

## DESIGN AND RESULTS OF AN AI-BASED FORECASTING OF AIR POLLUTANTS FOR SMART CITIES

L. Petry<sup>1</sup>, T. Meiers<sup>2</sup>, D. Reuschenberg<sup>2</sup>, S. Mirzavand Borujeni<sup>2</sup>, J. Arndt<sup>2</sup>, L. Odenthal<sup>2</sup>, T. Erbertseder<sup>3</sup>, H. Taubenböck<sup>3</sup>, I. Müller<sup>3</sup>, E. Kalusche<sup>3</sup>, B. Weber<sup>4</sup>, J. Käflein<sup>5</sup>, C. Mayer<sup>6</sup>, G. Meinel<sup>1</sup>, C. Gengenbach<sup>7</sup>, H. Herold<sup>1</sup> \*

<sup>1</sup> Leibniz Institute of Ecological Urban and Regional Development (IOER), Dresden, Germany - (l.petry, g.meinel, h.herold)@ioer.de

<sup>2</sup> Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI), Berlin, Germany - (thomas.meiers, david.reuschenberg, sara.mirzavand.borujeni, jost.arndt, luise.odenthal)@hhi.fraunhofer.de

<sup>3</sup> German Remote Sensing Data Center, German Aerospace Center (DLR), Weßling-Oberpfaffenhofen, Germany - (thilo.erbertseder, inken.mueller, elena.kalusche, hannes.taubenboeck)@dlr.de

<sup>4</sup> Hof University of Applied Sciences (HSH), Hof, Germany - beatrix.weber@hof-university.de

<sup>5</sup> Geomer GmbH, Heidelberg, Germany - julian.kaeflein@geomer.de

<sup>6</sup> Meggsimum, Mutterstadt, Germany - chris@meggsimum.de

<sup>7</sup> Software AG, Darmstadt, Germany - christian.gengenbach@softwareag.com

**KEY WORDS:** Air Pollutants, Forecasting, Artificial Intelligence, RNN, Smart City, Co-Design

### ABSTRACT:

This paper presents the design and the results of a novel approach to predict air pollutants in urban environments. The objective is to create an artificial intelligence (AI)-based system to support planning actors in taking effective and adequate short-term measures against unfavourable air quality situations. In general, air quality in European cities has improved over the past decades. Nevertheless, reductions of the air pollutants particulate matter (PM), nitrogen dioxide (NO<sub>2</sub>) and ground-level ozone (O<sub>3</sub>), in particular, are essential to ensure the quality of life and a healthy life in cities. To forecast these air pollutants for the next 48 hours, a sequence-to-sequence encoder-decoder model with a recurrent neural network (RNN) was implemented. The model was trained with historic in situ air pollutant measurements, traffic and meteorological data. An evaluation of the prediction results against historical data shows high accordance with in situ measurements and implicate the system's applicability and its great potential for high quality forecasts of air pollutants in urban environments by including real time weather forecast data.

### 1. INTRODUCTION

Air pollution is still one of the main public health issues in cities worldwide. It affects not only human health, but also the environment. Biodiversity as well as functioning terrestrial and aquatic ecosystems are threatened by air pollutants. Main sources for anthropogenic air pollution are traffic, industry, agriculture and combustion processes (EEA, 2019: 8, 15, 71 ff.; Arndt, 2012: 87; WHO, 2015; UBA, 2018). Due to legal regulations, new environmentally friendly technologies and the use of low-emission fuels, all air pollutants, except for ground-level ozone, could be decreased by at least 40 % since the base year 1990 (UBA, 2020a). Nevertheless, the European Environmental Agency focuses on three air pollutants which have currently the highest impact on human health of citizens in Germany. Particulate matter (PM), nitrogen dioxide (NO<sub>2</sub>) and ground-level ozone (O<sub>3</sub>) are responsible for serious neurological, cardiovascular and respiratory diseases (Dora et al., 2011: 2; EEA, 2019; Brunekreef and Holgate, 2002; Schneider et al., 2018). Therefore, one of the main objectives of municipal authorities and environmental agencies is to reduce air pollution with adequate measures. Because of the complex interaction between air pollutants among themselves (e.g., NO<sub>2</sub> – NO – O<sub>3</sub>) as well as with external influences (such as the weather conditions, traffic amount, and land use) it is not a trivial task to predict and simulate air pollution concentrations. However, in order to conduct adequate measures it is crucial and important

to support local decision makers in traffic, urban, and environmental planning with reliable information on the air pollution variability in time and space. The WHO air quality guidelines (WHO, 2006) assess the risks of air pollutants and suggest the following maximum annual mean values: 10 µg/m<sup>3</sup> for PM<sub>2.5</sub>; 20 µg/m<sup>3</sup> for PM<sub>10</sub>; and 40 µg/m<sup>3</sup> for NO<sub>2</sub>. The European Union enacted several agreements and directives. The Directive 2008/50/EC on ambient air quality and cleaner air for Europe, which was enacted in 2018, defines limiting and target values for each air pollutant. The daily mean of 50 µg/m<sup>3</sup> PM<sub>10</sub>, for example, is limited to 35 days per year, the annual mean should not exceed 40 µg/m<sup>3</sup>. NO<sub>2</sub> concentrations should not exceed the hourly limit value of 200 µg/m<sup>3</sup> for more than 18 hours per year, in addition the annual mean is limited to 40 µg/m<sup>3</sup>. The directive does not define limiting values for ground-level ozone. It only states a long-term objective of a maximum daily 8-hour mean of 120 µg/m<sup>3</sup> (39. BImSchV; EEA, 2019: 7, 12). Spot-based measuring stations detect and observe the air quality in Europe. Those measuring stations are mainly located in hotspot areas, such as roads and intersections with a high traffic volume (cf. Petry et al., 2020).

### 2. PROBLEM STATEMENT

Cities worldwide are facing similar problems concerning finding a balance between quality of life (including air quality) and providing the necessary traffic services, industrial and

\* Corresponding author

economic activities, which are, in turn, the main contributors to air pollution in urban areas. To employ adequate short-term measures, it is necessary to better understand the interactions between air pollutants and their influencing parameters.

There is a strong linkage between weather conditions and air pollutants. There are seasonal fluctuations (winter vs. summer) as well as short-term weather fluctuations. In general, a higher emission of PM and NO<sub>x</sub> in winter, due to low temperatures and increased heating, and an increased photochemical degradation of NO<sub>x</sub> in summer is observed (Plaß-Dülmer, 2020). During the observation period from 2015 to 2018 the Leibniz-Institute for Tropospheric Research (TROPOS) observed a strong influence of wind speed, height of mixing layer and humidity to the NO<sub>2</sub>-concentration as well as a strong influence of high temperatures and relative humidity on O<sub>3</sub>-concentration in their study areas in Saxony, Germany (Pixteren et al., 2020). Other studies observe additionally a strong linkage between global radiation and O<sub>3</sub>. High O<sub>3</sub>-concentration occur especially on hot and sunny days during summer (Stroh and Marb, 2020).

The retention period of PM in the air is strongly influenced by many weather parameters, for example precipitation, wind velocity, wind direction, nightly ground inversion and high-pressure weather conditions. Therefore, especially dry winter weather situations with limited air exchange can result in high PM-concentrations in Germany (UBA, 2020b). In addition, low temperatures lead to higher PM-concentrations due to increased emissions from domestic heating.

The complex chemical and photochemical interaction between NO, NO<sub>2</sub> and O<sub>3</sub> is not only influenced by weather parameters, but also by traffic volume. A high traffic volume, for example, results in high NO<sub>x</sub> emissions that react with O<sub>3</sub>. Nevertheless, O<sub>3</sub> concentrations are significantly higher in rural areas than in urban areas with high traffic volumes (Pixteren et al., 2020; Stroh and Marb, 2020; Behera and Balasubramanian, 2016).

The aim of this study was to design a system that forecasts pollutant values for 48 hours at the location of the pollutant measurement stations. This is important to inform decision makers for taking and announcing short-term measures, such as pop-up low emission zones, road cleaning, decreasing speed limits, increasing parking fees or free public transportation. To achieve this, a deep neural network was trained on historic air pollutant and weather data.

## 2.1 Research Questions

The main objective of this study is to develop and evaluate new AI-based approaches to support decision makers in finding the right measurements to improve air quality in cities and regions. Therefore, the following research questions are posed:

- How can an AI-based system be designed to integrate real time measurements with meteorological and traffic data to forecast and simulate the variability of air pollutants in urban environments?
- What input parameters are relevant? What is their influence on the quality and validity of the forecast?
- What variant of DNNs is best suited to model time series of air pollutants and their interdependencies with meteorological data?

## 2.2 State of the Art

Limited area models have been developed to assess and forecast air pollution concentrations considering the most relevant chemical and physical processes (e.g. Kukkonen et al., 2012, for an overview). The combined use of weather models, chemical transport models or particle dispersion models enables the spatiotemporally consistent (synoptic) simulation and forecast of chemical and physical quantities from global to regional and local scales. On the European scale, the Copernicus Atmosphere Monitoring Service (CAMS) delivers daily air quality forecasts at a spatial resolution of 0.1° (~ 10 km) (COP, 2021). The CAMS European Forecasting Ensemble is an ensemble of seven chemical transport models described in more detail by Marécal et al. (2015). For air quality forecasting at a higher spatial resolution for Germany and regions therein, the air quality modelling system Polyphemus/DLR operationally provides predictions of up to 72 hours at spatial resolutions of 2 km down to 500 m (Khorsandi et al., 2018; Mallet et al., 2007). Other air quality modelling systems applied in Germany are EURAD-IM (Elbern et al., 2007) and REMCAL-GRID (Stern, 2003).

Using data assimilation methods, models and observations from in-situ measurements or satellites can be merged to generate best estimates of the atmospheric state and to improve model forecasts (e.g. Talagrand, 1997). However, a large amount of the variability of air pollution occurs on very small scales (Lefebvre et al., 2013). Consequently, the limiting factors of chemical and physical models are: they require an accurate and highly resolved initialization of static (topography) and dynamic (meteorology, emissions, land use) fields, a good representation and parametrization of processes, a numerically stable solver and last but not least massive computing capacities to resolve effects in complex terrain or urban environments. The results are prone to model-based uncertainties that are difficult to quantify.

Using artificial intelligence this study proposes a different approach to simulate and forecast air quality at spot-based measurement stations as well as at city level. The quality of the prediction of an AI-system is based on the quality of input data. Therefore, only air pollutant data from official measuring stations with continuous data is considered, so far. For time series analyses, sequence-to-sequence models with recurrent neural network (RNN) have been widely used. RNNs are a special class of deep neural networks, which use an internal loop to propagate information between tokens of an input sequence. Since RNNs as such have problems in representing long term relationships in the input sequences due to vanishing or exploding error signals, various architectural improvements have been suggested to deal with this problem. Long-Short-Term-Memory (LSTM) (Hochreiter, 1997) and Gated-Recurrent-Unit (GRU) (Cho, 2014) both propagate an additional cell state between input tokens, that can be controlled via gating-parameters learned by the network to enable it to keep information over longer time intervals. Sequence-to-Sequence models, originally developed for language translation tasks (Sutskever, 2014, Cho, 2014), have also shown promising results in time-series-forecasting problems (Tang, 2016, Lai, 2018, Shih, 2019). The model consists of two parts: An encoder network processes past time steps and produces an encoding, the so-called hidden state. A decoder network receives this hidden state as input and produces the forecast step by step in an autoregressive loop.

### 2.3 Parameter Selection and Study Areas

The objective is to test, train and evaluate the system in two areas in Germany and to make it transferable to other areas in Germany. The study areas are the City of Stuttgart and the federal state of North Rhine-Westphalia (NRW), which have been described in Petry et al. (2020). As the data of the measuring stations in NRW is more consistent, the following examples refer to stations in NRW. The official measuring stations are classified on the basis of their location (such as urban, suburban, rural) as well as on the predominant emission source (such as traffic, industrial, or background, which are representative of the general exposure of the public and the vegetation) (EEA, 2019: 11).

The following figures display results from different measuring stations in NRW. The first station VWEL represents a typical urban station with a high volume of traffic, situated in Wuppertal-Elberfeld directly at a four-lane road. Multi-storey residential and commercial buildings are located on both sides of the street. Wuppertal is located close to Köln, Düsseldorf, Duisburg, Essen and Dortmund, which has an indirect impact on the traffic volume and emissions. The second station EIFE represents a rural background station. It is located in the municipality of Simmerath-Lammersdorf on a field path, the nearest federal road is 200 meters away. It is surrounded by spruce forests and meadows used for cattle-grazing (LANUV, 2021).

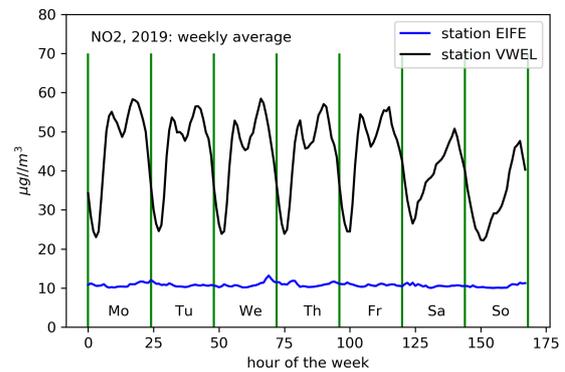
Urban traffic stations do often not collect data of  $O_3$ . Therefore, a third station, WALS, had to be selected to be able to compare  $O_3$  concentration of a rural with an urban station. WALS is located in front of a school building and in the middle of a residential area in Duisburg-Walsum. A coal-fired power station, a coking plant, a steel mill and a paper mill are nearby. This is why it is classified as an urban industrial station (LANUV, 2021).

To design a system that predicts the pollutant load for the next 48 hours, the relevant input parameters for determining the pollutant loads of NO,  $NO_2$ ,  $O_3$ ,  $PM_{10}$  have to be identified. For this purpose, the results of previous studies were used.

The main polluters are traffic, combustion, industry and agriculture. Since pollution is mostly of importance in urban regions, traffic and pollution from heating play a role here. Traffic is causally responsible for the pollution of NO,  $NO_2$  and  $PM_{10}$ . Heating systems mainly produce  $PM_{10}$ . These pollutants show a temporal characteristic that reflects the temporal behaviour of certain pollutants.

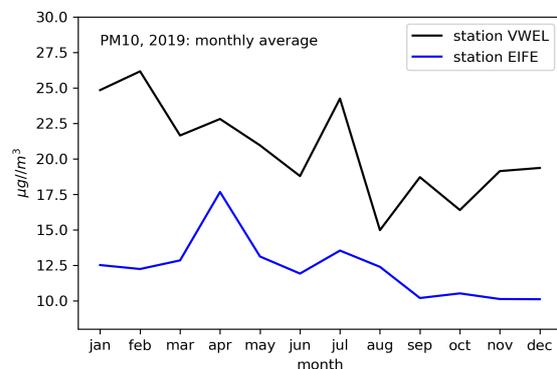
The weekly course of the pollutants NO and  $NO_2$  shows a behaviour that corresponds to the traffic volume (see figure 1; station VWEL). On weekdays there are usually two maxima representing rush hour traffic, on weekends this curve flattens out. Public holidays “behave” like Sundays.

The  $PM_{10}$  load has few fluctuations during the course of the day, but is higher in winter than in summer; this is due to the additional load caused by heating (see figure 2). The comparison of the values of the stations EIFE and VWEL shows that densely populated areas have a higher  $PM_{10}$  load.

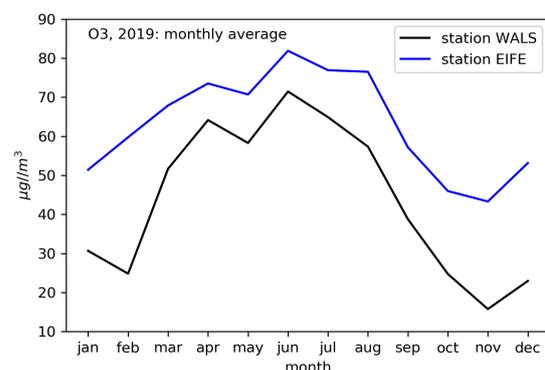


**Figure 1.** Weekly average values of  $NO_2$  (year 2019) for the urban station VWEL (high volume of traffic) and the rural station EIFE (surrounded by vegetation, no buildings, low traffic).

Ozone pollution depends on many factors; it has a specific daily pattern that is independent of the day of the week. There is rather a relationship to solar radiation and temperature. This also means that the ozone pollution is significantly higher in summer than in winter (see figure 3).



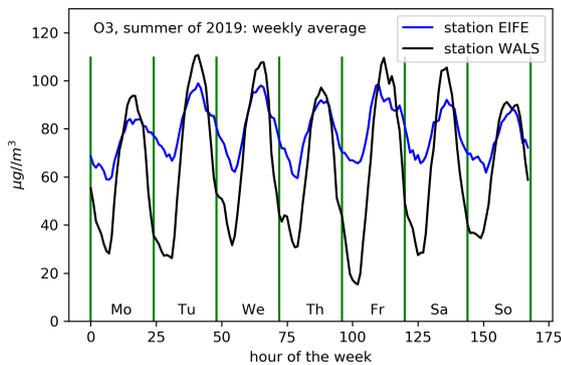
**Figure 2.** Monthly average values of  $PM_{10}$  (year 2019) for the urban station VWEL and the rural station EIFE.



**Figure 3.** Monthly average values of  $O_3$  (year 2019) for the urban station WALS and the rural station EIFE.

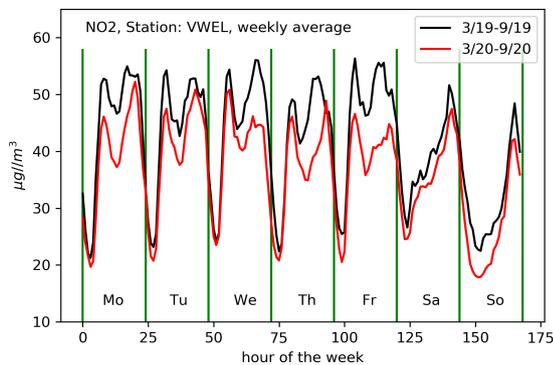
Other factors are complex chemical reactions, such as  $1 O_3 + 3 NO = 3 NO_2$  (Pixteren et al., 2020). This equilibrium reaction in turn depends on weather conditions, mainly temperature.

This effect means that ozone pollution in rural areas can be higher than in urban areas, since the NO, which is caused by traffic, binds ozone (see figure 4).



**Figure 4.** Weekly average values of O<sub>3</sub> (meteorological summer of the year 2019) for the urban station WALS (high volume of traffic) and the rural station EIFE (surrounded by vegetation, no buildings, scarcely any traffic).

The daily course correlates with a small delay with the solar radiation. However, the minima at the urban station are much lower than at the rural station. This is due to the interaction with NO, which binds O<sub>3</sub>.



**Figure 5.** Comparison of the weekly average values of NO<sub>2</sub> (from March 2019 until September 2019 versus the corona period March 2020 until September 2020) for the urban station VWEL (high volume of traffic)

Considering the long-term development is important, too, since the levels of NO, NO<sub>2</sub> and PM<sub>10</sub> have continuously decreased over the last ten years. This is probably due to more environmentally friendly technologies. The COVID-19 pandemic plays a special role here, as shown in Figure 5. The pandemic has led to a significant reduction in overall pollution (Erbertseder and Loyola, 2020). The temporal behaviour is therefore an important input parameter for predicting the pollution load. In detail, these are hour of day, day of week, holiday, month and year. Weather conditions play a significant role, directly and indirectly (Plaß-Dülmer, 2020).

The following conditions have a direct influence: temperature and radiation affect the ozone-nitrogen oxides equilibrium; precipitation washes out pollutants (UBA, 2020b); and higher wind speeds disperse pollutants faster and reduce peak loads.

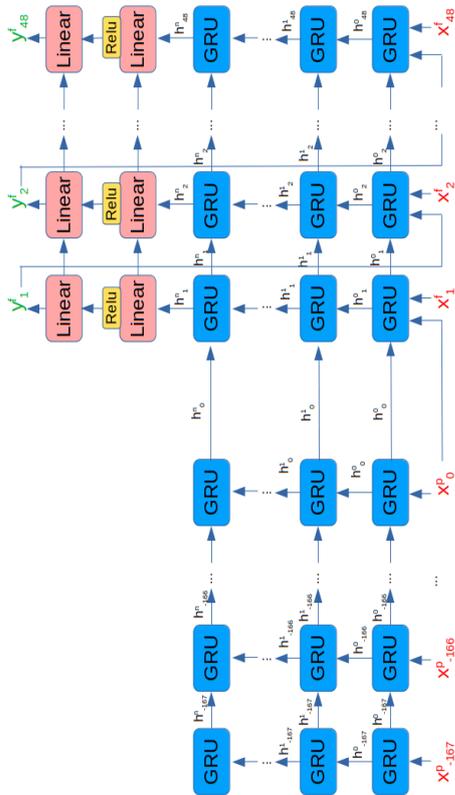
Indirect influence on the pollution load have the following conditions: combustion engines produce more nitrogen oxides when it is colder than at higher temperatures; at lower temperatures, more heating takes place, which leads to a higher PM<sub>10</sub>-concentration. Thus, temperature, precipitation, humidity, wind speed, wind direction, and sunshine duration are important input parameters for the analysis and forecast. The surroundings of the measuring stations such as the presence of roads, their traffic volume (number of lanes, traffic density), the density of buildings, the population density, the presence of vegetation in the form of forests, parks, green spaces and gardens are further influencing factors (cf. the values of the urban and rural station in figure 1).

In order to represent these factors, which are characteristic for the respective location, the course of the pollutant concentration together with the weather and time parameters of the previous week were used as input for the analysis. This also takes into account longer-term influences, such as the impact of the COVID-19 pandemic. Of course, this is only possible at the locations where measuring stations are available.

The influencing parameters are very diverse. There are interdependent and very complex effects that can hardly be represented with simple statistical methods. Newer methods that can depict such complex relationships are deep neural networks (DNNs). Therefore, a special DNN was designed, which is described in the next chapter.

### 3. PROPOSED MODEL

For this task, a sequence-to-sequence encoder-decoder model (see figure 6) has been designed and implemented, following roughly the approach described by Tang et al. (2016). Encoder and decoder are both  $n$ -layer GRUs (Gated Recurrent Units), and the decoder has in addition two linear output layers with a rectified linear unit (ReLU) in between. The network produces a so-called hidden state after each time step which is used as input for the computation of the next time step. Thus, information can flow from earlier to later time steps, enabling the model to learn dependencies and recurring patterns in the dataset. After having read all the inputs sequentially, the encoder produces  $n$  hidden states and the  $i$ th layer of the decoder takes the  $i$ th produced hidden state as an input hidden state. So in the first decoding time step, the decoder receives as input the hidden states from the encoder, the last measured pollutant values and additional time and weather input data (see next section). Then, in an autoregressive loop, the decoder takes its own output from the last time step as input for the prediction of the next time step, in addition to the last produced hidden state and additional input. In this way, the whole 48 hour output sequence is produced step by step. This approach has the advantage of being able to handle variable lengths of the input and output data, since the GRU in both encoder and decoder can be unrolled to the desired length. The split into an encoder and a decoder also provides the model with a natural modularity. Encoder and decoder can have completely different inputs and can be modified independently from one another. This is particularly relevant for our project, since the model has to deal with many heterogeneous data sources that have to be integrated into the model in multiple development cycles.



**Figure 6:** Architecture of the proposed sequence to sequence model.  $x^p$  is the input over the past 168 hours and  $x^f$  the input for the future 48 hours.  $h^k_i$  is the hidden state of the  $k$ -th layer at time step  $i$ .

## 4. EXPERIMENTS

### 4.1 Dataset Description

Table 1 gives an overview over the input data used to train the model. Historical pollutant load data from 2010 until 2020 have been acquired from 52 measuring stations across North-Rhine-Westphalia (NRW) and four measuring stations in Stuttgart. All the stations in NRW collect data of nitrogen monoxide (NO), nitrogen dioxide (NO<sub>2</sub>) and particulate matter (PM<sub>10</sub>), 20 stations collect ozone data (O<sub>3</sub>) in addition. The Stuttgart stations only have consistent measurements of NO<sub>2</sub>. For the weather data, data has been collected from stations of German Weather Service (DWD) across NRW and around Stuttgart. The values used for our dataset are temperature, precipitation, sunshine duration, wind speed, wind direction and amount of rain. Since the weather stations are in different locations than the pollutant measurement stations, the weather data at the pollutant stations had to be interpolated. For this, an inverse distance weighting has been applied to the values from weather stations in a 25 km radius around the pollutant load stations. All data is collected in hourly resolution, missing data has been linearly interpolated for up to 3 hour wide gaps.

As described in the previous section, the statistical analysis of the dataset suggests that recurring patterns over a year, month, week or day are characteristic. To provide information about this to the network, cyclical time features were created by encoding the month, day, weekday, hour and wind direction as sine and cosine values, so that the network can recognize, for

example, the hour 23 being close to hour 0. The cited weekend effect also applies to work free days in general, so a flag value was created stating if the considered day is a holiday or not.

Input Parameter	Source	Unit
NO	LANUV	µg/m <sup>3</sup>
NO <sub>2</sub>	LANUV/LUBW	µg/m <sup>3</sup>
O <sub>3</sub>	LANUV (20 stations)	µg/m <sup>3</sup>
PM <sub>10</sub>	LANUV	µg/m <sup>3</sup>
Air temperature	DWD	°C
Precipitation	DWD	mm
Sunshine duration	DWD	hour
Humidity	DWD	%
Wind direction	DWD	sine, cosine
Wind speed	DWD	m/sec
Year	Preprocessing	integer
Month	Preprocessing	sine, cosine
Weekday	Preprocessing	sine, cosine
Hour	Preprocessing	sine, cosine
Work free day	Preprocessing	Boolean

**Table 1.** Input parameters for training

For the training dataset, a mean normalization was applied for each feature. Overlapping sequences were created with 168 hours (one week) of past input data for the encoder with pollutant load, time and weather data and 48 hours of subsequent weather data and time input data for the decoder plus 48 hours of corresponding pollutant load labels. For splitting these data into training, validation and test data set, there was a trade-off between the amount of randomization and hence least possible bias in the split, and the amount of training data lost due to the splitting, because the 216 step long sequences have to be taken from continuous time regions and cut off at the borders between regions going to different datasets. We decided to split the data according to months, choosing for each station for each year one month at random going to validation and one month going to test data set, and the rest going to the training data set. Thereby the training, validation and test data set ratio is about 10/12 to 1/12 to 1/12. For all stations of NRW this added up to 2,299,349 samples in training, 169,523 in validation and 168,183 in test sets. These sequences have been shuffled and batched before each training epoch.

### 4.2 Training

Three separate models have been trained: One on data from all the NRW-stations producing a forecast of NO, NO<sub>2</sub> and PM<sub>10</sub>, the second on data from only the NRW stations where O<sub>3</sub> measurements were available, producing a forecast of O<sub>3</sub> in addition to the other values and the third model for the four stations in Stuttgart, producing a forecast of NO<sub>2</sub>.

Due to the unknown complexity of dependencies between the input parameters, many different combinations of hyper-parameters were explored, as shown in table 2. The network was trained with L2 loss function and the AdamW-optimization algorithm. It has 20,070,404 trainable parameters in total. Training the network on the complete dataset, which amounts to 77 GB of data, using a NVIDIA Tesla V100 GPU with 16 GB RAM, took about 35 hours.

Parameter	Value Range
Hidden State Size	64 – 1,280
RNN Cell Type	LSTM, GRU
RNN layers	1-6
Dropout rate	0.1-0.7
Learning rate	0.0001-0.01

**Table 2.** Explored parameters of the network and the training

### 4.3 Technical Details

The server for the analysis is running on Linux Ubuntu 18.04.1. All of the programming for the project was done with Python 3.8. The data are stored in a PostgreSQL database and accessed via the Python library SQLAlchemy. Data preparation and pre-processing was done using mainly Numpy and Pandas. The neural network model, training and prediction was implemented with PyTorch 1.4.

### 4.4 Results

The most important evaluation metric to compare the models is the mean absolute prediction error (L1) on the test data set. Additionally, the produced curves were evaluated visually see how good they capture the variability of the data. The following combination of hyper parameters yielded optimal results according to these criteria (see table 3):

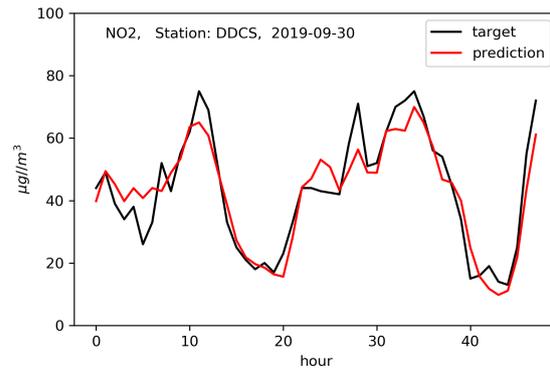
Parameter	Value
Hidden units	1024
RNN cell type	GRU
RNN layers	2
Dropout rate	0.3
Learning rate	0.0005

**Table 3.** Optimal combination of network and training parameters

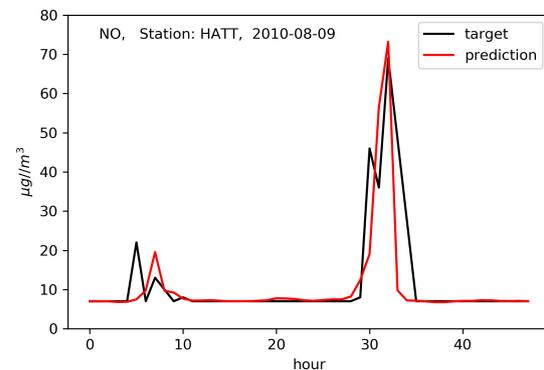
LSTM and GRU showed similar results, but LSTM has more parameters and is hence slower to train, so GRU was chosen. Increasing the number of hidden units boosted performance up to the value of 1024, where it reached a plateau. Increasing the number of layers above two, by contrast, did not change performance significantly. Increasing dropout above 0.3 resulted in a better proportion of train and validation loss, but decreased validation loss in total. Since dropout can be seen as making the network effectively smaller, and the total number of network parameters grows very fast (quadratically) with the number of hidden units, it is likely that the mere size of the network has a strong effect on performance. This is what we can expect for a dataset that is as large and diverse as ours. The network yielded an average L1 loss of 3.9 for a 48 hour prediction, distributed over the pollutants as shown in Table 4. Figures 7 to 10 show examples of predictions on the historical test data set in contrast to the measured values (named as “target”; scales vary between figures). In general, the network has the tendency to smoothen the predicted results. Positive and negative peaks in the curves are mostly detected by the network, whereas fluctuations with lower amplitude are smoothened.

Pollutant	L1 Test Loss ( $\mu\text{g}/\text{m}^3$ )
NO	2.96
NO <sub>2</sub>	4.26
O <sub>3</sub>	6.49
PM <sub>10</sub>	1.9
total	3.9

**Table 4.** L1 test loss for the pollutants

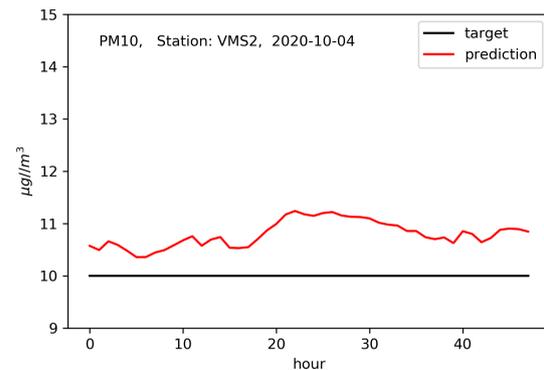


**Figure 7:** NO<sub>2</sub> prediction, station DDCS (L1-Loss: 5.22)

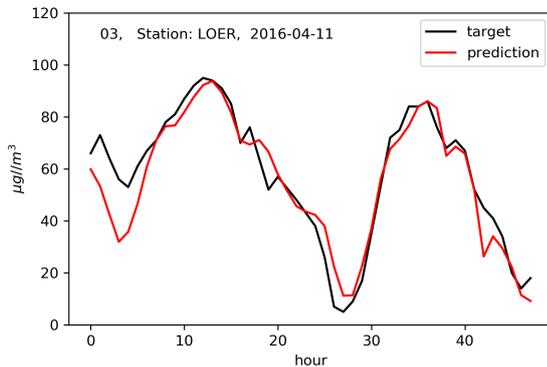


**Figure 8:** NO prediction, station HATT (L1-Loss: 3.12)

Figure 9 shows an example of a PM<sub>10</sub> prediction where the measured values were below the minimum detection threshold of 10  $\mu\text{g}/\text{m}^3$  PM<sub>10</sub>. Here, the network predicts a smooth curve but stays close to the constant minimum.



**Figure 9:** prediction of PM<sub>10</sub> near the minimum detection (L1-Loss: 0.79)



**Figure 10.** O<sub>3</sub> prediction, station LOER (L1-Loss: 6.04)

The initial results of the trained model show a consensus with the studies discussed in chapter 2. By adding the weather data to the model, the influence of the individual datasets becomes apparent and the prediction is getting more accurate.

To assess the influence of the weather conditions on the prediction quality, the model was run without weather data. Table 4 shows the L1 test loss without weather input. The average prediction accuracy decreases by 54%.

Pollutant	L1 Test Loss ( $\mu\text{g}/\text{m}^3$ )
NO	4.01
NO <sub>2</sub>	6.29
O <sub>3</sub>	10.89
PM <sub>10</sub>	2.99
total	6.04

**Table 1:** L1 test loss without weather input

## 5. CONCLUSION AND FUTURE RESEARCH

In this study we have successfully developed and trained a basic sequence-to-sequence model to forecast air pollutant concentrations at the location of ambient air monitoring sites. An initial evaluation with measurements (target) shows a fairly good agreement also considering the mean absolute prediction error.

In a next step, the model shall be further improved by first making use of additional data sources, namely topographic information, satellite-based observations and traffic data. Second, further architectural refinements shall be explored, such as adding residual connections and applying an attention mechanism. Third, the model shall be adapted to perform a forecast on arbitrary locations across the considered regions. For this, the network shall be retrained on the station locations, but this time the pollutant concentrations will be removed from the encoder input and additional data from other sources, namely topography, land use, weather, traffic and satellite observations of trace gases will be added. The measurement stations will provide real pollutant load values as labels for the training. Since the topographical, satellite, traffic and weather data are available all over the related regions, the network can then perform a forecast anywhere in those regions. The co-design with local experts and administrative authorities ensures a tailored system according to their needs and fosters the

applicability and acceptance of the system. Further research comprises the integration of additional data sources, the integration of the forecast capabilities in local applications, as well as the use of the trained model to simulate urban air quality under different scenario conditions.

## ACKNOWLEDGEMENTS

The research is supported by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) under the frame of mFUND, a research initiative funding R & D projects related to digital data-driven applications for Mobility 4.0 (grant number 19F2064F). We thank the four anonymous reviewers for their valuable comments.

## REFERENCES

- Arndt, W.-H., 2012: Verkehrsplanung und Gesundheit – Stadtverkehr und seine gesundheitlichen Folgen. In: Böhme et al.: *Handbuch Stadtplanung und Gesundheit*. Hans Huber Verlag, Bern. 85-102.
- Behera, S., Balasubramanian, R., 2016: The Air Quality Influences of Vehicular Traffic Emissions. <http://dx.doi.org/10.5772/6469>.
- Bundes-Immissionsschutzgesetz (BImSchG), Federal Emission Control Act from May 17<sup>th</sup>, 2013, last update April 8<sup>th</sup>, 2019.
- Brunekreef, B., Holgate, S.T., 2002: Air pollution and health. *The Lancet*. Vol. 360, 1233-1242.
- Cho, K., Merriënboer, B. van, Gülçehre, C., Bougares, F., Schwenk, H., Bengio, Y., 2014: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. <http://arxiv.org/abs/1406.1078>.
- COP Copernicus Atmosphere Monitoring Service, 2021. <http://www.regional.atmosphere.copernicus.eu> (31 March, 2021).
- Dora, C., Hosking, J., Mudu, P., Fletcher, E., 2011: Urban Transport and Health. Sustainable Transport: A Sourcebook for Policy-makers in Developing Cities. Hrsg: Deutsche Gesellschaft für Internationale Zusammenarbeit. <https://www.sutp.org/publications/urban-transport-and-health/> (15 April 2021).
- Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O., 2007: Emission rate and chemical state estimation by 4-dimensional variational inversion, *Atmospheric Chemistry and Physics*, 7, 3749–3769.
- Erbertseder, T., Loyola, D., 2020: Despite weather influence -- corona effect now indisputable, DLR-EOC press release, [https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-14195/24618\\_read-64626](https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-14195/24618_read-64626) (27 April, 2020).
- European Environment Agency (EEA), 2019: Air quality in Europe – 2019 report. <https://www.eea.europa.eu/publications/air-quality-in-europe-2019> (20 April 2021).
- Hochreiter, S., Schmidhuber, J., 1997: Long Short-Term Memory. In: *Neural Computation*, 9, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.

- Khorsandi, E., Baier, F., Erbertseder, T., Bittner, M., 2018: Air quality monitoring and simulation on urban scale over Munich, *SPIE Remote Sensing*, 2018, Berlin, Germany, *Proceedings Volume 10793, Remote Sensing Technologies and Applications in Urban Environments III*; 1079303.
- Kukkonen, J., Olsson, T., Schultz, D. M., Baklanov, A., Klein, T., Miranda, A. I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K. E. J., Karatzas, K., San Jose, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., and Eben, K., 2012: A review of operational, regional-scale, chemical weather forecasting models in Europe, *Atmospheric Chemistry and Physics*, 12, 1–87.
- Lai, G., Chang, W.C., Yang, Y., Liu, H., 2018: Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. <http://arxiv.org/abs/1703.07015>.
- LANUV, 2021: Messorte der Luftqualitätsüberwachung in NRW. <https://www.lanuv.nrw.de/umwelt/luft/immissionen/messorte-und-werte> (20 April 2021).
- Lefebvre, W., van Poppel, M., Maiheu, B., Janssen, S., and Dons, E., 2013: Evaluation of the RIO-IFDM-street canyon model chain, *Atmospheric Environment*, 77, 325–337.
- Mallet, V., Quelo, D., Sportisse, B., Ahmed de Biasi, M., Debry, E., Korsakissok, I., Wu, L., Roustan, Y., Sartelet, K., Tombette, M., and Foudhil, H., 2007: Technical Note: The air quality modeling system Polyphemus, *Atmospheric Chemistry and Physics*, 7, 5479–5487.
- Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadyrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouï, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., Ung, A., 2015: A regional air quality forecasting system over Europe: the MACC-II. *Geoscientific Model Development*. 8 p. 2777-2813.2015.
- Platz-Dülmer, C., 2020: Niedrige Konzentrationen in der Atmosphäre – Ein Corona-Pandemie Effekt? Hrsg.: Deutscher Wetterdienst (DWD), GAW-Brief Nr.75. [https://www.dwd.de/DE/forschung/atmosphaerenbeob/zusammensetzung\\_atmosphaere/hohenpeissenberg/download/gaw\\_briefe/gaw\\_brief\\_075\\_de.pdf.pdf?\\_\\_blob=publicationFile&v=4](https://www.dwd.de/DE/forschung/atmosphaerenbeob/zusammensetzung_atmosphaere/hohenpeissenberg/download/gaw_briefe/gaw_brief_075_de.pdf.pdf?__blob=publicationFile&v=4) (20 April 2021).
- Petry, L.; Herold, H.; Meinel, G.; Meiers, T.; Müller, Inken; Kalusche, E.; Erbertseder, T.; Taubenböck, H.; Zaunseder, E.; Srinivasan, V.; Osman, A.; Weber, B.; Jäger, S.; Mayer, C.; Gengenbach, C., 2020: Air quality monitoring and data management in Germany – Status quo and suggestions for improvement. In: ISPRS TC IV - 5th International Conference on Smart Data and Smart Cities, 2020, (*The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XLIV-4/W2-2020), 37-43. <https://doi.org/10.5194/isprs-archives-XLIV-4-W2-2020-37-2020>.
- Pixteren, D. van, Düsing, S., Wiedensohler, A., Hermann, H., 2020: Meteorologische Einflüsse auf Stickstoffdioxid. Hrsg.: Sächsisches Landesamt für Umwelt, Landwirtschaft und Geologie, Dresden. <https://publikationen.sachsen.de/bdb/artikel/35043> (20 April 2021).
- Schneider, A., Cyrus, J., Breiter, S., Kraus, U., Peters, A., Diegmann, V., Neunhäuserer L., 2018: Quantifizierung von umweltbedingten Krankheitslasten aufgrund der Stickstoffdioxid-Exposition in Deutschland. Hrsg. Umweltbundesamt. [https://www.umweltbundesamt.de/sites/default/files/medien/421/publikationen/abschlussbericht\\_no2\\_krankheitslast\\_final\\_2018\\_03\\_05.pdf](https://www.umweltbundesamt.de/sites/default/files/medien/421/publikationen/abschlussbericht_no2_krankheitslast_final_2018_03_05.pdf) (19 April 2021).
- Shih, S.Y., Sun, F.K., Lee, H.Y. 2019: Temporal Pattern Attention for Multivariate Time Series Forecasting. <http://arxiv.org/abs/1809.04206>.
- Stern, R., 2003: Entwicklung und Anwendung des chemischen Transportmodells REM/CALGRID. Abschlussbericht zum Forschungs- und Entwicklungsvorhaben 298 41 252 des Umweltbundesamts „Modellierung und Prüfung von Strategien zur Verminderung der Belastung durch Ozon“. <https://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/3604.pdf> (19 April 2021).
- Stroh, K., Marb, C., 2020: Bodennahes Ozon und Sommersmog. Hrsg. Bayerisches Landesamt für Umwelt. [https://www.lfu.bayern.de/buerger/doc/uw\\_47\\_bodennahes\\_ozon.pdf](https://www.lfu.bayern.de/buerger/doc/uw_47_bodennahes_ozon.pdf) (20 April 2021).
- Sutskever, I., Vinyals, O., Le, Q. V., 2014: Sequence to Sequence Learning with Neural Networks. <http://arxiv.org/abs/1409.3215>.
- Tang, Y., Xu, J., Matsumoto, K., Ono, C., 2016: Sequence-to-Sequence Model with Attention for Time Series Classification. <https://doi.org/10.1109/ICDMW.2016.0078>.
- Talagrand, O., 1997: Assimilation of observations, an introduction, *Journal of Meteorological Society Japan*, 75, p. 191–209.
- UBA, 2018: Wirkungen auf Ökosysteme. <https://www.umweltbundesamt.de/themen/luft/wirkungen-von-luftschadstoffen/wirkungen-auf-oekosysteme#einfuehrung> (30 March 2021).
- UBA, 2020a: Ozonbelastung. <https://www.umweltbundesamt.de/daten/luft/ozon-belastung#uberschreitung-von-schwellenwerten> (30 March 2021).
- UBA, 2020b: Feinstaub-Belastung. <https://www.umweltbundesamt.de/daten/luft/feinstaub-belastung#feinstaubkonzentrationen-in-deutschland> (30 March 2021).
- WHO, 2006: WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global update 2005. World Health Organization. <https://apps.who.int/iris/handle/10665/69477> (20 April 2021).
- WHO Regional Office for Europe, OECD, 2015: Economic cost of the health impact of air pollution in Europe: Clean air, health and wealth. Copenhagen: WHO Regional Office for Europe.