O3ResNet: A Deep Learning–Based Forecast System to Predict Local Ground-Level Daily Maximum 8-Hour Average Ozone in Rural and Suburban Environments

LUKAS HUBERT LEUFEN,[a,b] FELIX KLEINERT,[a,b] AND MARTIN G. SCHULTZ[a]

[a] *Jülich Supercomputing Centre, Research Centre Jülich, Jülich, Germany*
[b] *Institute of Geosciences, University of Bonn, Bonn, Germany*

(Manuscript received 29 November 2022, in final form 2 March 2023, accepted 10 April 2023)

ABSTRACT: With the impact of tropospheric ozone pollution on humankind, there is a compelling need for robust air quality forecasts. Here, we introduce a novel deep learning (DL) forecasting system called O3ResNet that produces a 4-day forecast for ground-level ozone. O3ResNet is based on a convolutional neural network with residual blocks. The model has been trained on 22 yr of ozone and nitrogen oxides in situ measurements and ERA5 reanalysis data from 2000 to 2021 at 328 stations in central Europe located in rural and suburban environments. Our model outperforms the state-of-the-art Copernicus Atmosphere Monitoring Service regional forecast model ensemble for ground-level ozone with respect to the mean-square error and mean absolute error of the daily maximum 8-h running-average ozone, thus marking a major milestone for DL-based ozone prediction. O3ResNet has a very small bias without requiring additional postprocessing, and it generalizes well so that new stations can be added with no need to retrain the neural network. Because the model works on hourly data, it can be easily adapted to output other air quality metrics. We conclude that O3ResNet is sufficiently advanced and robust to become a test application for operational air quality forecasting with DL.

SIGNIFICANCE STATEMENT: In this paper, we introduce a novel deep learning approach to forecast ground-level ozone for rural and suburban environments on a local scale. Our model is able to outperform the state-of-the-art Copernicus Atmosphere Monitoring Service regional ensemble forecast and is a major milestone toward a more reliable ozone prediction. This is important because local-scale ozone forecasts using conventional methods show significant bias or require site-dependent postprocessing. The findings suggest that the model presented in this article can become an important tool for air quality prediction.

KEYWORDS: Forecasting; Neural networks; Air quality; Ozone; Deep learning; Machine learning

1. Introduction

Data-driven methods like machine learning (ML) and in particular deep learning (DL) have the potential to replace or augment classical environmental modeling approaches because they can learn complex, intrinsic relationships among observed variables and because they exhibit small bias by design (Schultz et al. 2021). Especially at small local scales, atmospheric phenomena are often not well described by existing theories, and classical model predictions are therefore imprecise. As a complex interplay of meteorology, chemistry, emissions, and landscape is involved (Solberg et al. 2016), this is especially critical for ozone air pollution.

Exposure to ozone has a damaging effect on terrestrial life forms (U.S. EPA 2013; Monks et al. 2015; Mills et al. 2018). In particular, exposure to high ozone concentrations leads to adverse health effects in humans, especially in the pulmonary and cardiovascular systems (Fleming et al. 2018). Short-term exposure to high ozone concentrations has drastic effects (WHO 2013; Bell et al. 2014; U.S. EPA 2020), such as reduced lung function or triggering of asthma. Consequently, it is important to have reliable predictions of ozone concentrations several days in advance, in order to initiate appropriate countermeasures where necessary. Regulatory authorities around the world therefore define target and limit values for ozone. These are typically based on the daily maximum 8-h running average (dma8; Fleming et al. 2018), so the analysis and prediction of dma8 ozone is a task of high societal relevance.

Current forecast models are based on chemistry transport models (CTMs) built on chemical and physical relationships and equations to calculate air quality numerically. However, in such models, uncertainties arise due to various causes, such as parameterizations, simplification of relationships and equations, or other assumptions (Manders et al. 2012; Vautard et al. 2012; Brunner et al. 2015; Bessagnet et al. 2016). These, in turn, lead to systematic deviations between the model results and the observations (Otero et al. 2018). For example, the seasonal cycle of ozone is not well represented by the CTMs, nor do they capture the sensitivity of the models to meteorological drivers relevant for ozone formation and removal processes such as solar radiation and relative humidity well (Otero et al. 2018). Also, CTMs are too coarse scaled to resolve local phenomena (Stock et al. 2014) and they impose a substantial computational burden in solving chemical equations (Wang et al. 1999), which is critical when deployed operationally, where wall-clock time is a hard constraint (Baklanov et al. 2014).

To enable DL methods to learn how to reliably predict ozone concentrations, one needs to apply domain knowledge for constructing the input data and the DL model. Temperature has an

Corresponding author: Lukas Hubert Leufen, l.leufen@fz-juelich.de

DOI: 10.1175/AIES-D-22-0085.1 e220085

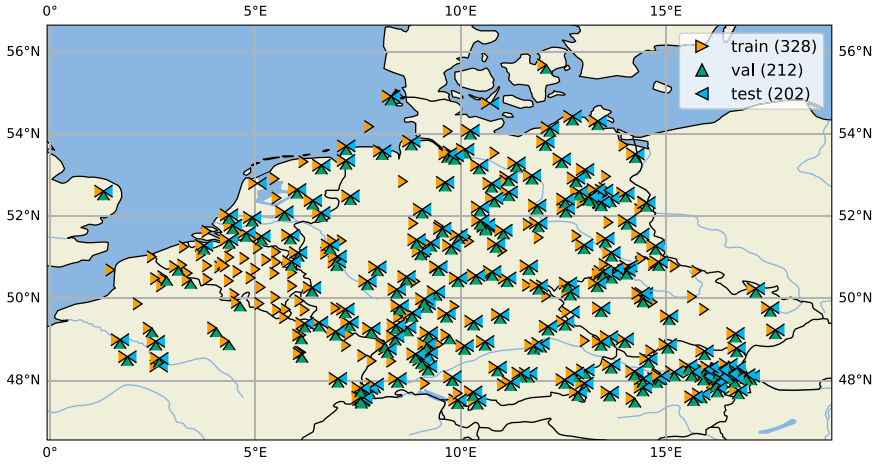

FIG. 1. Geographic overview of the ozone measurement stations in central Europe (47.5°–56°N and 1.3°–18°E). Of the 328 stations available for the 2000–21 period, all 328 stations were used for training O3ResNet (represented by orange triangles with apex oriented to the right). There are 212 stations available for validation (green triangle with apex oriented up) and 202 stations available for final testing (blue triangles with apex oriented left). The differences in the station numbers result from the data availability in the TOAR DB.

important influence on ozone, as chemical reactions are generally temperature dependent (Vautard et al. 2012). In particular, extreme ozone concentrations are mainly linked to high temperature periods (Fiore et al. 2015; Otero et al. 2016). Besides, persistence is also a strong predictor of high ozone levels, as it can indicate the presence of prolonged events and those with a day-by-day increase in concentrations (Jahn and Hertig 2021). Further meteorological factors that influence local ozone levels are solar radiation and cloud cover, as well as relative humidity and wind speed (Otero et al. 2016). Weng et al. (2022), based on random forest and ridge regression, identify, for example, temperature, surface solar radiation downward, and relative humidity as the key meteorological drivers of ozone. Their study also reveals, however, that the importance of individual variables can vary between different regions. Recent studies have shown that neural networks (NNs) are skillful methods for ozone-forecasting purposes and a variety of NN architectures have been explored in this context. For example, Seltzer et al. (2020) use fully connected networks (FCNs), Sayeed et al. (2020) use convolutional neural networks (CNNs), Kleinert et al. (2021) use CNNs with inception blocks, Ma et al. (2020) use long short-term memory networks (LSTMs), and He et al. (2022) and Kleinert et al. (2022) use U-Nets. However, to the best of our knowledge, there has been no study on forecasting of ozone at station locations that both reports good performance for lead times greater than 2 days and provides a direct comparison with a state-of-the-art CTM.

This paper presents the development of a generic DL-based ozone forecasting system called O3ResNet, that is based on a CNN architecture with residual blocks (He et al. 2016), to forecast ground-level dma8 ozone at individual stations. To showcase O3ResNet, we selected 328 stations in rural and suburban areas across central Europe for study, although the system can easily be adapted to other regions, provided that enough training data are available. Results of O3ResNet are more accurate than the Copernicus Atmosphere Monitoring Service (CAMS) regional ensemble forecast (CAMS 2020), which is the state-of-the-art air quality forecast system in Europe. Therefore, O3ResNet provides a reliable dma8 ozone forecast for the next 4 days, denoted D1–D4, which makes it a tool that is suitable for operational air quality forecasting.

This paper begins with a description of the data and methods used, followed by the results section, in which we draw a comparison with CAMS in addition to evaluating the overall performance of our model. Here, we also provide insights about the dependence of O3ResNet on its inputs and the lead time of a meteorological forecast. The paper concludes with a discussion of various aspects of O3ResNet, including a consideration of the benefits and limitations of O3ResNet, as well as thoughts on a road map toward operational deployment and the extension to forecasting other air pollutants.

2. Data and methods

a. Data

O3ResNet has been trained with data from 328 observation stations over central Europe (47.5°–56°N and 1.3°–18°E, see Fig. 1). We make use of the Tropospheric Ozone Assessment Report Database (TOAR DB; Schultz et al. 2017) and select all stations located in a rural or suburban environments and classified by the European Environmental Agency as background stations (European Parliament and Council of the European Union 2008). This means that there is no dominant air pollution source in the immediate vicinity. To prevent temporal data leakage, data are divided blockwise along the time axis into training (2000–15), validation (2016–18), and test (2019–21) data. Further details on the data split and a robustness analysis are presented in appendix A. Because of missing

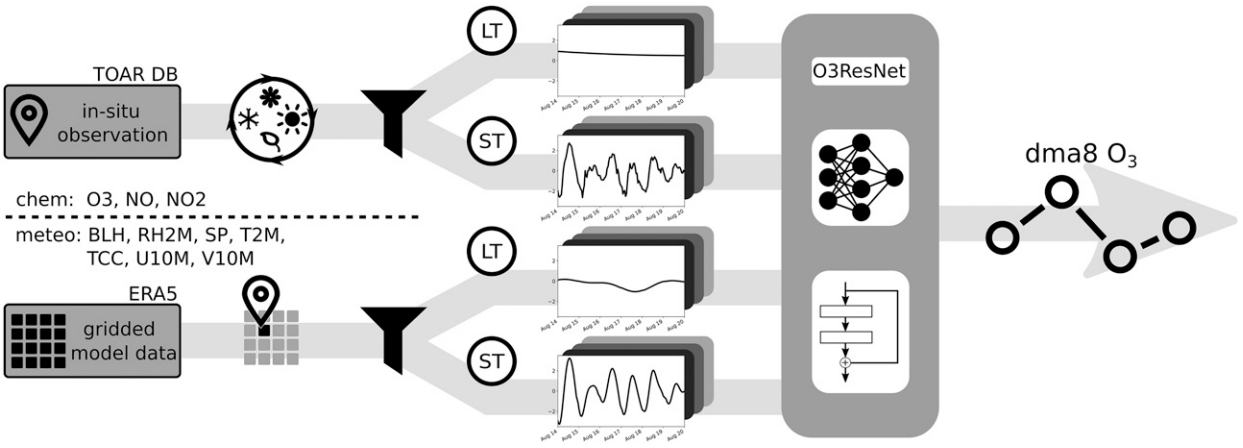

FIG. 2. Visualization of the training and inference workflow of O3ResNet as described in this paper. The chemical variables are taken from the TOAR DB (Schultz et al. 2017) as in situ observations and filtered into LT and ST components with the help of climatological statistics. The meteorological variables are obtained as gridded data from ERA5 and are mapped to the measurement stations by nearest neighbor and also split into LT and ST by filter. Note that variable names are listed according to the identifiers in the official documentation of TOAR DB (TOAR Data Team 2023) and ERA5 (Copernicus Climate Change Service 2022). All four branches are then input to O3ResNet, which makes a 4-day forecast of dma8 ozone.

or terminated observations, the number of stations varies for the validation (212) and test (202) subsets. In total, there are over 800 000 training samples, almost 200 000 for validation, and 170 000 for testing.

b. *Inputs*

For inputs, O3ResNet makes use of hourly time series of three chemical and seven meteorological variables at or near ground level: ozone (O_3), nitric oxide (NO), nitrogen dioxide (NO_2), cloud cover, planetary boundary layer height, pressure, relative humidity, temperature, and the zonal and meridional wind components. Relative humidity is calculated from temperature, dewpoint temperature, and pressure. The selection of these parameters is based on previous research in Leufen et al. (2022a), so we do not apply any new feature selection here. Chemical parameters (O_3, NO, and NO_2) are provided by the TOAR DB, and meteorological variables originate from the ERA5 reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al. 2020), with grid data mapped to station locations using nearest-neighbor interpolation. All time series are filtered into long-term (LT) and short-term (ST) components, with a finite impulse response (FIR) filter as in Leufen et al. (2022a). For causality reasons, we use the observations for lagged time steps ($t_i \leq t_0$) and climatology for time steps in the future ($t_i > t_0$) to calculate the LT and ST components of the chemical variables, as proposed in Leufen et al. (2022a). For the meteorological variables, we use reanalysis data as a pseudoforecast for all time steps t_i. A more detailed discussion on the time filtering can be found in appendix A. For the chemical inputs, we choose time steps of the past 3 days (72 h) from LT and ST components ($[t_0 - 3$ days, $t_0]$); the meteorological components cover, in addition, the forecast period on the interval of $[t_0 - 3$ days, $t_0 + 4$ days] with a total of 168 hourly values. All inputs are transformed by Z-score normalization to have 0 mean and a standard deviation of 1.

c. *Target*

The target variable of this study is dma8 ozone as defined by the European Parliament and Council of the European Union (2008) as the highest 8-h moving average of all ozone concentrations observed between 1700 local time of the previous day and 1600 local time of the current day. We predict dma8 ozone for the next 4 days ($[t_0 + 1$ day, $t_0 + 4$ days]). The daily resolved dma8 ozone for the model validation is obtained directly from TOAR DB. The temporal distribution of the target values in all subsets is shown in Fig. A2 in appendix A. Like the inputs, the targets are transformed by Z-score normalization. Figure 2 provides an overview of the entire workflow.

d. *Hyperparameter tuning*

We test different architectures like FCN, recurrent NN (RNN) based on LSTM and gated recurrent unit (GRU), CNN (with and without residual blocks), and U-Net. To find an optimal hyperparameter configuration for each NN architecture, we train NNs with various configurations over 100 epochs and evaluate the mean-square error (MSE) given by

$$\mathrm{MSE} = \frac{1}{n_i n_j} \sum_{i,j}^{n_i, n_j} (y_{i,j} - \hat{y}_{i,j})^2 \tag{1}$$

on the training and validation data, where n_i is the number of samples, n_j is the number of forecast steps, $y_{i,j}$ is the observed value, and $\hat{y}_{i,j}$ is the NN's forecast. After testing all alternative model architectures, we chose a CNN architecture with residual blocks as the best performing on validation data. In appendix B, we present details on the hyperparameter optimization and model selection strategies and provide technical background on the operating system, software, and the duration of preprocessing, training, and inference.

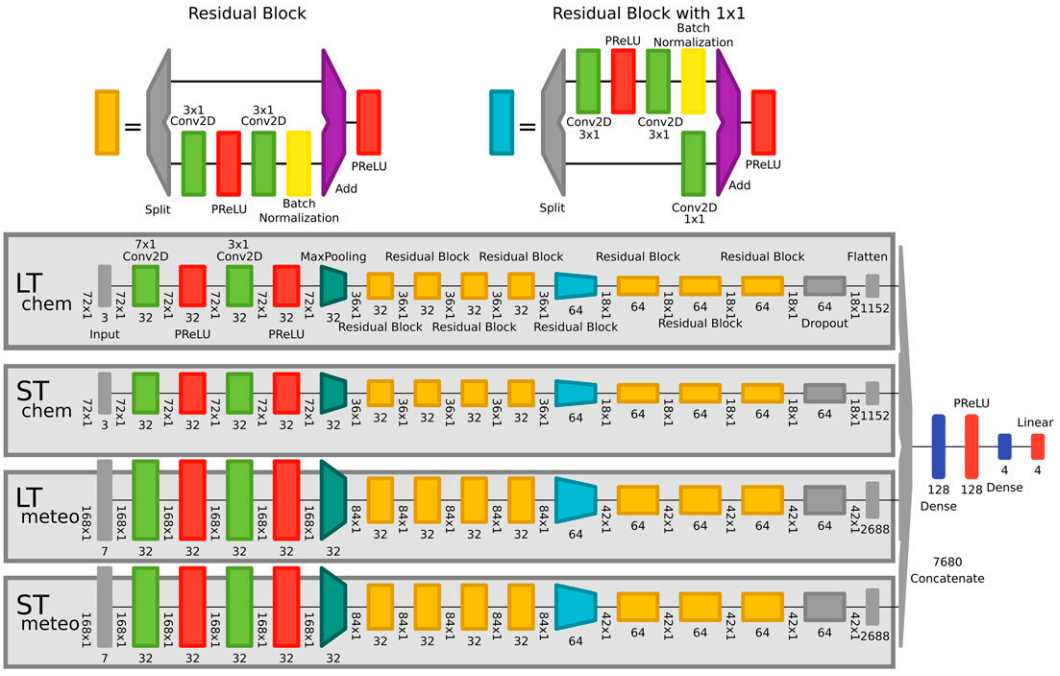

FIG. 3. Network architecture of O3ResNet consisting of convolutional layers (green), PReLU and linear activation (red), maxpooling layers (teal), batch normalization layers (yellow), residual blocks (orange), residual blocks with additional 1×1 filter to increase number of filters (cyan), dense layers (blue), add layer (purple), and input, dropout, flatten, concatenate, and split layers (all gray). Each branch is highlighted by a separate gray box. Numbers next to a layer show the numbers of filters and weights and the shape. Shapes of the inputs correspond to 72 hourly values for three chemical variables on the interval $[t_0 - 3 \text{ days}, t_0]$ and to 168 hourly values for seven meteorological variables on $[t_0 - 3 \text{ days}, t_0 + 4 \text{ days}]$. The graphic was created with Net2Vis (Bauerle et al. 2021) and edited afterward.

e. *CNN architecture*

Since we found the CNN architecture with residual blocks to be the best-performing DL architecture on validation data, we describe the exact architecture in more detail below. This is the model we refer to as O3ResNet. The O3ResNet architecture consists of eight residual blocks, 20 hidden layers, and a total of about 800 000 trainable parameters. A residual block consists of two convolutional layers, where the first layer is bypassed by a skip connection to stabilize the training and thus allow training of deeper networks, as gradients can propagate more directly during backpropagation (He et al. 2016). We follow Zheng et al. (2014) and apply all convolutions only along the time axis. A special feature of the O3ResNet architecture is the four input branches, consisting of an LT and ST component of the chemical and meteorological inputs. The motivation for these separate branches is that the NN can initially learn local features of the different variable types, chemical and meteorological variables, time scales, and LT and ST components, and later put this knowledge into a global context to make a prediction for ozone. The global context is learned in the tail of the network, starting from a concatenation layer up to the output layer. Each branch consists of two convolutional layers with thirty-two 7×1 and thirty-two 3×1 filters and a maxpooling operation (with pool size 2×1), succeeded by four residual blocks with thirty-two 3×1 filters and four residual blocks with sixty-four 3×1 filters. The outputs of each branch are flattened and concatenated into a layer, followed by a dense layer of 128 neurons and the output layer of four neurons, one for each day to be predicted. Except for the output layer, which features linear activation, all layers use a parametric rectified linear unit (PReLU; He et al. 2015) activation function. The architecture of O3ResNet is shown in Fig. 3. Appendix B gives further details on the O3ResNet's hyperparameters (see Table B4 of appendix B for a list of hyperparameters) and on the alternative network architectures.

f. *CAMS*

We compare O3ResNet with the state-of-the-art regional chemistry transport model ensemble with data assimilation from CAMS. The data are downloaded from the CAMS Atmosphere Data Store (ADS; ADS 2020) and preserved on local systems, as ADS hosts data on a rolling 3-yr archive. CAMS provides 96-h forecasts on a $0.1° \times 0.1°$ grid for Europe based on the median value approach of the nine ensemble members (Marécal et al. 2015). Details on the ensemble members are provided in appendix C. To produce the CAMS ensemble forecast, the median is calculated for each pixel individually using interpolated forecasts of all ensemble members. As CAMS provides a grid forecast, we apply nearest-neighbor interpolation to extract data at the station locations. We have

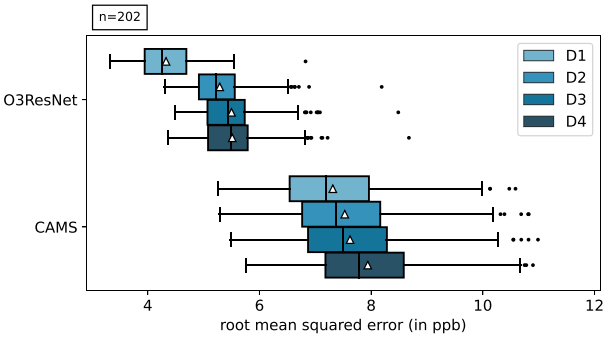

FIG. 4. Distribution of the RMSE of O3ResNet and CAMS over all test stations visualized as a box-and-whisker diagram. The different shades of blue correspond to the error from D1 (light blue) to D4 (dark blue). The boxes indicate the 25th and 75th quantile of the distribution, the line within the box shows the median, and the white triangle shows the mean.

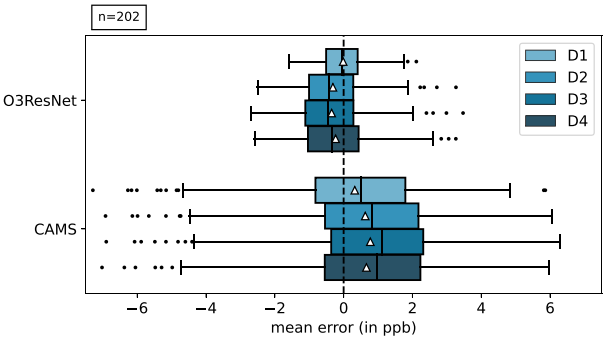

FIG. 5. As in Fig. 4, but for ME.

also tested bilinear interpolation as an alternative. Bilinear interpolation performed better at some stations and worse at others, so that on average, the choice of interpolation method has no discernable effect on the CAMS performance. Finally, dma8 ozone is calculated from the hourly data at each station. Peuch et al. (2022) provide a detailed overview on CAMS.

g. *Evaluation*

The final evaluation of the results is performed exclusively on the test data, which were used neither for training nor for hyperparameter optimization. For evaluation, we use the root-mean-square error (RMSE), which is given by the square root of the MSE from Eq. (1),

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}, \tag{2}$$

as well as the mean error (ME) given by

$$\mathrm{ME} = \frac{1}{n_i n_j} \sum_{i,j}^{n_i, n_j} (\hat{y}_{i,j} - y_{i,j}) = \bar{\hat{y}} - \bar{y}, \tag{3}$$

which can also be expressed as the difference between the means of forecast $\bar{\hat{y}}$ and observation $\bar{y}$. To compare models A and B with each other directly, we resort to the skill score given by

$$\mathrm{SS(A, B)} = 1 - \frac{\mathrm{MSE_A}}{\mathrm{MSE_B}}, \tag{4}$$

where $\mathrm{MSE_A}$ is the MSE of model A and $\mathrm{MSE_B}$ is the MSE of model B.

3. Results

Figure 4 shows the RMSE as a box-and-whisker diagram aggregated over all stations. O3ResNet yields a smaller RMSE for all forecast days when compared with CAMS. O3ResNet achieves the smallest error for the D1 forecast, with 4.3 ppb. The RMSE increases to 5.5 ppb for the D4 forecast, with almost identical RMSE on D3 and D4. Overall, the RMSE for O3ResNet lies between 3.9 and 5.8 ppb for the 25th and 75th percentiles of all stations. CAMS, on the other hand, shows a noticeably higher RMSE, with a mean RMSE ranging from 7.3 ppb on D1 to 7.9 ppb on D4. Moreover, a wider spread of RMSE across stations can be seen for CAMS. Thus, the 25th and 75th percentiles are 6.5 and 8.6 ppb, respectively. We also show the spatial distribution of the RMSE of O3ResNet and CAMS in appendix C (Figs. C2 and C3, respectively).

The ME shown in Fig. 5 provides insight into systematic biases between prediction and observation. For O3ResNet, we can see that the ME averaged over all stations is centered between −0.35 and −0.01 ppb for all forecast days, with an interquartile range (IQR) between 0.92 and 1.48 ppb. The ME for the CAMS predictions averages between +0.32 and +0.78 ppb, with the median for D2–D4 being larger than +0.83 ppb. Overall, the CAMS ME shows a wide variation, with an IQR of >2.6 ppb.

The analysis of the ME shows that CAMS suffers from a consistent high bias in relation to the observations. Therefore, we next correct all forecasts of CAMS and O3ResNet by 1) removing the averaged background value for each station and 2) subtracting a 30-day running mean from the forecasts for each station. This reveals what contribution to the total error is due to an improper accounting of the variability of ozone and what contribution is due to a systematic deviation. Figure 6 shows the results for bias-corrected predictions using method 1. Here, the

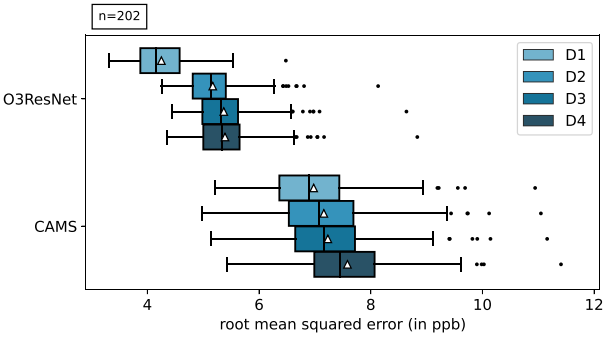

FIG. 6. As in Fig. 4, but for the bias-corrected RMSE. The correction is applied by removing the average background concentration at each station.

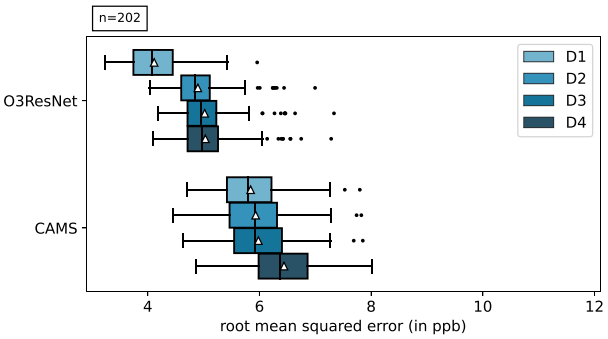

FIG. 7. As in Fig. 4, but for the seasonal bias-corrected RMSE. For bias correction, we remove a 30-day running mean for each station.

adjustment leads to a reduction in the RMSE for the CAMS predictions. Accordingly, since O3ResNet already exhibited a low ME, this postprocessing method does not lead to any improvement for O3ResNet. In contrast, the bias-corrected forecasts using method 2 lead to an improvement for CAMS and O3ResNet as measured by the RMSE (see Fig. 7). In all cases, it can be concluded that O3ResNet can better represent the variability of ozone.

Since ozone concentrations exhibit pronounced seasonal variation and the variance also varies with season, we next consider the seasonality of the error. Figure 8 shows the RMSE aggregated over all forecast steps for each month across all stations for the entire test period. For each individual month, O3ResNet has a lower RMSE than CAMS. In addition, the IQR indicated by the width of the band of quantiles is narrower for O3ResNet. Both findings are in line with the results presented so far. Indeed, we can identify a season-dependent performance for both O3ResNet and CAMS in Fig. 8. Overall, both models perform best in the spring months March–May (MAM), whereas the summer months June–August (JJA) show the highest error. Note that O3ResNet can provide notably better forecasts than CAMS for JJA 2019 but that for JJA 2021 neither model can provide decent forecasts, especially in July.

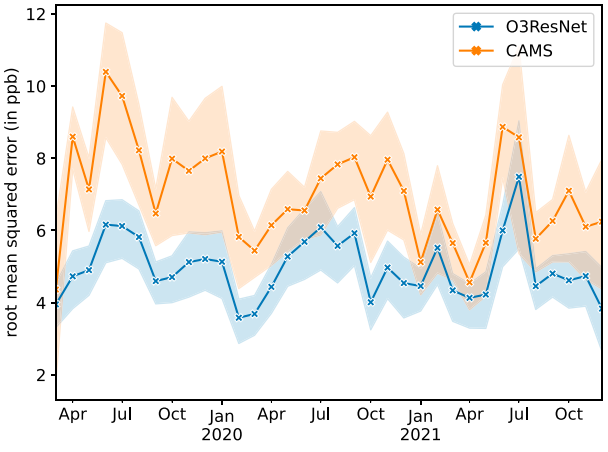

FIG. 8. Month-to-month variation of performance (RMSE) for O3ResNet (blue) and CAMS (orange) during the test period. Mean RMSE over all stations is shown as a thick line with crosses, and 25th and 75th quantiles are illustrated as bands.

To provide further insight into the quality of the O3ResNet forecasts, we use the likelihood-base rate factorization after Murphy and Winkler (1987). Figure 9 compares observation and prediction of O3ResNet. Shown in the dashed lines is the conditional distribution of the probability that, given a particular observation, O3ResNet can issue a proper forecast in advance. Considering the climatological distribution of the observations, represented by the gray bars (marginal distribution), this view allows us to draw conclusions about how well O3ResNet can discriminate between different observation events (Wilks 2006). It can be seen that the reference line and the median of the conditional quantile are in agreement within the interval between 20 and 55 ppb, and thus, O3ResNet can distinguish individual observations well in this interval. However, for small ozone values, the model tends to overestimate slightly, indicated by the fold in the lines of the conditional quantiles. Also, for ozone values exceeding 60 ppb, 03ResNet cannot fully follow the observations, tending to underestimate the ozone concentration. However, observations of high ozone concentrations are severely underrepresented in the training data, and regression approaches such as O3ResNet generally tend to favor values toward the mean. For the forecast horizon, increasing uncertainty with lead time is visible as the lines of the quantiles of the conditional distribution for D4 of the forecast are more widely spaced and both ends of the lines curve more pronouncedly than for D1. Between 20 and 50 ppb, however, the reference lines and median continue to be close to each other, indicating a reliable forecast issued by O3ResNet. The likelihood-base rate factorization for CAMS can be found in appendix C in Fig. C1. Here it can be seen that CAMS is not able to distinguish well between different observation events, because the slope of the conditional quantile lines deviates from the ideal reference line, meaning that smaller values are generally overestimated and high concentrations are underestimated.

a. Importance of input branches

To shed light on the robustness of the O3ResNet forecasts, we follow the single-pass approach (Breiman 2001). To understand the impact of each individual branch on O3ResNet, we fix all inputs of a single input branch to their average values and examine how much the resulting prediction differs from the unperturbed prediction. We measure this by the skill score as shown in Eq. (4). The more the skill score of the mean-fixed O3ResNet decreases with respect to the original O3ResNet forecasts, the greater the influence of the respective branch. Results are presented in Fig. 10. Considering all forecast days, the LT chemical and ST meteorological inputs have the strongest influence on the predicted ozone concentrations. The LT chemical inputs are particularly important for the D1 forecast and appear to be less important for D2 to D4. Moreover, for the D1 forecast, the ST component of the chemical inputs is important to some extent, whereas for the other days this is not evident. LT meteorological inputs only play a minor role for O3ResNet for all forecast days. In

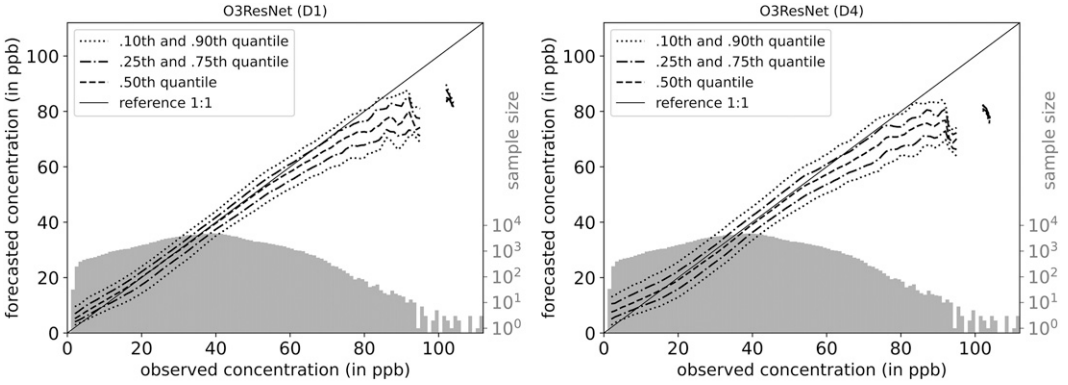

FIG. 9. Visualization of the likelihood-base rate factorization for the (left) D1 and (right) D4 forecast of O3ResNet. The factorization consists of the conditional distribution of the probability that a prediction is made in advance of an incoming observation and the frequency distribution of the observations. The conditional distribution is represented by the 10th, 25th, 50th, 75th, and 90th quantiles using different dashed lines and the optimal reference line. The frequency distribution of the observations is shown by a histogram (gray bars) with logarithmic scale on the right axis.

contrast, the ST meteorological components are relevant for all forecast days, and their importance even increases from D1 to the following days.

From a meteorological perspective, these sensitivities can be explained as follows. The LT chemical inputs allow O3ResNet to perform a bias correction, as they provide information about the long-term background concentration. In addition, these components also add information about the season, since, for example, average ozone concentrations are higher in summer than in winter. Note that O3ResNet has no explicit information about the day or month of the samples it is processing. The relevance of the ST chemical variables can be explained by the autocorrelation of ozone. As it decreases with lead time, the importance of past observations also drops. By contrast, the LT components of the meteorological variables cannot add any valuable information to O3ResNet since all information about seasonality is already contained in the LT chemical variables. However, the ST components of the meteorological inputs play an important role, since the deviations from long-term conditions contained therein characterize the current and future weather situation. For example, the ST meteorological variables provide information about the daily maximum temperature and humidity in the forecast horizon.

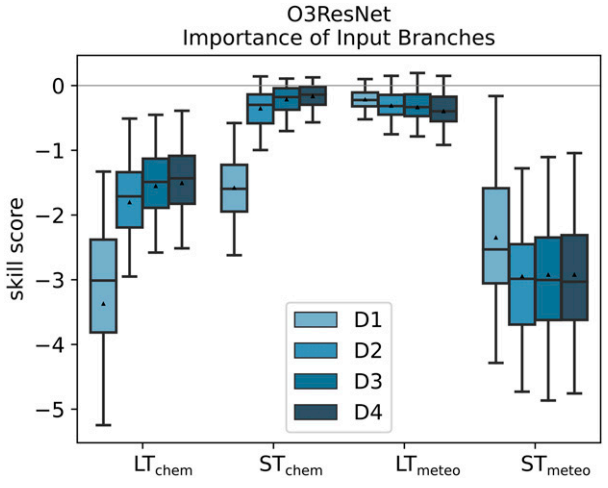

FIG. 10. Evaluation of the importance of each input branch for the prediction of O3ResNet generated using the single-pass approach. The skill score is calculated in reference to the unperturbed prediction. The impact on each prediction day is shown by blue colors from D1 (light blue) to D4 (dark blue). A large negative value indicates a strong dependence, whereas a value close to 0 describes a weaker dependence.

b. *Influence of the meteorological forecast lead time*

Since this study uses ERA5 data as a pseudoforecast and over an extended time horizon to calculate the LT and ST components (see appendix A), questions arise as to how O3ResNet would behave in an operational setting where meteorological forecasts have a more limited lead time and the forecast error tends to grow with increasing lead time. A sensitivity study, outlined subsequently, reveals that the forecast quality of O3ResNet is hardly affected by reducing the lead time of the meteorological forecast down to 4 days. To conduct this sensitivity study, we gradually decrease the maximum lead time for the meteorological variables. Values after this maximum lead time are replenished by the climatological statistics, as described in Leufen et al. (2022a) and as is done for the chemical variables. We do not retrain O3ResNet on these modified inputs but analyze how O3ResNet responds to this new information and whether the skill of the ozone prediction decreases in dependence on the meteorological forecast lead time. Results are shown in Fig. 11.

At large lead times, it can be seen that the reduction of the lead time of the meteorological variables from 168 to 93 h has no effect on the forecast performance of O3ResNet, as the skill score stays close to zero, indicating neither a gain nor a loss of skill. Note that we test with larger lead times than the 4-days forecast horizon of O3ResNet, as longer time series are mandatory to calculate an exact LT and ST decomposition (see appendix A). As this analysis shows, a blurred decomposition does not decrease the model's

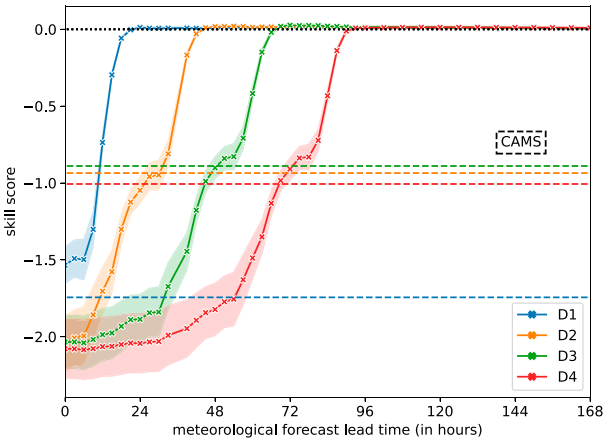

FIG. 11. Skill score of the forecast quality of O3ResNet depending on the lead time of the meteorological forecast in relation to a forecast based on quasi-unlimited lead time. The forecast days are individually colored for D1 (blue), D2 (orange), D3 (green), and D4 (red). The solid lines represent the mean skill scores, and the bands represent the range between the 25th and 75th quantiles. In addition, the skill scores for CAMS in relation to the original O3ResNet forecast are shown as dashed reference lines. Negative skill scores mean that the forecast for the corresponding meteorological forecast lead time is worse than the best case. At a skill score of 0, the difference disappears.

performance at all. A further decrease of the lead time up to the extreme case of 0 h results in a continuous decrease of the prediction skill for all days. Therefore, the forecast of O3ResNet always deteriorates only for the forecast days for which no meteorological forecast is available and climatology is fallen back on as a substitute. For example, when the lead time of the meteorological variables is 48 h, only the ozone forecasts for D3 and D4 worsen, with the D3 forecast having an equal skill to the CAMS forecast in this particular case. Conversely, the ozone forecasts for D1 and D2 are not affected at all and remain at their original skill level. This finding can be observed for all forecast days. Besides, results show that the D1 forecast of O3ResNet is more skillful than CAMS even at a lead time of 0 h.

4. Discussion and conclusions

This paper outlines the development of a skillful and reliable forecasting system for a 4-day point forecast of dma8 ozone based on DL methods. O3ResNet performs better than the state-of-the-art CAMS regional ensemble. O3ResNet was developed with data from central Europe but can easily be trained for other regions and, in principle, for other ozone metrics or even other air pollutants such as particulate matter or nitrogen oxides, provided sufficient data are available. The transferability of O3ResNet will be the subject of another study. The results above show that the combination of a CNN architecture with residual blocks, the temporal decomposition of inputs into long term and short term, and the integration of a weather forecast for all meteorological input parameters are the key ingredients for our new high-quality ozone forecasting system.

The outstanding advantages of O3ResNet are a nearly bias-free forecast as well as a low seasonal variation of the forecast quality. O3ResNet provides high-quality predictions, especially in the range of 20–55 ppb and for September–May. Only at the edges of the distribution and for forecasts during the summer season does the performance decrease a bit, although O3ResNet still outperforms the CAMS regional model ensemble. First, from a statistical point of view, this is related to heteroscedasticity since the variability of ozone is very high in summer and lower in winter. Second, ozone in summer is more determined by the local daily maximum temperature (Otero et al. 2016), which is less well reflected in the meteorological forecasts due to limited spatial model resolution. While such processes generally pose a problem for conventional CTMs as well (Stock et al. 2014; Young et al. 2018), O3ResNet can at least better accommodate them. The nearly bias-free forecast can be attributed to O3ResNet's understanding of the LT chemical variables, which allows O3ResNet to determine a correct concentration level at the target station. The ST meteorological inputs make a major contribution to the O3ResNet forecast quality, because they provide information about the current weather situation.

Analysis of the dependence on the horizon of the weather forecast shows that O3ResNet can already provide a fully reliable forecast of future ozone concentrations with a weather forecast of similar lead time. With a 48-h weather forecast, O3ResNet achieves an adequate 2-day forecast. This shows, with respect to previous studies such as Kleinert et al. (2021) or Leufen et al. (2022a), that ozone prediction with DL methods is limited not by a lack of understanding the relationship between weather and air quality but, in particular, by uncertainty about future weather, and that the inclusion of a skillful weather forecast contributes great value to DL-based ozone predictions.

In comparison with the CAMS regional ensemble median forecast, O3ResNet shows significant improvements for all forecast days. Moreover, CAMS requires additional postprocessing to deliver forecasts on a station level, whereas O3ResNet does not. Peuch et al. (2022) mention the development of various postprocessing methods, including ML, to adapt the raw CAMS forecasts to point forecasts with higher skill that are expected to be deployed in the coming years. O3ResNet demonstrates that high-quality ozone forecasts do not necessarily require the running of a complete CTM system but can alternatively also be produced using DL plus weather forecasts, which is much faster. A 4-day forecast at all 328 stations of this study takes about 10 s.

In conclusion, we suggest a number of tests and improvements before applying O3ResNet operationally. First, ERA5 is not a real forecast, but a reanalysis, meaning that the frequency of updates through data assimilation is much higher. Nevertheless, it can be reasonably expected that the forecast quality of O3ResNet would not drop dramatically, as relevant numerical weather prediction on comparable spatial and temporal resolution, such as the Integrated Forecasting System (IFS) operated by the ECMWF, already provides a very reliable forecast for one week ahead (see Bauer et al. 2015; Haiden et al. 2022). Second, O3ResNet is currently trained in rural and

suburban areas on stations classified as background. To provide a full range of forecasts, the model should also be tested in urban areas as well as in regions with dominant air pollution sources, which may require the integration of emissions data. Third, it is recommended that the predictive power for peak ozone concentrations be further investigated. Although O3ResNet is capable of simulating concentrations of dma8 ozone up to 80 ppb, the most extreme observed values are not reproduced satisfactorily. For example, O3ResNet for July 2021 does not match with observations well. Herein, uncertainty prediction, for example, using probabilistic DL architectures as in Foster et al. (2021) or following Barnes et al. (2021), who predict the parameters of a probability distribution instead of the deterministic values, could add useful information. Also, transformers (Vaswani et al. 2017), or more specifically, a temporal fusion transformer (Lim et al. 2021), harbors promising potential. In combination with suitable interpolation techniques, such DL models may even be able to generate useful forecasts at locations where no measurements of air pollutant concentrations are performed.

Acknowledgments. We thank Dr. Amirpasha Mozaffari and Enxhi Krespha for downloading the CAMS data, CAMS and the TOAR data center teams for providing the data, and the Jülich Supercomputing Center for computing time under the DeepACF project. This research has been supported by the European Research Council, H2020 Research Infrastructures (IntelliAQ; Grant 787576) and the DFG Funding Programme, Open Access Publication Funding (2022–24).

Data availability statement. Input data, forecasts on test data, and O3ResNet model are openly available online (http://doi.org/10.34730/76529959732a464486ec5b9277152233).

APPENDIX A

Cross Validation and Data Filtering

We perform a cross validation of the best model architecture (O3ResNet) by rotating the subsets, keeping the length of each subset, 3 yr for validation and testing and 12 yr for training, as well as the hyperparameter configuration. Data are always split blockwise along time. Therefore, in total, we test six different arrangements. Results are shown in Table A1 and Fig. A1. It can be seen that the RMSE is close for all orderings of subsets. Yet there is a deviation in performance when positioning the testing phase at the very beginning. Note that the number of samples varies from about 160 000 (train/validate/test) to 225 000 (test/train/validate) because of a large temporal and spatial variability of data coverage. In Fig. A2, we furthermore show the temporal distribution of the target dma8 ozone in the final subset ordering (train/validate/test). It can be seen that the temporal distribution is very similar for all subsets.

TABLE A1. Tabular results of cross validation implemented by rotating training, validation, and testing subsets, showing the RMSE (ppb), also visualized in Fig. A1.

Data split	D1	D2	D3	D4	Mean for D1–D4
Train/validate/test	4.55	5.43	5.69	5.73	5.35
Train/test/validate	4.59	5.56	5.64	5.72	5.38
Validate/train/test	4.25	5.14	5.25	5.38	5.01
Validate/test/train	4.71	5.71	5.87	5.90	5.55
Test/train/validate	5.13	6.28	6.45	6.51	6.09
Test/validate/train	5.25	6.42	6.69	6.73	6.27

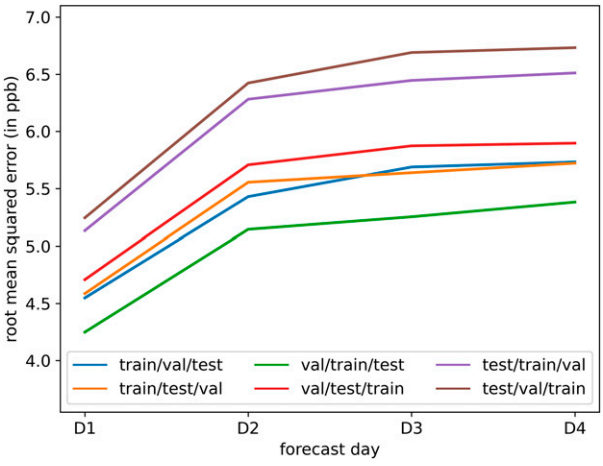

FIG. A1. Visualization of cross-validation results, as shown in Table A1.

To filter the data, all time series are split into LT and ST components by means of an FIR filter with a Kaiser window (Kaiser 1966) with parameter $\beta = 5$, a cutoff period of 21 days, and order of $N = 42$ days. For applying the FIR filter causally to all chemical variables, we follow the approach of Leufen et al. (2022a) and use climatology for time steps in the lead time, whereas reanalysis data are used as a pseudo-forecast for the meteorological variables.

The decomposition is formalized by the following steps. First, we calculate a climatological statistic a_i that contains the

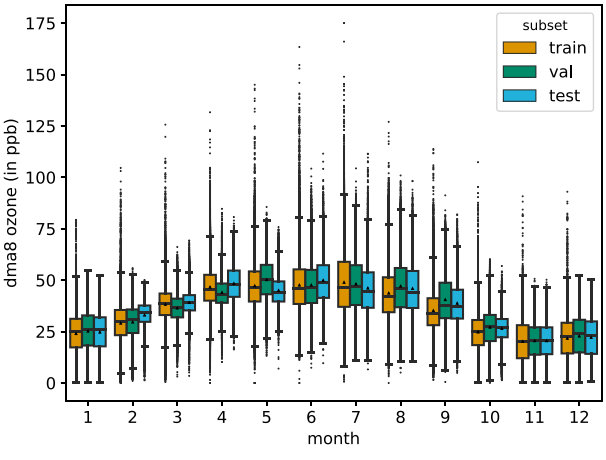

FIG. A2. Temporal distribution of dma8 ozone aggregated over all observations and stations illustrated as a box-and-whisker diagram. Distributions of the training (orange), validation (green), and testing (blue) data are highlighted in color.

seasonal cycle of the monthly mean as well as the diurnal cycle. Heteroscedasticity is taken into account by allowing this diurnal cycle to vary over the year:

$$a_i = f(x_i,\ t_i). \tag{A1}$$

A composite time series x_i is created from the raw time series x_i and the climatological statistic a_i for each time t_0 at which a forecast is initiated. The combination is done depending on the lead time t_l. For the chemical variables, $t_l = 0$ always applies, and for the meteorological variables, $t_l \to \infty$. For the analysis of the dependence of O3ResNet on the lead time of the meteorological variables, t_l is set to a lead time of between 0 and 168 h accordingly:

$$\breve{x}_i(t_0) = \begin{cases} x_i, & t_i \le t_0 + t_l \\ a_i, & t_i > t_0 + t_l \end{cases}. \tag{A2}$$

The properties b_i of the FIR filter are determined by the Kaiser window given for the order of $N = 42$ days. Applying the filter results in the LT components $x_i^{(\mathrm{LT})}$ of the time series:

$$x_n^{(\mathrm{LT})}(t_0) = \sum_{i=t_0-N/2}^{t_0+N/2} b_i \breve{x}_{n-i}(t_0). \tag{A3}$$

Last, the ST components $x_i^{(\mathrm{ST})}$ are calculated by the difference between the original time series x_i and the LT components $x^{(\mathrm{LT})}$:

$$x_i^{(\mathrm{ST})}(t_0) = x_i - x_i^{(\mathrm{LT})}(t_0). \tag{A4}$$

This means in reverse that the sum of LT and ST components always adds up to the original time series.

APPENDIX B

Technical Details

We train all NNs for this study on the Helmholtz Data Federation Machine Learning System (HDF-ML) at the Jülich Supercomputing Centre in Jülich, Germany. In total, HDF-ML is equipped with 15 computer nodes, each running 4 Nvidia Tesla V100 GPUs and 2 Intel Xeon Gold 6126 with 12 cores (24HT). For each training, we use a single node with all available GPUs since the computation times of the training are moderate (between half an hour and up to 4 h). Training as well as pre- and postprocessing are carried out with the research software MLAir (Leufen et al. 2022b). MLAir is based on the programming language Python, provides a complete workflow for performing ML experiments with a special focus on time series predictions (Leufen et al. 2021), and thereby makes use of TensorFlow (Abadi et al. 2015) for the ML training.

TABLE B1. Preprocessing and inference time.

Operation	Data	Duration (s)
Preprocessing	Station	108
Preprocessing	Sample	0.03
Inference	Station	2.8
Inference	Sample	0.0009

Preprocessing of the raw data of a single station covering the entire time period takes on average 108 s, which means that preprocessing of a single sample is about 0.03 s on average. Approximately 90% of the preprocessing time is spent calculating the decomposition into LT and ST components, as the data for each sample change with t_0. For this study, we use 12 parallel threads, so preprocessing is 12 times as fast on our systems. The inference time for a single station is approximately 2.8 s (0.0009 s per sample). Measured inference time includes losses due to input/output (I/O) operations such as loading the processed data from disk and storing the predictions locally. The actual NN prediction, without I/O operations, is performed on 4 GPUs in parallel. Numbers are also shown in Table B1.

With regard to hyperparameter tuning strategy, we apply a kind of evolutionary algorithm when searching for optimal hyperparameters. For the initial first generation, we randomly draw 70 combinations of hyperparameters according to the range of values, the sampling mode, and the variation properties shown in Tables B2 and B3 and measure the validation error. For the second generation, we select the top 10 performing hyperparameter combinations in terms of validation error and again draw random combinations from this new set, allowing all parameters to further vary according to the specified variation properties. We do not

TABLE B2. Overview of all hyperparameters tuned in this study. Each parameter is selected from the given range and with the indicated sampling method. Moreover, continuous parameters are varied according to the variation ratio. Details on the NN architectures are provided in Table B3. Parameters marked with a dagger symbol are not tested for ResNet and U-Net.

Parameter	Range	Sampling	Variation (%)
Learning rate	[0.0001, … , 0.1]	Power of 10	80
Learning rate decay	[0, 0.001, … , 0.1]	Power of 10	50
Batch size	{256, 512, 1024}	Discrete	–
Dropout	[0, … , 0.7]	Linear	50
Batch normalization	{true, false}	Discrete	–
l1 regularizer	[0, 0.001, … , 0.1]	Power of 10	50
l2 regularizer	[0, 0.001, … , 0.1]	Power of 10	50
Activation function	{relu, leakyrelu, prelu, elu†, selu†, tanh†}	Discrete	—
NN architecture	See Table B3	Discrete	—

TABLE B3. List of NN specific hyperparameters referring to the model architecture. The model column contains information about the chosen architecture and number of different configurations. A slash in the values column indicates the number of neurons' respective filters per layer.

Model	Parameter	Values
FCN (10×)	Hidden layers and neurons	{32, 64, 64/32, 128/32, 128/64, 128/64/32, … , 512/256/128}
	Dense layer (after concatenation)	{no, 256, 256/64, 256/64/16}
CNN (12×)	Layers	[1, 2, … , 6]
	Max pooling	{no, every second layer, always}
	Kernel size	{(3, 1), (5, 1)}
	Filter	{16, 32, 64, 128}
	Dense layer (after concatenation)	{no, 128, 256}
RNN (10×)	Recurrent layer	{10, 32, 32/32, 64, 64,/64, 64/32, … , 256/128}
	Unit type	{LSTM, GRU}
	Dense layer (after concatenation)	{no, 32, 64, 128}
ResNet (16×)	Residual blocks	[6, 7, … , 12]
	Kernel size	{(3, 1)}
	Filter	{16/32/64, 16/32, 32/64, 32/64/128}
	Consecutive layers with same filter	[2, 4]
	Dense layer (after concatenation)	{no, 128}
U-Net (9×)	Down blocks with filter	{16/32, 16/32/64, 16/32/64/128}
	Kernel size	{(3, 1)}
	Dense layer (before concatenation)	{no, 128}
	Dense layer (after concatenation)	{no, 128}
All	Dropout	{no, only final layer, every second layer, always}
	Output activation	{linear}

test the exact same combination a second time. For the second generation, we reduce the number of experiments by 30%. In each subsequent generation, we apply the same scheme but reduce the number of best performing combinations by 1 and the number of experiments by 30% each time. After running 10 generations, we consider the combination that leads to the lowest validation error across generations to be the optimal choice of hyperparameters. We apply this search strategy separately for each NN architecture. Table B4 gives the list of hyperparameters.

TABLE B4. Summary of the hyperparameters of O3ResNet.

Parameter	Range
Learning rate	0.0003
Learning-rate decay	0.0
Batch size	1024
Dropout	0.59
Batch normalization	False
l1 regularizer	0.095
l2 regularizer	0.12
Activation function	Prelu
NN architecture	See Fig. 3
Trainable parameters	807, 812

To select the best DL architecture/model, we look at the RMSE over all stations (Fig. B1). It can be seen that the forecasts of the CNN with residual blocks (ResNet) and U-Net are, with an average RMSE of 5.1 ppb, better than those of the other DL models (between 5.6 and 5.8 ppb). However, the distributions of the RMSE for ResNet and U-Net do not differ significantly in a Mann–Whitney U test. Therefore, we apply a bootstrap procedure with 1000 repetitions as a second evaluation step. We split the entire test dataset into monthly blocks, randomly sample 36 blocks with replacement for each iteration, and calculate the RMSE on each sample. In the bootstrap approach as shown in Fig. B2, the ResNet architecture performs slightly better, so we use it for further analysis.

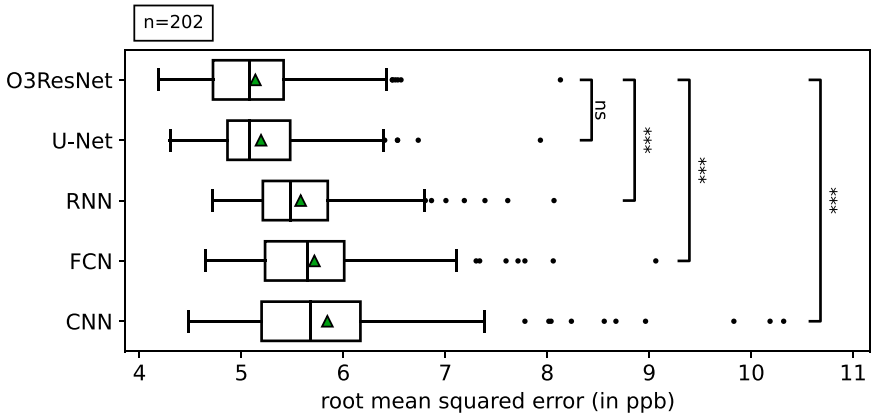

FIG. B1. Distribution of the RMSE aggregated over test data ($n = 202$ stations) visualized as a box-and-whisker diagram. Results from a Mann–Whitney U test are shown additionally. The presence of three asterisks indicates a significance level of $p < 0.001$, and "ns" (not significant) corresponds to $p > 0.05$.

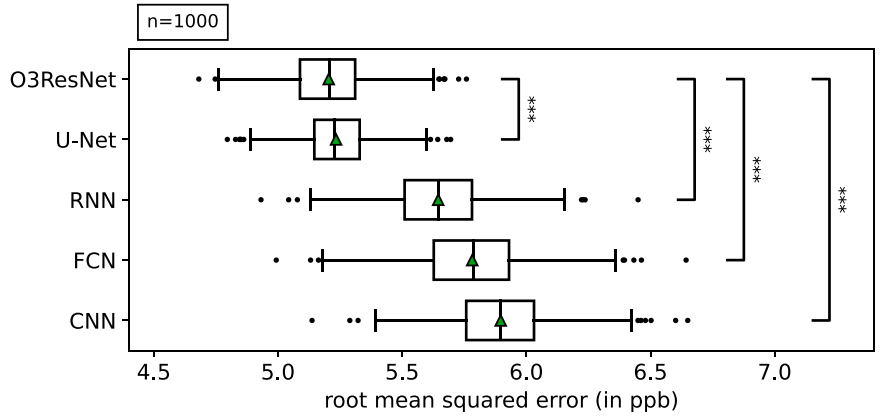

FIG. B2. Distribution of the RMSE calculated on $n = 1000$ bootstrap samples (with replacement) plotted as a box-and-whisker diagram. Significance levels are indicated as in Fig. B1.

APPENDIX C

Additional Information on CAMS

A good overview of the regional CAMS ensemble can be found in Peuch et al. (2022). The regional CAMS ensemble is composed of the nine members: "CHIMERE" (Menut et al. 2013), Danish Eulerian Hemispheric Model (DEHM; Christensen 1997), European Monitoring and Evaluation Programme (EMEP; Simpson et al. 2012), European Air Pollution Dispersion-Inverse Model (EURAD-IM; Hass et al. 1995; Memmesheimer et al. 2004), Global Environmental Multiscale Air Quality model (GEM-AQ; Kaminski et al. 2008), Long Term Ozone Simulation European Operational Smog (LOTOS-EUROS; Schaap et al. 2008), Multi-Scale Atmospheric Transport and Chemistry model (MATCH; Robertson et al. 1999; Andersson et al. 2015), Modèle de Chimie Atmosphérique à Grande Echelle (MOCAGE; Josse et al. 2004; Dufour et al. 2005), and System for Integrated Modelling of Atmospheric Composition (SILAM; Sofiev et al. 2008). Each model is first interpolated on a $0.1° \times 0.1°$ grid individually, and then the median is calculated for each grid cell. More information about this median value approach and in-depth details about the ensemble members involved are presented in Marécal et al. (2015).

With regard to the joint distribution of CAMS, Fig. C1 gives the likelihood-base rate factorization for CAMS; it can be seen that CAMS is not able to distinguish well between different observation events, with smaller values generally overestimated and high concentrations underestimated. In terms of error maps, the spatial distribution of the RMSE is given for O3ResNet (Fig. C2) and CAMS (Fig. C3).

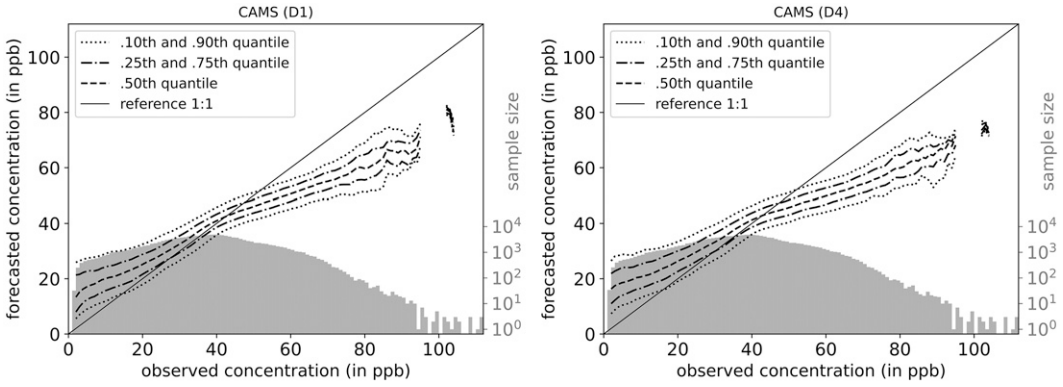

FIG. C1. As in Fig. 9, but for CAMS.

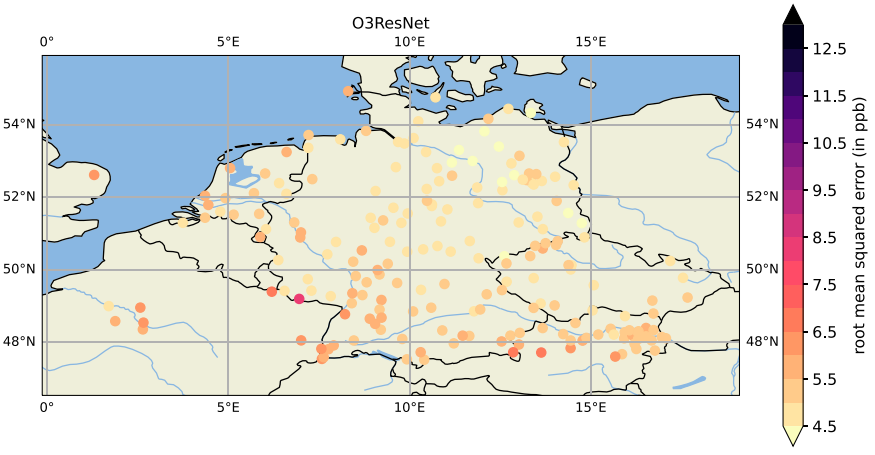

FIG. C2. Spatial distribution of the RMSE of O3ResNet averaged on all forecast days at each observation station.

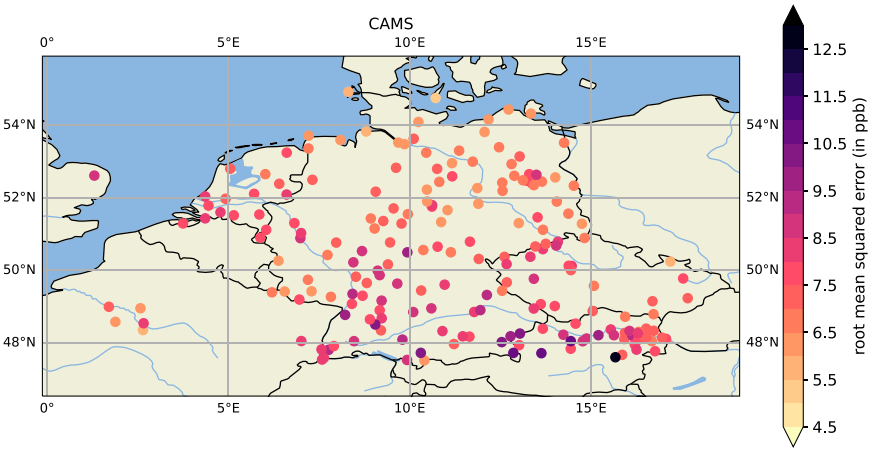

FIG. C3. As in Fig. C2, but for the CAMS forecast.

REFERENCES

Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous distributed systems. TensorFlow, 19 pp., https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf.

ADS, 2020: CAMS European air quality forecasts. Atmosphere Data Store, Copernicus Atmosphere Monitoring Service, accessed 31 May 2022, https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-forecasts?tab=overview.

Andersson, C., R. Bergström, C. Bennet, L. Robertson, M. Thomas, H. Korhonen, K. E. J. Lehtinen, and H. Kokkola, 2015: MATCH-SALSA—Multi-scale Atmospheric Transport and Chemistry model coupled to the SALSA aerosol microphysics model. Part 1: Model description and evaluation. *Geosci. Model Dev.*, **8**, 171–189, https://doi.org/10.5194/gmd-8-171-2015.

Baklanov, A., and Coauthors, 2014: Online coupled regional meteorology chemistry models in Europe: Current status and prospects. *Atmos. Chem. Phys.*, **14**, 317–398, https://doi.org/10.5194/acp-14-317-2014.

Barnes, E. A., R. J. Barnes, and N. Gordillo, 2021: Adding uncertainty to neural network regression tasks in the geosciences. arXiv, 2109.07250v1, https://doi.org/10.48550/ARXIV.2109.07250.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bauerle, A., C. van Onzenoodt, and T. Ropinski, 2021: Net2vis – A visual grammar for automatically generating publication-tailored CNN architecture visualizations. *IEEE Trans. Visualization Comput. Graphics*, **27**, 2980–2991, https://doi.org/10.1109/tvcg.2021.3057483.

Bell, M. L., A. Zanobetti, and F. Dominici, 2014: Who is more affected by ozone pollution? A systematic review and meta-analysis. *Amer. J. Epidemiol.*, **180**, 15–28, https://doi.org/10.1093/aje/kwu115.

Bessagnet, B., and Coauthors, 2016: Presentation of the EURODELTA III intercomparison exercise – Evaluation of the chemistry transport models' performance on criteria pollutants and joint analysis with meteorology. *Atmos. Chem. Phys.*, **16**, 12 667–12 701, https://doi.org/10.5194/acp-16-12667-2016.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brunner, D., and Coauthors, 2015: Comparative analysis of meteorological performance of coupled chemistry-meteorology models in the context of AQMEII phase 2. *Atmos. Environ.*, **115**, 470–498, https://doi.org/10.1016/j.atmosenv.2014.12.032.

CAMS, 2020: Regional production, updated documentation covering all regional operational systems and the ENSEMBLE. Copernicus Atmosphere Monitoring Service, ECMWF Copernicus Tech. Rep., 81 pp., https://atmosphere.copernicus.eu/sites/default/files/2020-09/CAMS50_2018SC2_D2.0.2-U2_Models_documentation_202003_v2.pdf.

Christensen, J. H., 1997: The Danish Eulerian hemispheric model—A three-dimensional air pollution model used for the Arctic. *Atmos. Environ.*, **31**, 4169–4191, https://doi.org/10.1016/S1352-2310(97)00264-1.

Copernicus Climate Change Service, 2022: ERA5: Data documentation. ECMWF, accessed 18 January 2023, https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation.

Dufour, A., M. Amodei, G. Ancellet, and V.-H. Peuch, 2005: Observed and modelled "chemical weather" during ESCOMPTE. *Atmos. Res.*, **74**, 161–189, https://doi.org/10.1016/j.atmosres.2004.04.013.

European Parliament and Council of the European Union, 2008: 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. EUR-Lex, https://data.europa.eu/eli/dir/2008/50/oj.

Fiore, A. M., V. Naik, and E. M. Leibensperger, 2015: Air quality and climate connections. *J. Air Waste Manage. Assoc.*, **65**, 645–685, https://doi.org/10.1080/10962247.2015.1040526.

Fleming, Z. L., and Coauthors, 2018: Tropospheric ozone assessment report: Present-day ozone distribution and trends relevant to human health. *Elementa*, **6**, 12, https://doi.org/10.1525/elementa.273.

Foster, D., D. J. Gagne, and D. B. Whitt, 2021: Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002474, https://doi.org/10.1029/2021MS002474.

Haiden, T., M. Janousek, F. Vitart, Z. Ben-Bouallegue, L. Ferranti, F. Prates, and D. Richardson, 2022: Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Tech. Memo. 902, 56 pp., https://www.ecmwf.int/en/elibrary/81321-evaluation-ecmwf-forecasts-including-2021-upgrade.

Hass, H., H. J. Jakobs, and M. Memmesheimer, 1995: Analysis of a regional model (EURAD) near surface gas concentration predictions using observations from networks. *Meteor. Atmos. Phys.*, **57**, 173–200, https://doi.org/10.1007/BF01044160.

He, K., X. Zhang, S. Ren, and J. Sun, 2015: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proc. IEEE Int. Conf. on Computer Vision*, Santiago, Chile, IEEE, 1026–1034, https://doi.org/10.1109/ICCV.2015.123.

——, ——, ——, and ——, 2016: Deep residual learning for image recognition. *2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, IEEE, 770–778, https://doi.org/10.1109/CVPR.2016.90.

He, T.-L., and Coauthors, 2022: Deep learning to evaluate US NO_x emissions using surface ozone predictions. *J. Geophys. Res. Atmos.*, **127**, e2021JD035597, https://doi.org/10.1029/2021JD035597.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Jahn, S., and E. Hertig, 2021: Modeling and projecting health-relevant combined ozone and temperature events in present and future central European climate. *Air Qual. Atmos. Health*, **14**, 563–580, https://doi.org/10.1007/s11869-020-00961-0.

Josse, B., P. Simon, and V.-H. Peuch, 2004: Radon global simulations with the multiscale chemistry and transport model MOCAGE. *Tellus*, **56B**, 339–356, https://doi.org/10.3402/tellusb.v56i4.16448.

Kaiser, J. F., 1966: Digital filters. *System Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, Eds., John Wiley and Sons, 218–285.

Kaminski, J. W., and Coauthors, 2008: GEM-AQ, an on-line global multiscale chemical weather modelling system: Model description and evaluation of gas phase chemistry processes. *Atmos. Chem. Phys.*, **8**, 3255–3281, https://doi.org/10.5194/acp-8-3255-2008.

Kleinert, F., L. H. Leufen, and M. G. Schultz, 2021: IntelliO3-ts v1.0: A neural network approach to predict near-surface

ozone concentrations in Germany. *Geosci. Model Dev.*, **14**, 1–25, https://doi.org/10.5194/gmd-14-1-2021.

——, ——, A. Lupascu, T. Butler, and M. G. Schultz, 2022: Representing chemical history in ozone time-series predictions—A model experiment study building on the MLAir (v1.5) deep learning framework. *Geosci. Model Dev.*, **15**, 8913–8930, https://doi.org/10.5194/gmd-15-8913-2022.

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

——, ——, and ——, 2022a: Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-hour average ozone prediction. *Environ. Data Sci.*, **1**, E10, https://doi.org/10.1017/eds.2022.9.

——, ——, F. Weichselbaum, V. Gramlich, and M. G. Schultz, 2022b: MLAir—A tool to enable fast and flexible machine learning on air data time series, version 2.3.0. GitLab, https://gitlab.jsc.fz-juelich.de/esde/machine-learning/mlair/.

Lim, B., S. Ö. Arik, N. Loeff, and T. Pfister, 2021: Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecasting*, **37**, 1748–1764, https://doi.org/10.1016/j.ijforecast.2021.03.012.

Ma, J., Z. Li, J. C. P. Cheng, Y. Ding, C. Lin, and Z. Xu, 2020: Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. *Sci. Total Environ.*, **705**, 135771, https://doi.org/10.1016/j.scitotenv.2019.135771.

Manders, A. M. M., E. van Meijgaard, A. C. Mues, R. Kranenburg, L. H. van Ulft, and M. Schaap, 2012: The impact of differences in large-scale circulation output from climate models on the regional modeling of ozone and PM. *Atmos. Chem. Phys.*, **12**, 9441–9458, https://doi.org/10.5194/acp-12-9441-2012.

Marécal, V., and Coauthors, 2015: A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. *Geosci. Model Dev.*, **8**, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015.

Memmesheimer, M., E. Friese, A. Ebel, H. Jakobs, H. Feldmann, C. Kessler, and G. Piekorz, 2004: Long-term simulations of particulate matter in Europe on different scales using sequential nesting of a regional model. *Int. J. Environ. Pollut.*, **22**, 108–132, https://doi.org/10.1504/IJEP.2004.005530.

Menut, L., and Coauthors, 2013: CHIMERE 2013: A model for regional atmospheric composition modelling. *Geosci. Model Dev.*, **6**, 981–1028, https://doi.org/10.5194/gmd-6-981-2013.

Mills, G., and Coauthors, 2018: Tropospheric ozone assessment report: Present-day tropospheric ozone distribution and trends relevant to vegetation. *Elementa*, **6**, 47, https://doi.org/10.1525/elementa.302.

Monks, P. S., and Coauthors, 2015: Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmos. Chem. Phys.*, **15**, 8889–8973, https://doi.org/10.5194/acp-15-8889-2015.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338, https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Otero, N., J. Sillmann, J. L. Schnell, H. W. Rust, and T. Butler, 2016: Synoptic and meteorological drivers of extreme ozone concentrations over Europe. *Environ. Res. Lett.*, **11**, 024005, https://doi.org/10.1088/1748-9326/11/2/024005.

——, and Coauthors, 2018: A multi-model comparison of meteorological drivers of surface ozone over Europe. *Atmos. Chem. Phys.*, **18**, 12 269–12 288, https://doi.org/10.5194/acp-18-12269-2018.

Peuch, V.-H., and Coauthors, 2022: The Copernicus Atmosphere Monitoring Service: From research to operations. *Bull. Amer. Meteor. Soc.*, **103**, E2650–E2668, https://doi.org/10.1175/BAMS-D-21-0314.1.

Robertson, L., J. Langner, and M. Engardt, 1999: An Eulerian limited-area atmospheric transport model. *J. Appl. Meteor.*, **38**, 190–210, https://doi.org/10.1175/1520-0450(1999)038<0190:AELAAT>2.0.CO;2.

Sayeed, A., Y. Choi, E. Eslami, Y. Lops, A. Roy, and J. Jung, 2020: 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.

Schaap, M., R. M. Timmermans, M. Roemer, G. Boersen, P. J. Builtjes, F. J. Sauter, G. J. Velders, and J. P. Beck, 2008: The LOTOS EUROS model: Description, validation and latest developments. *Int. J. Environ. Pollut.*, **32**, 270–290, https://doi.org/10.1504/IJEP.2008.017106.

Schultz, M. G., and Coauthors, 2017: Tropospheric ozone assessment report: Database and metrics data of global surface ozone observations. *Elementa*, **5**, 58, https://doi.org/10.1525/elementa.244.

——, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philos. Trans. Roy. Soc.*, **379**, 20200097, https://doi.org/10.1098/rsta.2020.0097.

Seltzer, K. M., D. T. Shindell, P. Kasibhatla, and C. S. Malley, 2020: Magnitude, trends, and impacts of ambient long-term ozone exposure in the United States from 2000 to 2015. *Atmos. Chem. Phys.*, **20**, 1757–1775, https://doi.org/10.5194/acp-20-1757-2020.

Simpson, D., and Coauthors, 2012: The EMEP MSC-W chemical transport model—Technical description. *Atmos. Chem. Phys.*, **12**, 7825–7865, https://doi.org/10.5194/acp-12-7825-2012.

Sofiev, M., M. Galperin, and E. Genikhovich, 2008: A construction and evaluation of Eulerian dynamic core for the air quality and emergency modelling system SILAM. *Air Pollution Modeling and Its Application XIX*, C. Borrego and A. I. Miranda, Eds., NATO Science for Peace and Security Series, Series C: Environmental Security, Springer, 699–701.

Solberg, S., A. Colette, and C. Guerreiro, 2016: Discounting the impact of meteorology to the ozone concentration trends. ETC/ACM Tech. Paper 2015/9, 34 pp.

Stock, Z. S., M. R. Russo, and J. A. Pyle, 2014: Representing ozone extremes in European megacities: The importance of resolution in a global chemistry climate model. *Atmos. Chem. Phys.*, **14**, 3899–3912, https://doi.org/10.5194/acp-14-3899-2014.

TOAR Data Team, 2023: TOAR data infrastructure. TOAR Data Team, accessed 18 January 2023, https://toar-data.fz-juelich.de/.

U.S. EPA, 2013: Final report: Integrated science assessment for ozone and related photochemical oxidants (final report, Feb 2013). U.S. Environmental Protection Agency Rep. EPA 600/R-10/076F, 1251 pp., https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=247492.

——, 2020: Integrated Science Assessment (ISA) for ozone and related photochemical oxidants (final report, April 2020). U.S. Environmental Protection Agency Rep. EPA/600/R-20/012, 1468 pp., https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=348522.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017: Attention is all

you need. *31st Conf. on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, ACM, 5998–6008, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vautard, R., and Coauthors, 2012: Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations. *Atmos. Environ.*, **53**, 15–37, https://doi.org/10.1016/j.atmosenv.2011.10.065.

Wang, S. W., H. Levy, G. Li, and H. Rabitz, 1999: Fully equivalent operational models for atmospheric chemical kinetics within global chemistry-transport models. *J. Geophys. Res.*, **104**, 30 417–30 426, https://doi.org/10.1029/1999JD900830.

Weng, X., G. L. Forster, and P. Nowack, 2022: A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019. *Atmos. Chem. Phys.*, **22**, 8385–8402, https://doi.org/10.5194/acp-22-8385-2022.

WHO, 2013: Review of evidence on health aspects of air pollution – REVIHAAP Project: Technical report. Tech. Doc. WHO/EURO: 2013-4101-43860-61757, 309 pp., https://apps.who.int/iris/bitstream/handle/10665/341712/WHO-EURO-2013-4101-43860-61757-eng.pdf?sequence=1&isAllowed=y.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.

Young, P. J., and Coauthors, 2018: Tropospheric ozone assessment report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elementa*, **6**, 10, https://doi.org/10.1525/elementa.265.

Zheng, Y., Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, 2014: Time series classification using multi-channels deep convolutional neural networks. *Web-Age Information Management*, D. Hutchison et al., Eds., Lecture Notes in Computer Science, Vol. 8485, Springer, 298–310.