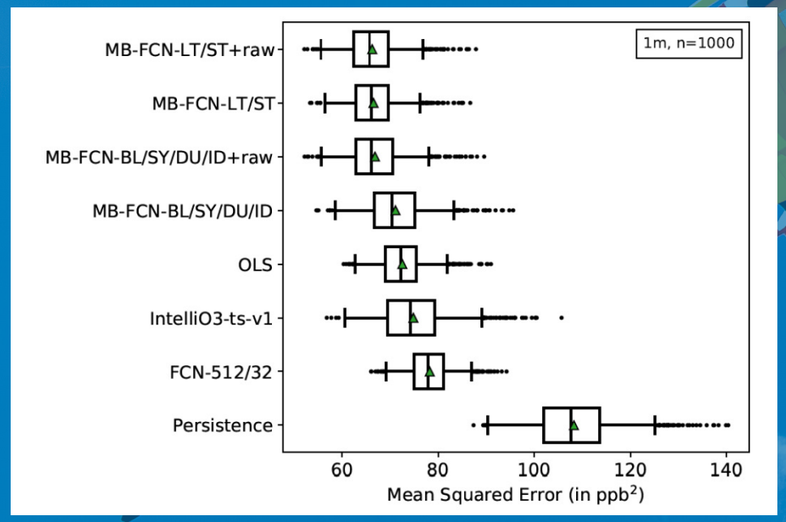

Figure 2: Error metrics (mean squared error) of different neural network and reference models predicting the daily maximum 8-hour ozone average concentrations at 32 air quality monitoring sites in Northern Germany (training was performed on the data from 55 sites). Lower MSE values are better. The top four boxes and whiskers are results from a model with time filtering. OLS is a linear least square model and IntelliO3-ts-v1 is the original model described in Kleinert, Leufen and Schultz (2021). FCN-512/32 is a feed forward network without time filtering and Persistence is a persistence forecast, simply repeating the last values. Figure from Leufen, Kleinert & Schultz, 2021, submitted.

A user-friendly, adaptable software tool for time series forecasting through deep learning was developed and published together with the source code (Leufen, Kleinert & Schultz, 2021). Future work on time series forecasting will look into making use of meteorological predictions (instead of prescribing only data until 'now') and of a rich variety of geospatial data that has been exploited in the spatial mapping approach described next.

Mapping

Here, we tested the use of a wide variety of freely available geospatial datasets (e.g. population density, digital elevation models, nighttime light intensity, landcover classes) to infer annual ozone average concentrations at locations without monitoring sites. A novel mapping method based on random forests has been developed, where the high-resolution geospatial datasets serve as predictors for the air quality metrics calculated from the TOAR database (Betancourt *et al.*, 2021). The study showed that about 60 per cent of the ozone variability can be explained by the geospatial predictors. In a follow-up study (Betancourt *et al.*, in preparation), the mapping approach was extended across the whole globe (Figure 3). Recent techniques which make machine learning explainable (e.g. Lundberg & Lee, 2017; Meyer & Pebesma, 2020) were employed and further improved to assess the robustness and credibility of the generated global ozone concentration maps. The combination of these techniques is unique and can serve as a blueprint for future mapping studies with machine learning, which are not limited to air quality-related topics (Stadtler *et al.*, 2021, submitted).

Video prediction

Initial attempts to apply relatively simple video prediction methods based on convolutional neural networks to forecasting spatiotemporal weather fields yielded unsatisfactory results. The resulting fields looked rather blurry, and important features (i.e. gradients in the data) were lost. This has led to the adoption of generator models (Goodfellow *et al.*, 2014), which draw samples from a learned distribution and can therefore produce sharp and realistic looking images.

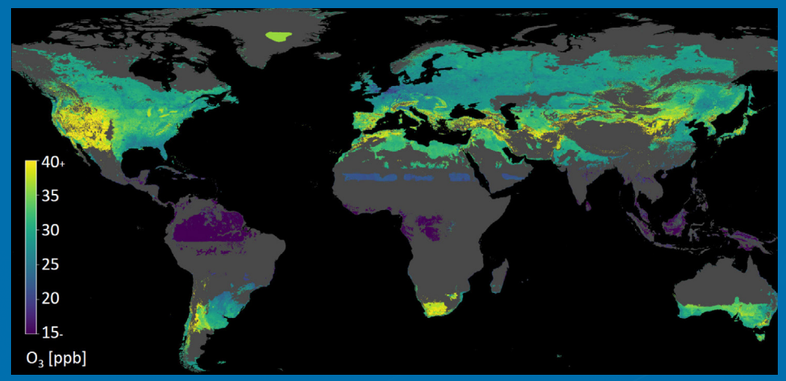

Figure 3: High-resolution map of annual mean ozone concentrations estimated based on the spatial mapping of several geospatial datasets to ozone concentrations measured at several thousands of stations. Grey areas are regions where explainable AI techniques indicated too little confidence in the extrapolation results. Figure from Betancourt et al., manuscript in preparation.

In Gong *et al.* (2021, submitted), we explored a combination of GANs with a variational autoencoder model (SAVP, Lee *et al.*, 2018) for use as a meteorological forecasting model. The model was trained on 11 years of reanalysis data and generated skilful predictions of the 2-metre temperature over Europe over 12 hours. Even though the model was only trained with temperature at two metres and at 850 hPa, it clearly outperformed a simpler convolutional LSTM model and a persistence forecast. However, the SAVP model exhibited some difficulties in correctly predicting the temperatures over mountains, and the model error is still substantially larger than that of state-of-the-art weather prediction models. We expect that additional input variables and an embedding of the surface orography would lead to improved forecasts, but to beat numerical weather prediction, one would also have to find ways to incorporate some physical knowledge into the machine learning tools.

Synthesis

The different deep learning approaches described above have led to valuable insights with respect to the preparation of input data, the strengths and weaknesses of different deep learning architectures and the technical hurdles to implement and train complex deep learning models on high-end supercomputer systems. The ultimate goal of the IntelliAQ project is to design a deep learning model which can either generate trustworthy maps of global air pollution based on the heterogeneous and scattered data that are available or produce skilful forecasts of air pollutant concentrations over several days. It appears that so-called transformer models (e.g. Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2020) could be trained to learn spatiotemporal representations of atmospheric data and then be employed in various ways to accomplish the interpolation and forecasting tasks that were defined in the IntelliAQ proposal. Such transformer models belong to the class of unsupervised machine learning systems, and they require substantial skill and computing power to achieve competitive results. We accept the challenge and hope that we can bring deep learning for air quality to a level where it outperforms state-of-the-art chemistry transport models in a variety of applications.

References click here

IntelliAQ

PROJECT NAME
IntelliAQ

PROJECT SUMMARY
IntelliAQ is an ERC Advanced Grant project to explore the application of cutting-edge machine learning techniques to global air quality data in combination with high resolution geospatial and weather data. It combines novel data management and data science approaches to build the foundation for innovative air quality information services.

PROJECT PARTNERS
The IntelliAQ project is hosted by the Jülich Supercomputing Centre at Forschungszentrum Jülich and benefits from the world-class high-performance computing infrastructure at this institution. IntelliAQ has strong ties to the international Tropospheric Ozone Assessment Report activity and thus involves broad scientific collaboration.

PROJECT LEAD PROFILE
Dr Schultz works at the interface between atmospheric and computer science. He obtained his PhD at Forschungszentrum Jülich in 1995 and worked at Harvard University and the Max Planck Institute for Meteorology before he returned to Jülich in 2006. Since 2017 he established the research group Earth System Data Exploration, which develops new machine learning methods for Earth system science at the Jülich Supercomputing Centre.

CONTACTS
Dr Martin G. Schultz
Jülich Supercomputing Centre,
Forschungszentrum Jülich, 52425 Jülich,
Germany.
+49 2461 61 96870
m.schultz@fz-juelich.de
https://go.fzj.de/martinschultz

FUNDING
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No.787576.