https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework

Felix Kleinert[1,2], Lukas H. Leufen[1,2], Aurelia Lupascu[3], Tim Butler[3], and Martin G. Schultz[1]

[1]Forschungszentrum Jülich GmbH, Jülich Supercomputing Centre (JSC) , Jülich, Germany
[2]Institute of Geosciences, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
[3]Institute for Advanced Sustainability Studies, Potsdam, Germany

Correspondence: Felix Kleinert (f.kleinert@fz-juelich.de)

Abstract. Tropospheric ozone is a secondary air pollutant that is harmful to living beings and crops. Predicting ozone concentrations at specific locations is thus important to initiate protection measures, i.e. emission reductions or warnings to the population. Ozone levels at specific locations result from emission and sink processes, mixing and chemical transformation along an air parcel's trajectory. Current ozone forecasting systems generally rely on computationally expensive chemistry transport models (CTMs). However, recently several studies have demonstrated the potential of deep learning for this task. While a few of these studies were trained on gridded model data, most efforts focus on forecasting time series from individual measurement locations. In this study, we present a hybrid approach which is based on time series forecasting (up to four days) but uses spatially aggregated meteorological and chemical data from upstream wind sectors to represent some aspects of the chemical history of air parcels arriving at the measurement location. To demonstrate the value of this additional information we extracted pseudo observation data for Germany from a CTM to avoid extra complications with irregularly spaced and missing data. However, our method can be extended so that it can be applied to observational time series. Using one upstream sector alone improves the forecasts by 10% during all four days while the use of three sectors improves the mean squared error (MSE) skill score by 14% during the first two days of the prediction but depends on the upstream wind direction. Our method shows its best performance in the northern half of Germany for the first two prediction days. Based on the data's seasonality and simulation period, we shed some light on our models' open challenges with i) spatial structures in terms of decreasing skill scores from the northern German plain to the mountainous south and ii) concept drifts related to an unusually cold winter season. Here we expect that the inclusion of explainable artificial intelligence methods could reveal additional insights in future versions of our model.

1 Introduction

Near surface ozone (O_3) is a secondary air pollutant which is harmful for living beings (WHO, 2013; Fleming et al., 2018) and crops (Avnery et al., 2011; Mills et al., 2018). The first Tropospheric Ozone Assessment Report (TOAR, https://igacproject.org/activities/TOAR/TOAR-I (last access: 2022-02-22)) provided the first globally consistent analysis of the global distribution

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

and trends of tropospheric ozone. Key aspects of the assessment were among others: changes in the tropospheric ozone burden and its budget (Archibald et al., 2020), observed long-term trends and their uncertainties (Tarasick et al., 2019), present day tropospheric ozone distribution and trends of metrics that are relevant to health (Fleming et al., 2018), vegetation (Mills et al., 2018) and climate (Gaudel et al., 2018). Furthermore, the capabilities of current atmospheric chemistry models were reviewed (Young et al., 2018).

Field and time series forecasting are two examples, where earth system scientists start picking up deep learning (DL) models to enhance the quality or performance of air pollution forecasts or explore novel analyses of air quality, weather and climate data. The success of these DL models is largely due to two factors: (1) improved model architectures that can capture spatiotemporal relations in the data, and (2) the increasing amount of data that has become available in recent years. While new DL methods in application areas like image or speech recognition or video frame prediction are typically developed with the help of specific benchmark datasets, thus greatly accelerating the adoption of new concepts, such benchmark datasets are only now beginning to be developed for atmospheric applications. For example, Rasp et al. (2020) developed the first meteorological benchmark data set called *WeatherBench* for medium-range weather forecasting based on ERA5 data. In terms of air quality Betancourt et al. (2021) developed a benchmark data set called *AQ-Bench* focusing on long-term ozone metrics from time-independent local features of ozone measurement sites by using the TOAR database (Schultz et al., 2017).

Concerning the problem of air quality and specifically ozone forecasts, researchers have tried out different DL models for a range of lead times in hourly (for example Eslami et al., 2019; Sayeed et al., 2020) or daily (Kleinert et al., 2021) resolution. Hybrid forecasting models combining chemical transport models with DL (Sayeed et al., 2021) and data-driven Bayesian neural network ensemble method for (numerical) geophysical models (Sengupta et al., 2020) were developed. He et al. (2022) used a deep neural network to evaluate NO_x emissions by forecasting ozone concentrations over several years and showed that their model reproduced ozone concentrations in low emission regions best when using satellite-derived NO_x trends. Sayeed et al. (2022) recently developed a DL model for postprocessing by mapping ozone precursors and meteorological information from models to observed ozone concentrations at monitoring stations. The available ozone forecasting studies employed different selections of input variables and different methods to preprocess the input data in order to help the DL methods extract the most relevant information. Often, environmental scientists use their knowledge of atmospheric processes to select variables or design the preprocessing strategies. As one example out of many a decomposition of input time-series can help neural networks to learn features like seasonality much easier - especially when the amount of training data is limited (Leufen et al., 2021b, in review).

The present study focuses on the extraction of spatiotemporal features in the context of time-series predictions. We re-use the DL set-up of Kleinert et al. (2021) who developed a time-series prediction model based on an inception architecture (Szegedy et al., 2015; Ismail Fawaz et al., 2020). That model was trained on daily aggregated ozone data from more than 300 German air quality measurement stations for a lead time of up to four days. In contrast to this earlier study the present work uses a U-shaped architecture and is based on simulation output from the chemical transport model (CTM) WRF-Chem (Grell et al., 2005) instead of the TOAR observations to demonstrate the added value of upstream wind sector information without having

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions
Open Access
EGU

to cope with irregularly spaced and sometimes missing observations. Using WRF-Chem data also as target data $\mathbf{y}$ avoids representation problems when comparing gridded model data and point measurements.

Through the adoption of an aggregated upstream wind sector approach following Yi et al. (2018) we are aiming to capture the *chemical history* of air masses, i.e. the fact that air pollutant concentrations at a given observation site are a result of emission and sink processes, mixing and chemical transformations along the transport pathways of air. Yi et al. (2018) defined multiple wind sectors around a measurement station and used the spatially aggregated information as additional input for their time-series model. In our study we condense this method by using one or three upstream sectors and we identify the influence of those sectors on the air quality forecast accuracy.

In reality, the chemical history of air parcels must be expressed as a multi-dimensional integral over a wide spectrum of chemical ages (i.e. the product of loss rates and time) and including different mixing rates. Lagrangian particle models such as FLEXPART (Pisso et al., 2019) have been used to disentangle these processes and allow for attribution of air pollutant concentrations to specific source regions (cf. Stohl et al., 1998, 2013; Wenig et al., 2003; Yu et al., 2020; Aliaga et al., 2021). However, such simulations are not straightforward to run and would have added a lot of complexity to this study. The much simpler aggregation of input variables in one or three upstream wind sectors should at least capture the first-order effects of the air parcel history and thus add valuable information to the input data for our DL network.

This article is structured as follows: In Sect. 2 we briefly introduce the WRF-Chem model data. In Sect. 3 we introduce our preprocessing methods (Sect. 3.1) and the MLAir framework (Sect. 3.2) which we use for our study. In Sect.3.3 we present the DL model architecture and training procedure, followed by a description of our reference models (Sect. 3.4). We present the our results in Sect. 4 and discuss them in Sect. 5 with a special focus on the DL method's loss (Sect. 5.1), the benefits of using additional upstream information (Sect. 5.2), the sensitivity of our DL model with respect to input variables (Sect. 5.3) and the implications of an unusually cold winter season for the evaluation of our DL model (Sect. 5.4). Finally, Sect. 6 provides conclusions.

2 Data: WRF-Chem

We use numerical model data from WRF-Chem (Grell et al., 2005), version 3.9, as the foundation of our study. We carried out a simulation for 2009 and spring 2010 over Europe because a model set-up existed for this period (Galmarini et al., 2021). The simulation domain covers 400 grid points in west-east direction, 360 grid points in south-north direction across 35 vertical levels from the surface to 50hPa. This corresponds to a horizontal grid spacing of $\sim$ 12km. Initial and boundary conditions for the meteorological fields are taken from the ERA-Interim reanalyses (Dee et al., 2011), and the chemical initial and boundary conditions were derived from the CAMS reanalysis (Inness et al., 2019). To force the model toward the spatial and temporal analyses, the model was run with grid-nudging. Single simulations were performed for 3.5 months periods, leaving out the first 15 days. Thus we ensure that the model does not deviate from the observed synoptic events and the impact of meteorological errors on atmospheric chemistry simulations is reduced. The anthropogenic emissions data are taken from the TNO-MACC III emission inventory (Kuenen et al., 2014), and the MEGAN model (Guenther et al., 2006) was employed to

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

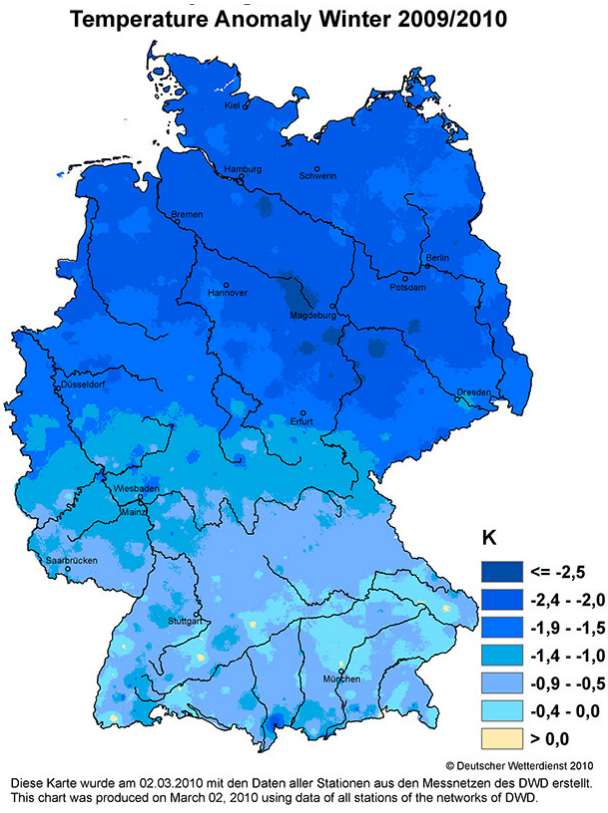

Figure 1. Temperature anomaly in the winter 2009/10 with respect to the multi year average from 1961 to 1990. Figure created by Deutscher Wetterdienst (German weather service, DWD) and downloaded from https://www.dwd.de/EN/ourservices/klimakartendeutschland/klimakartendeutschland.html?nn=495490 (last access: 2022-04-01).

estimate biogenic species emissions. As in Kuik et al. (2016), land cover classes have been updated with the CORINE dataset (CLC, 2012, Copernicus Land Monitoring Service). The emissions from open burning are based on the Fire INventory from NCAR (FINNv1.5) (Wiedinmyer et al., 2011). The main physics options are described in Lupaşcu and Butler (2019). The gas-phase mechanism is MOZART (Emmons et al., 2010; Knote et al., 2014) coupled with the MOSAIC aerosol module (Zaveri et al., 2008).

To train and evaluate the deep learning model, we extracted WRF-Chem data at the locations of the air quality measurement stations of the German Umweltbundesamt as in Kleinert et al. (2021) (see also Sect. 3.1).

The winter 2009/10 in Europe was special in the sense that it was an unusually cold winter. The average winter temperature in Germany was $-1.3°\mathrm{C}$ which is $1.5°\mathrm{C}$ below the average winter temperatures from 1961 to 1990 (DWD, 2022). Figure 1 shows that the anomalies in Northern Germany were stronger than those in Southern Germany. We will discuss the implications of this anomaly for the evaluation of our DL model in Sect. 5.4.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

3 Method

Data driven machine learning applications, require substantial effort to select and preprocess the data to be used. In the case of environmental data researchers have to find a compromise between available data and the independence of data used for training, validation and testing mostly due to limitations in data availability (Schultz et al., 2021). To incorporate the information from one or three upstream wind sectors, the preprocessing of data had to be expanded compared to Kleinert et al. (2021) and is presented in Sect. 3.1. Moreover, we briefly introduce the *Machine Learning on Air data* framework (Leufen et al., 2021a, MLAir) which we have used to carry out the experiments (Sect. 3.2), and the model architecture and training procedure (Sect. 3.3).

3.1 Data Preprocessing

The main focus of this study is to assess the added value of incorporating upstream information in the sense of a chemical air parcel history on the characteristics of air quality, i.e. ozone, forecasts with a deep neural network. For this, we define three sets of experiments; firstly, a baseline (NNb) experiment with no upstream information, secondly, a single sector (NN1s), and thirdly a multi-sector (NN3s) approach. These sets of experiments necessitate different preprocessing steps, which we introduce in the following (see also Table 3 for a summary of acronyms). As the ultimate goal of our experiments and methods is to apply them to air quality observations, we selected the locations of German air quality stations as our input and target data and applied nearest-neighbour sampling to extract the specific WRF-chem model grid box 'representing' the location of a given measurement site. From here on, we will refer to those grid boxes as pseudo-observations and pseudo-measurement stations.

All in all, we use data from 332 pseudo-measurement stations for training, validating and testing of the neural network by following Kleinert et al. (2021). As Schultz et al. (2021) and Kleinert et al. (2021) propose, we use a consecutive data split. We train the model with data ranging from January 1st 2009 to October 15th 2009 and use a short validation period ranging from October 16th 2009 to December 14th 2009. The final test period, which we used for all results presented below, covers December 16th 2009 to March 31st 2010. This split is motivated by the fact that ozone concentrations during winter are primarily determined by transport processes as opposed to summertime, when photochemistry plays a more active role. Hence, the effects of incorporating upstream information should become more apparent if we evaluate (i.e. 'test') our model for winter months (DJF(M)). However, the model is trained with data from all seasons and thus needs to generalise sufficiently well to capture the strong seasonal cycle of ozone concentrations. An overview of our data split is shown in Fig. 2 and Tab. 1.

We use the following meteorological variables as model input ($\mathbf{X}$): 2m-temperature (T2), 2m-water vapour mixing ratio (Q2), surface pressure (PSFC), planetary boundary layer height (PBLH), 10m-horizontal wind speed (wspd10ll) and direction (wdir10ll). We convert the horizontal wind components from the Lambert conformal conic projection of WRF-Chem to the geographic coordinate system. We further convert the winds u- and v-component to wind-speed and wind direction to determine the upstream wind direction. Moreover, we use the following chemical variables from the lowest model level: ozone (O_3), nitrogen oxide (NO), nitrogen dioxide, (NO_2) and carbon monoxide (CO). We aggregate the hourly model values to daily

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

Table 1. Number of pseudo-stations and number of samples for the training (train), validation (val) and testing (test) data sets.

	no. ps.-stations	×	no. days	=	no. samples
train	332		278		92296
val	332		50		16600
test	332		97		32204

Table 2. Input and target variables with applied daily aggregation and scaling method

	variable	daily metric	scaling method
input	NO	dma8eu	z-standardised
	NO_2	dma8eu	z-standardised
	O_3	dma8eu	z-standardised
	CO	dma8eu	z-standardised
	2m temperature (T2)	mean	z-standardised
	2m water vapour (Q2)	mean	z-standardised
	10m wind direction (wdir10ll)	mean	min-max scaled
	10m wind speed (wspd10ll)	mean	z-standardised
	surface pressure (PSFC)	mean	z-standardised
	planetary boundary layer height (PBLH)	mean	z-standardised
target	O_3	dma8eu	z-standardised

means (meteorological variables) and maximum daily 8-hour means (dma8eu) (chemical variables[1], see also Tab. 2) according to the EU definitions (European Parliament, 2008) by using the toar-stats python-package (Selke et al., 2021). Afterwards, we z-standardise all data from the training set except the wind direction to mean zero with unit variance. The wind direction is min-max scaled. Subsequently, we apply the scaling parameters obtained from the training set to the validation and testing data sets. This approach allows for the most rigorous evaluation of the model's generalisation capability, because no implicit information of the validation and testing sets contaminates the training set. Table 2 summarises all variables, their aggregated statistics and the applied scaling method.

For the baseline experiment (NNb), we take the daily meteorological and chemical variables (see also Sect. 2) at a pseudo-measurement station of the previous $N = 6$ time steps (t_{-6} to t_0) to create the input tensor ($\mathbf{X}$) for our neural network. Accordingly, we use the next $M = 4$ timesteps (t_1 to t_4) at the same pseudo-measurement station to create the labels ($\boldsymbol{y}$). We repeat this procedure for all $K = 332$ pseudo-measurement stations.

[1]We decided to use dma8eu for all chemical variables to sample all chemical quantities during the same time periods. dma8eu is calculated based on data starting at 17:00 local time (LT) on the previous day, while the mean is calculated based on data starting at 00:00 LT at the current day.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions
Open Access
EGU

For NN1s and NN3s we use additional chemical and meteorological information of the surrounding area as inputs to the neural network. We divide the wind directions into eight sectors corresponding to i) north, ii) north-east, iii) east, iv) south-east, v) south, vi) south-west, vii) west, and viii) north-west. Consequently, each section covers $45°$. For each of our pseudo-observations, we identify the upstream wind sector at time t_0. We then select all WRF grid boxes (center points) which fall into this sector and have a maximal distance of 200km to the central point of interest. Next, we compute the spatial mean for all chemical and meteorological variables in this set of grid boxes and use this aggregated information as additional input on top of the local pseudo-observational input that is used in NNb. For NN3s experiments, the same processing is applied to data in the sectors to the left and right of the upstream wind sector. Algorithm 1 depicts the preprocessing strategy.

Figure 3 shows the distribution of upstream sectors, i.e. number of samples, within the test set. The south-western sector dominates ($\sim$ 8200 samples) followed by the western sector ($\sim$ 4800 samples). The northern and north-western sectors contain less that 2000 samples each.

Algorithm 1 Data preprocessing: NNb, NN1s, NN3s

1: Apply toar_stats to all variables
2: Convert horizontal wind components from Lambert conformal to geographic coordinate system
3: Rescale time series with parameters from training set
4: **for all** Pseudo-stations: Create samples **do**
5: Identify grid box representing pseudo station
6: **if** NN1s **or** NN3s **then**
7: detect upstream wind sector at t_0
8: average across all grid boxes within the sector
9: **if** NN3s **then**
10: detect 'left' and 'right' upstream wind sector at t_0
11: average across all grid boxes within each sector
12: **end if**
13: **end if**
14: Create inputs $\mathbf{X}$ with variables of seven days $(-6d)$ to $(0d)$).
Shape of $\mathbf{X}$: $7 \times 1 \times (10 + 10s)$: number of days, 1, number of pseudo-measurement variables + aggregated sector variables
#s(NNb) = 0, s(NN1s) = 1, s(NN3s) = 3
15: Create labels $\boldsymbol{y}$ with ozone (dma8eu) concentrations for the next four days (1d to 4d).
Shape of $\boldsymbol{y}$: 1×4 (1, lead time)
16: **end for**
17: Permute samples in the training set
18: Create batches of size 256

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

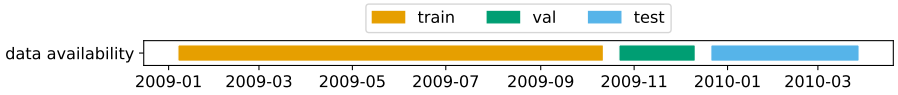

Figure 2. Data availability and split into training (orange), validation (green) and testing (blue) data set. Each day contains data from all 332 pseudo-stations.

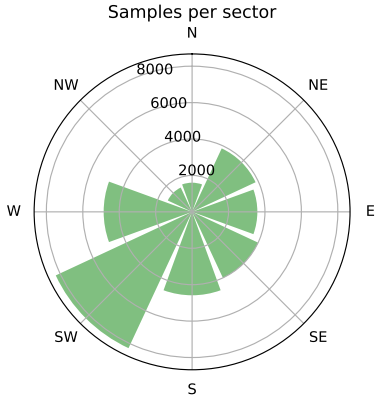

Figure 3. Number of samples per (main)-upstream wind sector at time step t_0 within the testing set.

3.2 MLAir Framework

We use the *Machine Learning on Air data* (MLAir, version 1.5) framework (Leufen et al., 2021a) as the backbone for our experiments. MLAir provides a workflow framework for machine learning based atmospheric forecasts with easily extensible modules for data preprocessing, training, hyperparameter optimisation and evaluation. MLAir uses the TensorFlow (Abadi et al., 2015), dask (Rocklin, 2015) and xarray (Hoyer and Hamman, 2017) libraries. For each of the mentioned preprocessing methods (see Sect. 3.1), we implemented individual `DataHandlers` that allow us to modify the data preprocessing steps while maintaining the same training and evaluation procedures for all experiments in spite of the different data structures.

3.3 Model architecture and training procedure

In contrast to our previous study (Kleinert et al., 2021) we use a U-shaped convolutional neural network (CNN) with two parallel long short term memory (LSTM) cells in the lowest level. Ronneberger et al. (2015) first introduced the U-Net architecture for biomedical image segmentation, but many other disciplines picked up this architecture. U-shaped or U-Net architectures have already been used to successfully tackle spatio-temporal problems (He et al., 2022). As outlined in Sect 3.1, we train three different models according to the three preprocessing variants, i) baseline (NNb), ii) one upstream sector (NN1s), iii) three upstream sectors (NN3s). The model architecture for the NN3s model is shown in Fig. 4. The first layer is the input layer. During training the individual samples in $\mathbf{X}$, representing short seven-day time series (t_{-6} to t_0) for each variable, are passed

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

to the network. We use the variables as channel dimension. We use 2D convolutional layers with a degenerated width dimension. Thus, the overall dimension of our training set is #samples $\times$ prev. N days $\times 1 \times$ #(sector) variables. We use a kernel size of 3×1 for convolutions and a pooling kernel size of 2×1. We apply exponential linear units (Clevert et al., 2016, ELU) as activation function for inner layers and a linear activation function for the output layer. Due to the different input data structures between the three experiment sets, the model architectures differ from each other in terms of the number of channels. Other design parameters - like the number of layers and kernel sizes - are identical across all runs. As the time series are relatively short, we use symmetric padding to ensure that the temporal dimension does not shrink. In the U-structure's lowest level, we use two LSTM branches after the third convolution block. While we used a max-pooling layer before the first LSTM branch to capture highly dominant features, the second LSTM branch uses the whole temporal dimension of size seven. We use an upsampling layer to expand the temporal dimension of the first LSTM branch and concatenate the resulting tensor with the symmetric padded skip connection of the third convolution block. We apply an additional convolutional layer for information compression before we append (concatenate) the second LSTM branch. Afterwards, we use altering concatenation and convolution layers to reconstruct the U's right slope. Finally, we use a simple dense layer with four nodes as the output layer. Here each node corresponds to the prediction for time step t_1 to t_4.

The original U-net proposed by Ronneberger et al. (2015) does not include padding as their input images are large enough for multiple convolutions. For each convolutional operation with a kernel size of k, the processed input data's shape shrinks by $k-1$ (assuming no padding and strides $s=1$). To prevent the inputs from becoming too small, we implement the mentioned paddings and use the standard *UpConv*-layer after the unpadded LSTM branch only. We train all models for 300 epochs using Adam (Kingma and Ba, 2014) as optimiser and the mean squared error as loss function. Detailed information on the choice of hyperparameters can be found in Tab. A1.

3.4 Reference Models

As in Kleinert et al. (2021), we use persistence and an ordinary least squares (OLS) model as references to provide a meaningful baseline evaluation of our machine learning models. The persistance forecast is built simply by using the latest observation at t_0 as forecast for all lead times t_1 to t_4. The persistence model serves as a relatively strong competitor on short lead times, and should be outperformed by all other models which add any value to the forecasting solution. To train the OLS model, we use the same data set as for the NN3s network. The OLS model serves as a linear competitor and therefore as an indicator on how well the neural networks differ from a 'simple' linear forecast. A detailed description of these two reference forecasts is provided in Leufen et al. (2021a, MLAir) and Kleinert et al. (2021, IntelliO3-ts). We also reran the experiments from this study with the IntelliO3-ts (Kleinert et al., 2021) model architecture which is based on inception layers (Szegedy et al., 2015) for comparison. As described in section 3.1 three different variants of the IntelliO3 architecture had to be built because of the different input data dimensions. All models were trained from scratch using the same data as in our main experiments.

Tab. 3 summarises the experiments described in this paper and introduces the labels that are used to denote these experiments in the results section.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development
Discussions
Open Access
EGU

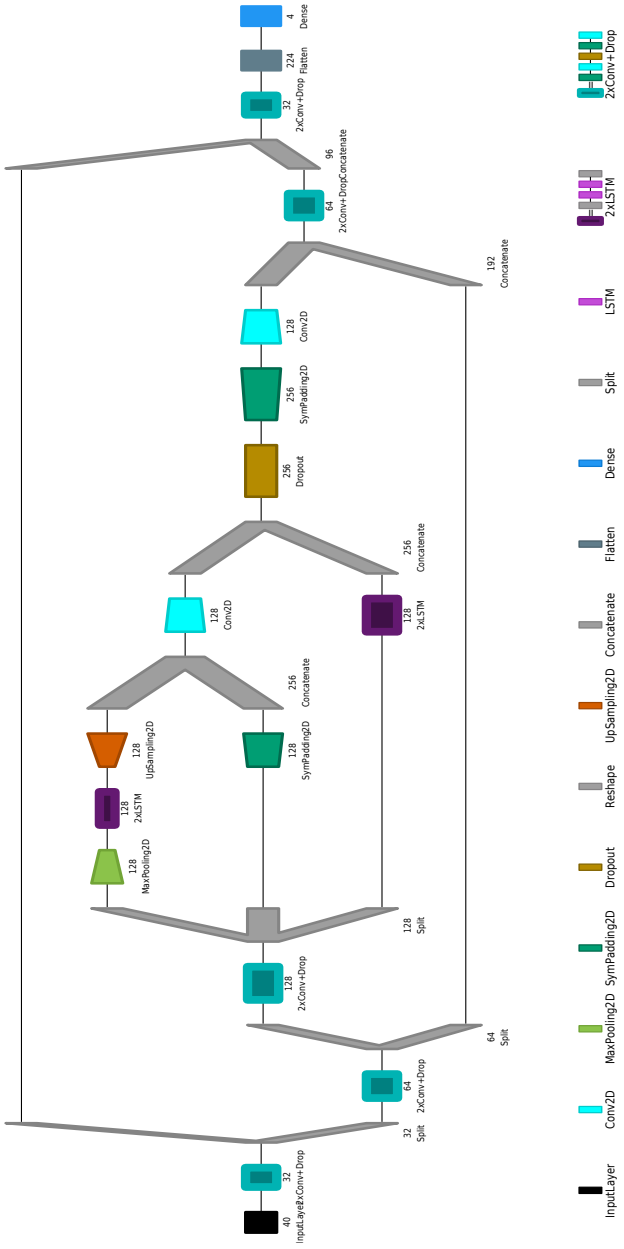

Figure 4. Summary of NN3s model architecture with input (black, bottom), hidden, and output (blue, top) layer. Several layers are grouped for better visualizations. Colour-coding of individual layers and groups is shown on the right. This figure is created with Net2Vis (Bauerle et al., 2021).

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

Table 3. Naming of deep learning experiments and reference models described in this paper

label	description
NNb	neural network baseline experiment; U-Net architecture, no upstream information
NN1s	as NNb, but with aggregated information from one upstream wind sector
NN3s	as NN1b, but with aggregated information from three upstream wind sectors (Fig. 4)
persi	persistence as reference model
ols	ordinary least square regression model as reference
IntelliO3-b	as NNb, but with IntelliO3 network architecture
IntelliO3-1s	as NN1s, but with IntelliO3 network architecture
IntelliO3-3s	as NN3s, but with IntelliO3 network architecture

4 Results

We first present the results of the two sectorial approaches (NN1s, NN3s) in comparison to the baseline method (NNb), which is not using any upstream information. These three models are compared to OLS, persistence and the IntelliO3_[b, 1s, 3s] variants. Secondly, we compare the individual losses of NN3s based on the training, validation and testing loss. Afterwards, we present more detailed results for the multi-sector approach (NN3s) including an exemplary sensitivity analysis for input variables.

Figure 5 shows the overall mean squared error (MSE) for all models and reference models across all lead times. We can clearly distinguish between different experiments ($p < .001$, see below). The U-Net architecture networks (NN3s, NN1s, NNb) show the lowest RMSE. Contrary, the multi-sector, baseline IntelliO3 and persistence experiments (IntelliO3_3s, IntelliO3_b, persi) show the largest RMSE. We estimate the uncertainty by performing a block-wise bootstrapping with a block length of seven days, and 1000 draws with replacement. Additionally, we perform a non-parametric two-sided Mann-Whitney u-test (5% significance level) for NN3s and all competitors. The NN3s approach shows the lowest RMSE and the null hypothesis of the performed u-test can be rejected ($p < .001$) for all competitors (detailed test statistics and corresponding p-values are shown in Appendix Table B1). NN3s performs better than the baseline (NNb) and the single sector (NN1s) approach, which in turn exhibits a lower RMSE compared to OLS, the InitelliO3 variants and persistence.

Figure 6 shows a monthly comparison of the forecasts distributions per lead times for the pseudo-observation. The results are shown for the test data set period ranging from Dec 2009 to Mar 2010 (see also Sect. 3.1). The forecasts show a narrower distribution when compared to the pseudo- observation's distribution. While we can observe that the NN3s model captures the changing monthly structure in terms of varying mean and median concentrations, the network is not able to adequately reproduce the variability and forecasts converge towards the monthly means. Such a behaviour was already found in Kleinert et al. (2021).

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development
Discussions
Open Access
EGU

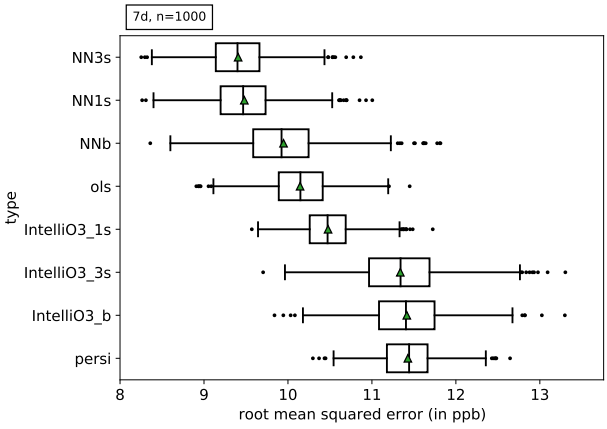

Figure 5. Estimated uncertainty of the mean squared error using block wise bootstrapping (1000 realisations with replacement with block length of seven days). IntelliO3_* correspond to the model architecture as presented in Kleinert et al. (2021, IntelliO3-ts v1.0).

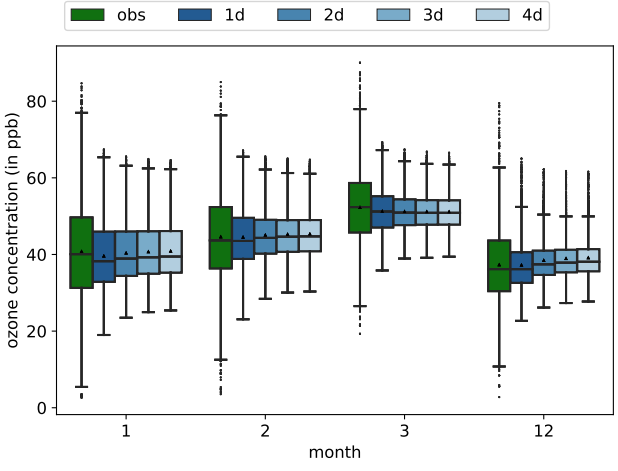

Figure 6. Monthly dma8eu ozone concentration for all test pseudo-stations as boxplots. Pseudo-observation are denoted by obs (green), forecasts are denoted by 1d (dark blue) to 4d (light blue). Triangles display the arithmetic means.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

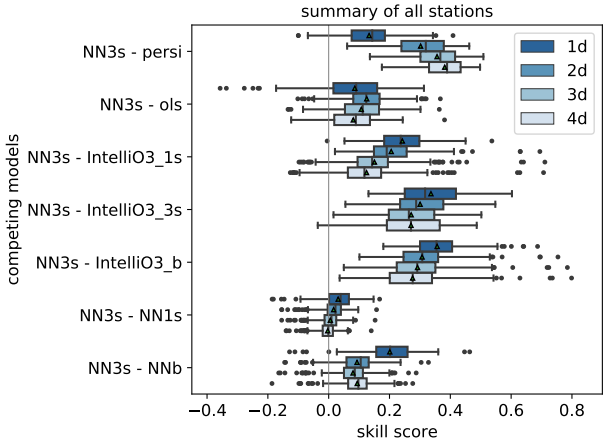

Figure 7. Skill scores of the NN3s model versus the reference models persistence (persi), ordinary least square (ols), single upstream sector model (NN1s) and pseudo-station model (NNb) based on the mean squared error; separated for all lead times (1d (dark blue) to 4d (light blue)). Positive values denote that NN3s performs better that the given references. Triangles display the arithmetic means.

More insights can be gained from evaluating the skill scores of the various experiments for the individual lead times (i.e. day 1 to day 4). As we base the comparison of our models on the mean squared error (MSE), the skill scores (S) take the form of

$$S(\boldsymbol{m},\boldsymbol{r},\boldsymbol{o}) = 1 - \frac{MSE(\boldsymbol{m},\boldsymbol{o})}{MSE(\boldsymbol{r},\boldsymbol{o})}, \tag{1}$$

where $\boldsymbol{m}$ is a vector containing the model's forecast, $\boldsymbol{o}$ is a vector containing the corresponding observations and $\boldsymbol{r}$ is a vector containing the reference forecast (Murphy, 1988). Here a positive value of $S > 0$ corresponds to an improvement of the model over the reference. Consequently, a negative value ($S < 0$) corresponds to a deterioration of skill with respect to the reference forecast.

Figure 7 shows the skill scores separated for all lead times. Here the uncertainties (boxes and whiskers) are calculated based on the individual pseudo-stations. As expected, the NN3s approach shows an increasing skill score with increasing lead time when the persistence forecast is used as reference (mean from $\sim .13$ on 1d to $\sim .38$ on 4d). When we compare NN3s with respect to the OLS model, the mean skill score is positive throughout (overall mean $\sim .10$) and has its maximum for a lead time of 2d (mean $\sim .13$), with the smallest interquartile range (IQR). For 3d and 4d lead times the skill score decreases. For the comparison of NN3s with respect to the IntelliO3 variants (_b, _1s _3s) we can observe a general pattern with the highest skill score on 1d which decreases with increasing lead time. The mean skill score of NN3s vs. IntelliO3_3s decreases from $\sim .34$ to $\sim .27$, NN3s vs. IntelliO3_1s decreases from $\sim .24$ to $\sim .12$, and NN3s vs. IntelliO3_b from $\sim .35$ to $\sim .26$, respectively. The additional two upstream sectors of IntelliO3_3s do not add any additional information compared to its single sector approach (IntelliO3_1s).

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

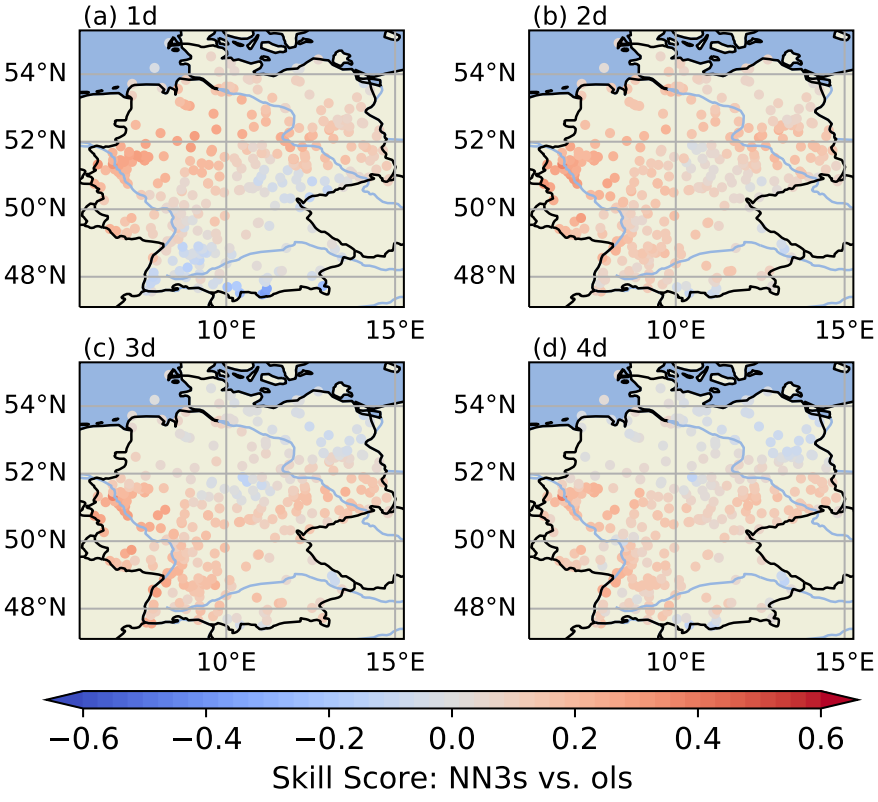

Figure 8. Skill score (NN3s vs. ols) per station for lead time 1d (a) to 4d (d). This figure is created with Cartopy (Met Office, 2010 - 2015). Map data © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Even though NN3s provides a significant performance improvement over NN1s as shown in Figure 5 the skill score of the NN3s vs. NN1s comparison is close to zero during all four forecast days (mean $\sim .03$ on 1d, $\sim .02$ on 2d, $\sim .01$ on 3d and $\sim .0$ on 4d). The added value of neighbouring upstream sectors is apparently lost after 1d.

For the comparison of NN3s with respect to NNb we see the largest skill score for a lead time of 1d (mean $\sim .2$) which decreases up do 3d ($\sim .09$, $\sim .08$) and slightly increases again on 4d ($\sim .09$).

As we originally expected a greater benefit of using upstream wind sector information for ozone forecasts, we have further analysed these skill scores and looked at their spatial distribution. From Fig. 7 we can identify some pseudo-stations for which the ols model performs better then the NN3s model. To capture potential differences in spatial properties Fig. 8 shows the skill NN3s vs. ols skill scores for each pseudo-station. The skill score (NN3s vs. ols) is mostly positive in the northern part of Germany on 1d, and is negative mostly in mountainous regions in the south of Germany. With increasing lead time the skill score becomes positive also in the mountainous regions but tend to turn negative in the north east.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

5 Discussion

Based on the results presented in Sect. 4 we can observe that NN3s outperforms the (simplistic) persistence and OLS forecasts as well as NNb, which only uses the pseudo observations of a specific grid box. The increase of the NN3s-persi skill score with increasing lead time (Fig. 7) is in line with expected behaviour as the persistence forecast has its most valuable predictions on short lead times. Consequently, the increased skill score for NN3s-persi is mostly caused by the worsening of the persistence forecast with increasing lead time. However, the positive skill score on lead-time 1d shows that the NN3s model has a genuine added value over the persistence forecast on short lead times. When comparing NN3s and NN1s, we see that the skill score's lower bound of the IQR is close to zero for the first two days of prediction and that skill score's mean and median converge towards zero for the remaining lead times; meaning that NN3s behaves like NN1s for 3d and 4d. Consequently, NN3s can not extract any additional helpful information from the input fields for 3d and 4d. Most likely this effect is caused by the definition of neighbouring sectors as described in Sect. 3.1, where we use the upstream wind sector and the two adjacent sectors with a radius of 200km at time step t_0 to calculate the spatial means for all input time steps t_{-6} to t_0. Thus, depending on the average wind speed, the static upstream wind sectors cannot provide all relevant information. Moreover, the sectorial approach is a cruder approximation of a streamline than a backwards trajectory, and differences between both of them tend to increase with increasing lead time. Based on the NN3s and NN1s sample uncertainty estimate in Fig. 5 and the competitive skill score per lead time in Fig. 7 we can conclude, that the statistical significance of the u-test (Appendix Tab. B1) is related to the better performance on the first two lead times (1d and 2d).

In the following, we discuss differences in the training, validation and testing loss of NN3s, the influence of the upstream wind direction and implications based on our data split in more detail.

5.1 Evaluation of training, validation and testing differences

We use different parts of a year to train, validate and test our models (see Sect. 3.1). Consequently, the network encounters different meteorological and chemical conditions. Therefore, we compare the differences in the MSE (which we used as the loss function during training) for the training, validation and testing set for each station individually. Averaged over all stations, the mean rescaled losses for NN3s are $64.08\,\mathrm{ppb}^2$ (train, scaled: 1.04); $63.34\,\mathrm{ppb}^2$ (val, scaled: 0.98) and $64.62\,\mathrm{ppb}^2$ (test, scaled: 1.08). Figure 9 shows how the losses are distributed geographically. In general, we can identify two main patterns. Firstly, the test set's loss is mostly lower than the validation and train loss in Germany's western and southwestern parts. In contrast, we can observe the opposite in Germany's northern and northeastern regions. The highest MSE on all sets is located in the mountainous south.

5.2 Sectorial results - NN3s

In the following we further analyse the influence of the upstream wind direction on the skill score. As mentioned in Sect. 2, the SW direction is the dominant upstream sector in the testing set, while the fewest cases are found in the NW sector.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

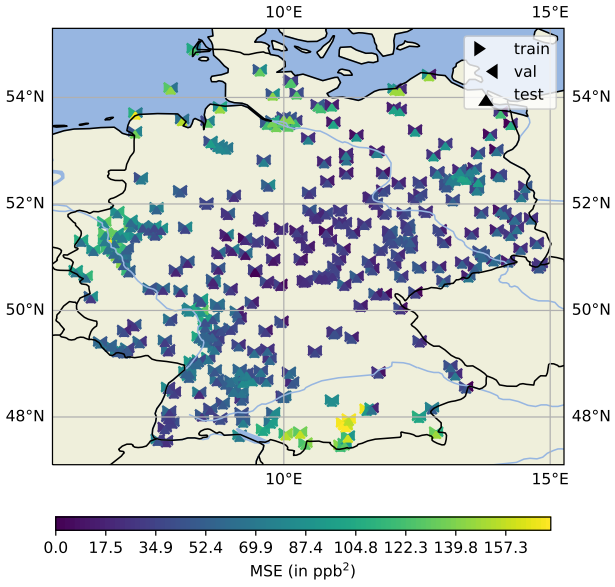

Figure 9. MSE of the training (left triangle), validation (right triangle), and testing (bottom triangle) data set for each station. This figure is created with Cartopy (Met Office, 2010 - 2015). Map data © OpenStreetMap contributors 2022. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Figure 10 shows the skill score separated for each wind sector (NN3s vs. NNb). We can observe that the skill score is mostly negative for the northern sector ($\sim -.2$), having a low number of samples (Fig. 3). On the contrary, the southwestern sector's skill score is -on average- positive, but close to zero, indicating that NNb performs equally well in the dominant upstream direction. On average the skill scores for E, SE, S and W are positive indicating that the NN3s model can extract some useful information in contrast to using the pseudo-observation at the pseudo location only. However, it becomes also obvious, that this is not the case for all stations (negative skill scores indicated by dots). It seems, that the additional information provided by the upstream sectors might sometimes confuse the network and therefore does not have the desired effect at all instances.

5.3 Non-linearities

As we saw in Fig. 8, there are pseudo stations in Germany, where a simpler OLS model produces a better forecast than NN3s. Naively, one might expect a neural network to generate better forecasts because of its ability to capture non-linear dependencies between variables. To better understand these non-linearities, we iteratively modified the input variables of each sample (in all sectors and the pseudo-observation). Afterwards, we feed the modified input samples into the trained NN3s model and detect how the ozone forecasts for all lead times change. Figure 11 shows the sensitivity of ozone inputs for one example station. For the O_3 sensitivity test, the resulting sample distribution is very narrow. We can identify three regions where the first one ranges from 0 ppb to $\sim$ 50 ppb; the second from $\sim$ 50 ppb to $\sim$ 80ppb, and the third one >80ppb. The shallower increase within

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

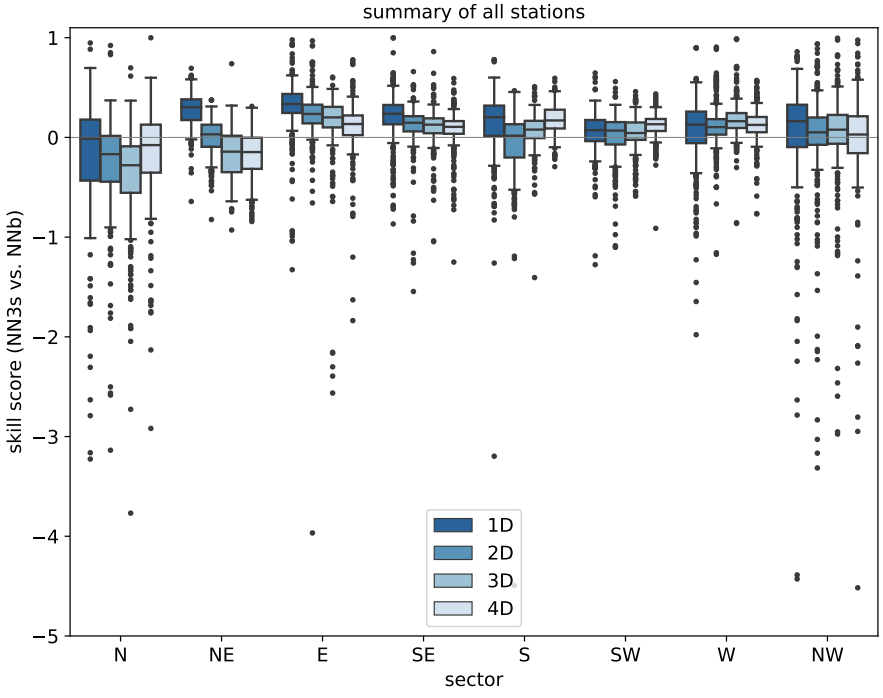

Figure 10. Skill scores (NN3s vs. NNb) based on the upstream wind direction at time step t_0.

the third region aligns with the findings from Fig. 6 where we see that NN3s cannot adequately reproduce high ozone levels. Figure 12 shows the sensitivity towards 10m wind speed. The wind speed analysis results in a broader distribution with a flat shape from 0 to 10 ms^{-1} and a linear increase for larger wind speeds.

An increase of NO levels leads to an increase of the resulting ozone level (Fig. 13). The increase is strongest for low NO levels and flattens for larger NO levels. Contrary, an increase on NO_2 leads to a decrease of the resulting ozone levels (Fig. 14) forecasted by NN3s.

5.4 Concept drift

We trained all our models and reference models with data ranging from January 1st 2009, to October 15th 2009. We selected a short validation period from mid-autumn to early winter and finally based our analysis on data from the winter and early spring seasons. Thus, as depicted in Fig. 1 the testing set contains data from an unusually cold winter which was not reflected adequately in the training set. Therefore, the network encounters some input patterns and combinations of features during testing, which are not similar to any data used for training, and thus the DL network has to generate predictions outside of its 'comfort zone' (see, for example, Leonard et al. (1992) and Pastore and Carnini (2021) or Ziyin et al. (2020) for periodic data). As presented in Sect. 2, the temperature anomaly during winter 2009/10 was largest in northern Germany and less, but still

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

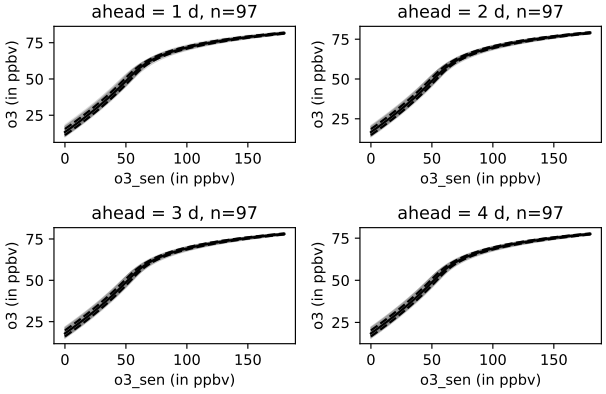

Figure 11. Sensitivity of NN3s forecasts for an exemplary station located at $50.770382°\mathrm{N}$ and $9.459403°\mathrm{E}$ with respect to ozone. Light grey lines correspond to individual test set's samples. Solid black line represents the mean, while dashed lines represent +- one standard deviation.

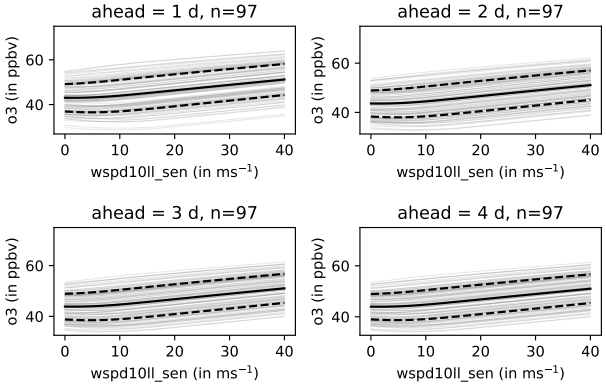

Figure 12. Same as Fig. 11, but for 10m wind speed instead of ozone input

negative, in the Southern parts of Germany. We can observe a similar north-south pattern for the skill scores (NN3s vs ols) per station (Fig. 8) and a skill scores' change of sign with increasing lead time. Nonetheless, our analysis cannot attribute these
315 phenomena and the geographic height to each other, and further analysis using explainable machine learning techniques like for example in Stadtler et al. (2022) would be required.

6 Conclusions

In this study, we explored the potential benefit of using spatially aggregated upstream information to improve point-wise predictions of near-surface ozone concentrations. Even though this analysis was based on pseudo observations sampled from

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

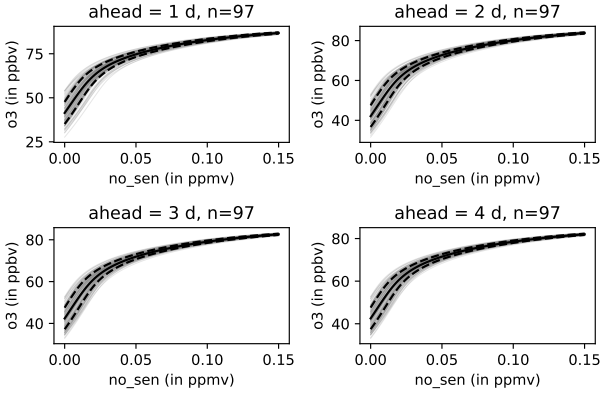

Figure 13. Same as Fig. 11, but for NO instead of ozone input

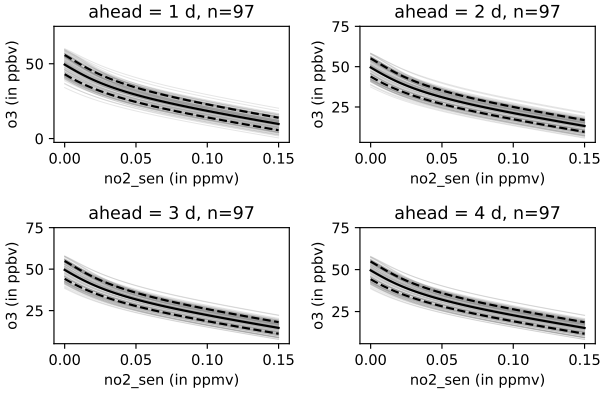

Figure 14. Same as Fig. 11, but for NO_2 instead of ozone input

chemistry transport model the results should apply to real observation data as well if they are not confounded by the inhomogeneous spatial distribution and missing data occurrences in observational data.

The first result from this study is that a U-net architecture with a combination of convolutional and LSTM cells is superior to the inception block architecture presented in Kleinert et al. (2021) in this explicit forecasting setting. The second, and main result is that the additional information provided by the central upstream wind sector (NN1s) improves the forecast for all lead times (1d to 4d) with respect to the baseline model (NNb) by 10%, which has not seen any upstream information during training. Moreover, we show that further information provided by the left and right upstream sectors (NN3s) improve the forecasts only at the first two days ($\sim$ 14%) and that there is no further improvement at the remaining days (3d and 4d) with respect to the central upstream model (NN1s). The non-linearity provided by the neural network is essential to extract meaningful upstream information as NN3s outperforms the ols model which are both trained on exactly the same data.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022


Geoscientific Model Development Discussions

Nonetheless, we showed that the NN3s model does not outperform the ols model at all pseudo-stations and lead times. This can in part be attributed to a sampling bias, because the winter in 2010 (test data) was much colder than the winter of 2009 (part of the training data). Besides these limitations, we conservatively evaluated the generalisation capability in the sense that the testing set has the largest possible 'distance' to the training data. The improvement in forecast quality arising from the upstream sector information can therefore be seen as a lower limit.

The previous study of Yi et al. (2018) showed that their deep neural network model (*DeepAir*) that makes use of spatially aggregated information outperforms several other network architectures. In complement to their study, we investigated different preprocessing variants. Our variants use various amounts of data to encapsulate the influence on the forecasting performance of one U-shaped network type.

Using Lagrangian particle modelling to derive the area of influence of a pseudo-observation (see e.g Yu et al., 2020) and sampling the history based on this area of influence instead of fixed wind sectors could be a promising direction for further studies.

Code and data availability. The current version of MLAir and its additional features are available from the project website: https://gitlab.jsc.fz-juelich.de/esde/machine-learning/mlair/-/tree/Kleinert_et_al_2022_Representing (last access: 28 February 2022) under the MIT licence (http://opensource.org/licenses/mit-license.php, last access: 28 February 2022). The exact versions of the model and data used to produce the results in this paper are archived on b2share at https://doi.org/10.34730/19c94b0b77374395b11cb54991cc497d (Kleinert et al., 2022a) and Kleinert et al. (2022b, c, d, e)

Appendix A: NN3s - Hyperparameters

Table A1 shows the hyper-parameters used to train the NN3s model. The table is automatically pooled and generated by MLAir.

Appendix B: Additional statistics

Table B1 summarises the results of two-sided Mann-Whitney u-tests (5% significance level) for NN3s and all our competitive models.

Author contributions. FK and MGS developed the concept of the study. FK implemented the preprocessing variants with contributions of LHL. FK had the lead in writing the manuscript with contributions from LHL and MGS. AL and TB conducted the wrf-chem simulations and contributed the model description and to the data section. All authors revised the final manuscript and accepted to submit to GMD.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions
Open Access
EGU

Table A1. Hyper-parameters used for NN3s

parameter	model setting
_input_shape	(7, 1, 40)
_output_shape	4
activation	elu
amsgrad	False
beta_1	0.9
beta_2	0.999
bias_regularizer	tf.python.keras.regularizers.L1L2
count params	885252
decay	0.0
epsilon	1e-07
first_filter_size	32
initial_lr	0.001
kernel_initializer	he_normal
kernel_regularizer	tf.python.keras.regularizers.L1L2
kernel_size	(3, 1)
learning_rate	0.001
loss	mean_squared_error
lstm_units	128
metrics	mse, mae
model_name	NN3s
name	Adam
optimizer	tf.python.keras.optimizer_v2.adam.Adam
pool_size	(2, 1)

Competing interests. TB is member of the editorial board of GMD. The peer-review process was guided by an independent editor, and the authors have also no other competing interests to declare.

Acknowledgements. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (http://www.gauss-centre.eu, last access: 20 April 2022) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC). We further thank Michael Langguth and Clara Betancourt for fruitful discussions and Amirpahsa Mozaffari for helping us with data curation. FK, LHL and MGS acknowledge funding from ERC-2017-ADG#787576 (IntelliAQ). AL and TB acknowledge funding from IASS Potsdam which is supported financially by the Federal Ministry of Education and Research of Germany (BMBF) and the Ministry for Science, Research and Culture of the State of Brandenburg (MWFK).

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Geoscientific Model Development Discussions

Table B1. Results of two-sided Mann-Whitney u-test (5% significance level) of NN3s and listed competitor models

type	statistics	p-value
NN1s	550094.0	0.0001
NNb	7.965250e+05	$\ll$ .0001
ols	9.024260e+05	$\ll$ .0001
IntelliO3_1s	9.790170e+05	$\ll$ .0001
IntelliO3_3s	998110.0	$\ll$ .0001
IntelliO3_b	998892.0	$\ll$ .0001
persi	999849.0	$\ll$ .0001

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/, software available from tensorflow.org, 2015.

Aliaga, D., Sinclair, V. A., Andrade, M., Artaxo, P., Carbone, S., Kadantsev, E., Laj, P., Wiedensohler, A., Krejci, R., and Bianchi, F.: Identifying source regions of air masses sampled at the tropical high-altitude site of Chacaltaya using WRF-FLEXPART and cluster analysis, 21, 16 453–16 477, https://doi.org/10.5194/acp-21-16453-2021, 2021.

Archibald, A. T., Neu, J. L., Elshorbany, Y. F., Cooper, O. R., Young, P. J., Akiyoshi, H., Cox, R. A., Coyle, M., Derwent, R. G., Deushi, M., Finco, A., Frost, G. J., Galbally, I. E., Gerosa, G., Granier, C., Griffiths, P. T., Hossaini, R., Hu, L., Jöckel, P., Josse, B., Lin, M. Y., Mertens, M., Morgenstern, O., Naja, M., Naik, V., Oltmans, S., Plummer, D. A., Revell, L. E., Saiz-Lopez, A., Saxena, P., Shin, Y. M., Shahid, I., Shallcross, D., Tilmes, S., Trickl, T., Wallington, T. J., Wang, T., Worden, H. M., and Zeng, G.: Tropospheric Ozone Assessment Report: A critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100, Elementa: Science of the Anthropocene, 8, 034, https://doi.org/10.1525/elementa.2020.034, 2020.

Avnery, S., Mauzerall, D. L., Liu, J., and Horowitz, L. W.: Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage, Atmospheric Environment, 45, 2284–2296, https://doi.org/10.1016/j.atmosenv.2010.11.045, 2011.

Bauerle, A., van Onzenoodt, C., and Ropinski, T.: Net2Vis – A Visual Grammar for Automatically Generating Publication-Tailored CNN Architecture Visualizations, 27, 2980–2991, https://doi.org/10.1109/TVCG.2021.3057483, 2021.

Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G., and Stadtler, S.: AQ-Bench: a benchmark dataset for machine learning on global air quality metrics, Earth System Science Data, 13, 3013–3033, https://doi.org/10.5194/essd-13-3013-2021, number: 6, 2021.

CLC: Copernicus Land Monitoring Service: Corine Land Cover, http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012/, 2012.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), arXiv:1511.07289 [cs], http://arxiv.org/abs/1511.07289, arXiv: 1511.07289, 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haim-

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

berger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

DWD, D.: Monthly description, https://www.dwd.de/EN/ourservices/klimakartendeutschland/klimakartendeutschland_monatsbericht.html?nn=495490#buehneTop, 2022.

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4), Geoscientific Model Development, 3, 43–67, https://doi.org/10.5194/gmd-3-43-2010, 2010.

Eslami, E., Choi, Y., Lops, Y., and Sayeed, A.: A real-time hourly ozone prediction system using deep convolutional neural network, Neural Computing and Applications, https://doi.org/10.1007/s00521-019-04282-x, 2019.

European Parliament, C. o. t. E. U.: Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, 51, 1–44, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32008L0050&qid=1637580240302, 2008.

Fleming, Z. L., Doherty, R. M., von Schneidemesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., Lefohn, A. S., Hamad, S., Moolla, R., Solberg, S., and Feng, Z.: Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health, Elementa: Science of the Anthropocene, 6, 12, https://doi.org/10.1525/elementa.273, 2018.

Galmarini, S., Makar, P., Clifton, O. E., Hogrefe, C., Bash, J. O., Bellasio, R., Bianconi, R., Bieser, J., Butler, T., Ducker, J., Flemming, J., Hodzic, A., Holmes, C. D., Kioutsioukis, I., Kranenburg, R., Lupascu, A., Perez-Camanyo, J. L., Pleim, J., Ryu, Y.-H., San Jose, R., Schwede, D., Silva, S., and Wolke, R.: Technical note: AQMEII4 Activity 1: evaluation of wet and dry deposition schemes as an integral part of regional-scale air quality models, Atmospheric Chemistry and Physics, 21, 15 663–15 697, https://doi.org/10.5194/acp-21-15663-2021, 2021.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P.-F., Cuesta, J., Cuevas, E., Doniki, S., Dufour, G., Ebojie, F., Foret, G., Garcia, O., Granados-Muñoz, M. J., Hannigan, J. W., Hase, F., Hassler, B., Huang, G., Hurtmans, D., Jaffe, D., Jones, N., Kalabokas, P., Kerridge, B., Kulawik, S., Latter, B., Leblanc, T., Le Flochmoën, E., Lin, W., Liu, J., Liu, X., Mahieu, E., McClure-Begley, A., Neu, J. L., Osman, M., Palm, M., Petetin, H., Petropavlovskikh, I., Querel, R., Rahpoe, N., Rozanov, A., Schultz, M. G., Schwab, J., Siddans, R., Smale, D., Steinbacher, M., Tanimoto, H., Tarasick, D. W., Thouret, V., Thompson, A. M., Trickl, T., Weatherhead, E., Wespes, C., Worden, H. M., Vigouroux, C., Xu, X., Zeng, G., and Ziemke, J.: Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation, Elementa: Science of the Anthropocene, 6, 39, https://doi.org/10.1525/elementa.291, 2018.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B.: Fully coupled "online" chemistry within the WRF model, 39, 6957–6975, https://doi.org/10.1016/j.atmosenv.2005.04.027, 2005.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmospheric Chemistry and Physics, 6, 3181–3210, https://doi.org/10.5194/acp-6-3181-2006, 2006.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

He, T., Jones, D. B. A., Miyazaki, K., Huang, B., Liu, Y., Jiang, Z., White, E. C., Worden, H. M., and Worden, J. R.: Deep Learning to Evaluate US NO $_x$ Emissions Using Surface Ozone Predictions, 127, https://doi.org/10.1029/2021JD035597, 2022.

Hoyer, S. and Hamman, J.: xarray: N-D labeled arrays and datasets in Python, Journal of Open Research Software, 5, https://doi.org/10.5334/jors.148, 2017.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmospheric Chemistry and Physics, 19, 3515–3556, https://doi.org/10.5194/acp-19-3515-2019, 2019.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., and Petitjean, F.: InceptionTime: Finding AlexNet for time series classification, Data Mining and Knowledge Discovery, 34, 1936–1962, https://doi.org/10.1007/s10618-020-00710-y, 2020.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs], http://arxiv.org/abs/1412.6980, arXiv: 1412.6980, 2014.

Kleinert, F., Leufen, L. H., and Schultz, M. G.: IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany, Geoscientific Model Development, 14, 1–25, https://doi.org/10.5194/gmd-14-1-2021, 2021.

Kleinert, F., Leufen, L. H., Lupaşcu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework: Experiments and source code, https://doi.org/10.34730/19c94b0b77374395b11cb54991cc497d, 2022a.

Kleinert, F., Leufen, L. H., Lupaşcu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework: Data 1/4, https://doi.org/10.34730/c799f04beb644e38a575fa20c2dd8d40, 2022b.

Kleinert, F., Leufen, L. H., Lupaşcu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework: Data 2/4, https://doi.org/10.34730/d5f34ae6a8e34d4c8ac33f75b993e8a9, 2022c.

Kleinert, F., Leufen, L. H., Lupaşcu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework: Data 3/4, https://doi.org/10.34730/a423ec9003194209989726a95a1a490c, 2022d.

Kleinert, F., Leufen, L. H., Lupaşcu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework: Data 4/4, https://doi.org/10.34730/718262bd2c894fd6aadce19a08040f69, 2022e.

Knote, C., Hodzic, A., Jimenez, J. L., Volkamer, R., Orlando, J. J., Baidar, S., Brioude, J., Fast, J., Gentner, D. R., Goldstein, A. H., Hayes, P. L., Knighton, W. B., Oetjen, H., Setyan, A., Stark, H., Thalman, R., Tyndall, G., Washenfelder, R., Waxman, E., and Zhang, Q.: Simulation of semi-explicit mechanisms of SOA formation from glyoxal in aerosol in a 3-D model, Atmospheric Chemistry and Physics, 14, 6213–6239, https://doi.org/10.5194/acp-14-6213-2014, 2014.

Kuenen, J. J. P., Visschedijk, A. J. H., Jozwicka, M., and Denier van der Gon, H. A. C.: TNO-MACC_II emission inventory; a multi-year (2003–2009) consistent high-resolution European emission inventory for air quality modelling, Atmospheric Chemistry and Physics, 14, 10 963–10 976, https://doi.org/10.5194/acp-14-10963-2014, 2014.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Kuik, F., Lauer, A., Churkina, G., Denier van der Gon, H. A. C., Fenner, D., Mar, K. A., and Butler, T. M.: Air quality modelling in the Berlin–Brandenburg region using WRF-Chem v3.7.1: sensitivity to resolution of model grid and input data, Geoscientific Model Development, 9, 4339–4363, https://doi.org/10.5194/gmd-9-4339-2016, 2016.

Leonard, J., Kramer, M., and Ungar, L.: A neural network architecture that computes its own reliability, 16, 819–835, https://doi.org/10.1016/0098-1354(92)80035-8, 1992.

Leufen, L. H., Kleinert, F., and Schultz, M. G.: MLAir (v1.0) – a tool to enable fast and flexible machine learning on air data time series, Geoscientific Model Development, 14, 1553–1574, https://doi.org/10.5194/gmd-14-1553-2021, 2021a.

Leufen, L. H., Kleinert, F., and Schultz, M. G.: Exploring decomposition of temporal patterns to facilitate learning of neural networks for near-surface dma8eu ozone prediction, Submitted to: Environmental Data Science, 2021b.

Lupaşcu, A. and Butler, T.: Source attribution of European surface O₃ using a tagged O₃ mechanism, Atmospheric Chemistry and Physics, 19, 14 535–14 558, https://doi.org/10.5194/acp-19-14535-2019, 2019.

Met Office: Cartopy: a cartographic python library with a Matplotlib interface, Exeter, Devon, https://scitools.org.uk/cartopy, 2010 - 2015.

Mills, G., Pleijel, H., Malley, C. S., Sinha, B., Cooper, O. R., Schultz, M. G., Neufeld, H. S., Simpson, D., Sharps, K., Feng, Z., Gerosa, G., Harmens, H., Kobayashi, K., Saxena, P., Paoletti, E., Sinha, V., and Xu, X.: Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation, Elementa: Science of the Anthropocene, 6, 47, https://doi.org/10.1525/elementa.302, 2018.

Murphy, A. H.: Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, Monthly Weather Review, 116, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2, 1988.

Pastore, A. and Carnini, M.: Extrapolating from neural network models: a cautionary tale, 48, 084 001, https://doi.org/10.1088/1361-6471/abf08a, publisher: IOP Publishing, 2021.

Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaaftink, C. D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J. F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEXPART version 10.4, 12, 4955–4997, https://doi.org/10.5194/gmd-12-4955-2019, number: 12, 2019.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, Journal of Advances in Modeling Earth Systems, 12, https://doi.org/10.1029/2020MS002203, 2020.

Rocklin, M.: Dask: Parallel computation with blocked algorithms and task scheduling, in: Proceedings of the 14th python in science conference, 130-136, Citeseer, 2015.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., pp. 234–241, Springer International Publishing, Cham, 2015.

Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., and Jung, J.: Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance, Neural Networks, 121, 396–408, https://doi.org/10.1016/j.neunet.2019.09.033, 2020.

Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., Lee, J.-B., Park, H.-J., and Choi, M.-H.: A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance, Scientific Reports, 11, 10 891, https://doi.org/10.1038/s41598-021-90446-6, 2021.

Sayeed, A., Eslami, E., Lops, Y., and Choi, Y.: CMAQ-CNN: A new-generation of post-processing techniques for chemical transport models using deep neural networks, 273, 118 961, https://doi.org/10.1016/j.atmosenv.2022.118961, 2022.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O., Galbally, I., Petropavlovskikh, I., Von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., Naja, M., Seguel, R., Dauert, U., Eckhardt, P., Feigenspahn, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Christian Kjeld, P., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolla, R., Wang, T., Sharps, K., Adame, J. A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M. Y., Gros, V., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Huber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L., McClure-Begley, A., Mohamad, M., Murovic, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xu, X., Xue, L., and Zhiqiang, M.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elementa: Science of the Anthropocene, 5, 58, https://doi.org/10.1525/elementa.244, 2017.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A, https://doi.org/10.1098/rsta.2020.0097, 2021.

Selke, N., Schröder, S., and Schultz, M. G.: toarstats, https://gitlab.jsc.fz-juelich.de/esde/toar-public/toarstats, 2021.

Sengupta, U., Amos, M., Hosking, J. S., Rasmussen, C. E., Juniper, M., and Young, P. J.: Ensembling geophysical models with Bayesian Neural Networks, arXiv:2010.03561 [physics, stat], http://arxiv.org/abs/2010.03561, arXiv: 2010.03561, 2020.

Stadtler, S., Betancourt, C., and Roscher, R.: Explainable Machine Learning Reveals Capabilities, Redundancy, and Limitations of a Geospatial Air Quality Benchmark Dataset, 4, 150–171, https://doi.org/10.3390/make4010008, 2022.

Stohl, A., Hittenberger, M., and Wotawa, G.: Validation of the lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data, Atmospheric Environment, 32, 4245–4264, https://doi.org/10.1016/S1352-2310(98)00184-8, 1998.

Stohl, A., Klimont, Z., Eckhardt, S., Kupiainen, K., Shevchenko, V. P., Kopeikin, V. M., and Novigatsky, A. N.: Black carbon in the Arctic: the underestimated role of gas flaring and residential combustion emissions, Atmospheric Chemistry and Physics, 13, 8833–8855, https://doi.org/10.5194/acp-13-8833-2013, 2013.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, IEEE, Boston, MA, USA, https://doi.org/10.1109/CVPR.2015.7298594, 2015.

Tarasick, D., Galbally, I. E., Cooper, O. R., Schultz, M. G., Ancellet, G., Leblanc, T., Wallington, T. J., Ziemke, J., Liu, X., Steinbacher, M., Staehelin, J., Vigouroux, C., Hannigan, J. W., García, O., Foret, G., Zanis, P., Weatherhead, E., Petropavlovskikh, I., Worden, H., Osman, M., Liu, J., Chang, K.-L., Gaudel, A., Lin, M., Granados-Muñoz, M., Thompson, A. M., Oltmans, S. J., Cuesta, J., Dufour, G., Thouret, V., Hassler, B., Trickl, T., and Neu, J. L.: Tropospheric Ozone Assessment Report: Tropospheric ozone from 1877 to 2016, observed levels, trends and uncertainties, Elementa: Science of the Anthropocene, 7, 39, https://doi.org/10.1525/elementa.376, 2019.

Wenig, M., Spichtinger, N., Stohl, A., Held, G., Beirle, S., Wagner, T., Jähne, B., and Platt, U.: Intercontinental transport of nitrogen oxide pollution plumes, Atmospheric Chemistry and Physics, 3, 387–393, https://doi.org/10.5194/acp-3-387-2003, 2003.

WHO: Health risks of air pollution in Europe – HRAPIE project. Recommendations for concentration–response functions for cost–benefit analysis of particulate matter, ozone and nitrogen dioxide, Tech. rep., WHO Regional Office for Europe, UN City, Marmorvej 51 DK-2100 Copenhagen Ø, Denmark, http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/2013/health-risks-of-air-pollution-in-europe-hrapie-project.

https://doi.org/10.5194/gmd-2022-122
Preprint. Discussion started: 11 May 2022

-recommendations-for-concentrationresponse-functions-for-costbenefit-analysis-of-particulate-matter,-ozone-and-nitrogen-dioxide, 2013.

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J., and Soja, A. J.: The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning, Geoscientific Model Development, 4, 625–641, https://doi.org/10.5194/gmd-4-625-2011, 2011.

Yi, X., Zhang, J., Wang, Z., Li, T., and Zheng, Y.: Deep Distributed Fusion Network for Air Quality Prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 965–973, ACM, https://doi.org/10.1145/3219819.3219822, 2018.

Young, P. J., Naik, V., Fiore, A. M., Gaudel, A., Guo, J., Lin, M. Y., Neu, J. L., Parrish, D. D., Rieder, H. E., Schnell, J. L., Tilmes, S., Wild, O., Zhang, L., Ziemke, J., Brandt, J., Delcloo, A., Doherty, R. M., Geels, C., Hegglin, M. I., Hu, L., Im, U., Kumar, R., Luhar, A., Murray, L., Plummer, D., Rodriguez, J., Saiz-Lopez, A., Schultz, M. G., Woodhouse, M. T., and Zeng, G.: Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends, Elementa: Science of the Anthropocene, 6, 10, https://doi.org/10.1525/elementa.265, 2018.

Yu, C., Zhao, T., Bai, Y., Zhang, L., Kong, S., Yu, X., He, J., Cui, C., Yang, J., You, Y., Ma, G., Wu, M., and Chang, J.: Heavy air pollution with a unique “non-stagnant” atmospheric boundary layer in the Yangtze River middle basin aggravated by regional transport of PM_{2.5} over China, Atmospheric Chemistry and Physics, 20, 7217–7230, https://doi.org/10.5194/acp-20-7217-2020, number: 12, 2020.

Zaveri, R. A., Easter, R. C., Fast, J. D., and Peters, L. K.: Model for Simulating Aerosol Interactions and Chemistry (MOSAIC), Journal of Geophysical Research, 113, D13 204, https://doi.org/10.1029/2007JD008782, 2008.

Ziyin, L., Hartwig, T., and Ueda, M.: Neural Networks Fail to Learn Periodic Functions and How to Fix It, in: Advances in Neural Information Processing Systems, edited by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., vol. 33, pp. 1583–1594, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2020/file/1160453108d3e537255e9f7b931f4e90-Paper.pdf, 2020.