Environmental Data Science (2022), 1: e10, 1–31
doi:10.1017/eds.2022.9

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-hour average ozone prediction

Lukas Hubert Leufen[1,2,*], Felix Kleinert[1,2] and Martin G. Schultz[1]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[2]Institute of Geosciences, University of Bonn, Bonn, Germany
*Corresponding author. Email: l.leufen@fz-juelich.de

Received: 15 November 2021; **Revised:** 05 May 2022; **Accepted:** 31 May 2022

Keywords: Air quality; deep learning; machine learning; ozone; temporal decomposition; time series prediction

Abstract

Exposure to ground-level ozone is a concern for both humans and vegetation, so accurate prediction of ozone time series is of great importance. However, conventional as well as emerging methods have deficiencies in predicting time series when a superposition of differently pronounced oscillations on various time scales is present. In this paper, we propose a meteorologically motivated filtering method of time series data, which can separate oscillation patterns, in combination with different multibranch neural networks. To avoid phase shifts introduced by using a causal filter, we combine past observation data with a climatological estimate about the future to be able to apply a noncausal filter in a forecast setting. In addition, the forecast in the form of the expected climatology provides some a priori information that can support the neural network to focus not merely on learning a climatological statistic. We apply this method to hourly data obtained from over 50 different monitoring stations in northern Germany situated in rural or suburban surroundings to generate a prediction for the daily maximum 8-hr average values of ground-level ozone 4 days into the future. The data preprocessing with time filters enables simpler neural networks such as fully connected networks as well as more sophisticated approaches such as convolutional and recurrent neural networks to better recognize long-term and short-term oscillation patterns like the seasonal cycle and thus leads to an improvement in the forecast skill, especially for a lead time of more than 48 hr, compared to persistence, climatological reference, and other reference models.

Impact Statement

Exposure to ground-level ozone harms humans and vegetation, but the prediction of ozone time series, especially by machine learning, encounters problems due to the superposition of different oscillation patterns from long-term to short-term scales. Decomposing the input time series into long-term and short-term signals with the help of climatology and statistical filtering techniques can improve the prediction of various neural network architectures due to an improved recognition of different temporal patterns. More reliable and accurate forecasts support decision-makers and individuals in taking timely and necessary countermeasures to air pollution episodes.

This research article was awarded Open Data and OpenMaterials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

1. Introduction

Human health and vegetation growth are impaired by ground-level ozone (REVIHAAP, 2013; US EPA, 2013; Monks et al., 2015; Maas and Grennfelt, 2016; Fleming et al., 2018). High short-term ozone exposures cause worsening of symptoms, a need for stronger medication, and an increase in emergency hospital admissions, for people with asthma or chronic obstructive pulmonary diseases in particular (US EPA, 2020). More broadly, ozone exposure also increases susceptibility to respiratory diseases such as pneumonia in general, which in turn leads to an increased likelihood of hospitalization (US EPA, 2020). Findings of Di et al. (2017) further support earlier research that short-term exposure to ozone, even below regulatory limits, is highly likely to increase the risk of premature death, particularly for the elderly. Since the 1990s, there have been major changes in the global distribution of anthropogenic emissions (Richter et al., 2005; Granier et al., 2011; Russell et al., 2012; Hilboll et al., 2013; Zhang et al., 2016), which in turn has an influence on the ozone concentrations. Although reductions in peak concentrations have been achieved (Simon et al., 2015; Lefohn et al., 2017; Fleming et al., 2018), the negative effects of ground-level ozone remain (Cohen et al., 2017; Seltzer et al., 2017; Zhang et al., 2018; Shindell et al., 2019). Recent studies show that within the European Union, for example, ozone has the greatest impact on highly industrialized countries such as Germany, France, or Spain (Ortiz and Guerreiro, 2020). For all these reasons, it is therefore of utmost importance to be able to predict ozone as accurately as possible in the short term.

In light of these impacts, it is desirable to accurately forecast ozone concentrations for a couple of days so that protection measures can be initiated in time. Chemical transport models (CTMs), which explicitly solve the underlying chemical and physical equations, are commonly used to predict ozone (e.g., Collins et al., 1997; Wang et al., 1998a, 1998b; Horowitz et al., 2003; von Kuhlmann et al., 2003; Grell et al., 2005; Donner et al., 2011). Even though CTMs are equipped with the most up-to-date knowledge of research, the resulting estimates for exposure to and impacts of ozone may vary enormously between different CTM studies (Seltzer et al., 2020). Since CTMs operate on a computational grid and are thus always dependent on simplification of processes, parameterizations, and further assumptions, CTMs are themselves affected by large uncertainties (Manders et al., 2012). The deviations in the output of CTMs result accordingly from chemical and physical processes, fluxes such as emissions or deposition, as well as meteorological phenomena (Vautard et al., 2012; Bessagnet et al., 2016; Young et al., 2018). Finally, in order to use the predictions of the CTMs at the level of measuring stations, either model output statistics have to be applied (Fuentes and Raftery, 2005) or statistical methods are required (Lou Thompson et al., 2001).

In addition to simpler methods such as multilinear regressions, statistical methods that can map the relationship between time and observations in the time series are also suitable for this purpose. In general, time series can be characterized by the fact that values that are close in time tend to be similar or correlated (Wilks, 2006) and that the temporal ordering of these values forms an essential property of the time series (Bagnall et al., 2017). Autoregressive models (ARs) use this relationship and calculate the next value of a series x_{i+1} as a function ϕ of past values $x_i, x_{i-1}, \ldots, x_{i-n}$ where ϕ is simply a linear regression. Autoregressive moving average models extend this approach by additionally considering the error of past values, which is not described by the AR model. In the case of nonstationary time series, autoregressive integrated moving average models are used. However, these approaches are mostly limited to univariate problems and can only represent linear relationships (Shih et al., 2019). Alternative developments of nonlinear statistical models, such as Monte Carlo simulations or bootstrapping methods, have therefore been used for nonlinear predictions (De Gooijer and Hyndman, 2006).

In times of high availability of large data and increasingly efficient computing systems, machine learning (ML) has become an excellent alternative to classical statistical methods (Reichstein et al., 2019). ML is a generic term for data-driven algorithms like decision trees, random forests, or neural networks (NNs), which usually determine their parameters in a data-hungry and time-consuming learning process and can then be applied to new data at relatively low cost in terms of time and computational effort.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Fully connected networks (FCNs) are the pioneers of NNs and were already successfully applied around the turn of the millennium, for example, for the prediction of meteorological and air quality problems (Comrie, 1997; Gardner and Dorling, 1999; Elkamel et al., 2001; Kolehmainen et al., 2001). Simply put, FCNs extend the classical method of multilinear regression by adding the properties of nonlinearity as well as learning of knowledge. From a theoretical point of view, a sufficiently large network can be assumed to be a universal approximator of any function (Hornik et al., 1989). Nevertheless, it also shows that the application of FCNs is limited because they ignore the topology of the inputs (LeCun et al., 1999). In terms of time series, this means that FCNs will not be able to understand the abstract concept of unidirectional time.

These shortcomings have been overcome to some extent by deep learning (DL). In general, any NN that has a more sophisticated architecture or is based on more than three layers is classified as a deep NN. As Schultz et al. (2021) describe, the history of DL has been marked by highs and lows, as both computational cost and the size of datasets have always been tough adversaries. Since the 2010s, DL's more recent advances can be attributed to three main points: First, the acquisition of new knowledge has been drastically accelerated by massive parallel computation using graphics processing units. Second, so-called convolutional neural networks (CNNs; LeCun et al., 1999) became popular, whose strength lies in their ability to contextualize individual data points better than previous neural networks by sharing weights within the network and thus learning more information while maintaining the same network size. Finally, due to ever-increasing digitization, more and more data are available in ever-improving quality. Since DL methods are purely data-based compared to classical statistics, greater knowledge can be built up within a neural network simply through the greater availability of data.

Various newer NN architectures have been developed and also applied to time series forecasting in recent years. In this study, we focus on CNNs and recurrent neural networks (RNNs) as competitors to an FCN. For the prediction of time series, CNNs offer an advantage over FCNs due to their ability to better map relationships between neighboring data points. In Earth sciences, time series are typically multivariate, since a single time series is rarely considered in isolation, but always in interaction with other variables. However, multivariate time series should not be treated straightforwardly as two-dimensional images, since a causal relationship between different time series does not necessarily exist at all times and a different order of these time series would influence the result. Multivariate time series are therefore better to be understood as a composite of different one-dimensional data series (Zheng et al., 2014). Following this fact, multivariate time series can best be considered as a one-dimensional picture with different color channels. To extract temporal information with a CNN, so-called inception blocks (Szegedy et al., 2015) are frequently used, as, for example, in Fawaz et al. (2020) and Kleinert et al. (2021). These blocks consist of individual convolutional elements with different filter sizes that are applied in parallel and are intended to learn features with different temporal localities.

RNNs offer the possibility to model nonlinear behavior in temporal sequences in a nonparametric way. Frequently used RNNs are long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and gated recurrent unit networks (Chung et al., 2014) or hybrids of RNNs and CNNs such as in Liang and Hu (2015) and Keren and Schuller (2016). RNNs find intensive application in natural language processing, speech recognition, and signal processing, although they appear to have been largely replaced by transformer architectures more recently. However, these applications are mostly analysis problems and not predictive tasks. For time series prediction, especially for the prediction of multiple time steps into the future, there is little research evaluating the predictive performance of RNNs (Chandra et al., 2021). Moreover, Zhao et al. (2020) question the term *long* in LSTMs, as their research shows that LTSMs do not have long-term memory from a purely statistical point of view because their behavior hardly differs from that of a first-order AR. Furthermore, Cho et al. (2014) were able to show that these network types, for example, have difficulties in reflecting an annual pattern in daily-resolved data. Thus, the superposition of different periodic patterns remains a critical issue in time series prediction, as RNNs have fundamental

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

difficulties with extrapolating and predicting periodic signals (Ziyin et al., 2020) and therefore tend to focus on short-term signals only (Shih et al., 2019).

In order to deal with the superposition of different periodic signals and thus help the learning process of the NN, digital filters can be used. So-called finite impulse response (FIR) filters are realized by convolution of the time series with a window function (Oppenheim and Schafer, 1975). In fact, FIR filters are widely used in meteorology without being labeled as such, since a moving average is nothing more than a convolution with a rectangular window function. With the help of such FIR filters, it is possible to extract or remove a long-term signal from a time series or to directly divide the time series into several components with different frequency ranges, as applied, for example, in Rao and Zurbenko (1994), Wise and Comrie (2005), and Kang et al. (2013). In these studies, so-called Kolmogorov–Zurbenko (KZ) filters (Žurbenko, 1986) are used, which were specially developed for use in meteorology and promise a good separation between long-term and short-term variations of meteorological and air quality time series (Rao and Zurbenko, 1994).

There are examples of the use of filters in combination with NNs, for example, in Cui et al. (2016) and Jiang et al. (2019), but these are limited purely to analysis problems. The application of filters in a predictive setting is more complicated, because, for a prediction, filters may only be applied causally to past values, which inevitably produces a phase shift and thus a delay in the filtered signal (Oppenheim and Schafer, 1975). The lower the chosen cutoff frequency of the low-pass filter, for example, to extract the seasonal cycle, the more the resulting signal becomes delayed. This in turn leads to the fact that values in the recent past cannot be separated, as no information is yet available on the long-term components.

In this work, we propose an alternative way to filter the input time series using a composite of observations and climatological statistics to be able to separate long-term and short-term signals with the smallest possible delay. By dividing the input variables into different frequency ranges, different NN architectures are able to improve their understanding of both short-term and long-term patterns.

This paper is structured as follows: First, in Section 2, we explain and formalize the decomposition of the input time series and give details about the NN architecture used. Then, in Section 3, we describe our conducted experiments in detail, describing the data used, their preparation, the training setup, and the evaluation procedures. This is followed by the results in Section 4. Finally, we discuss our results in Section 5 and draw our conclusions in Section 6.

2. Methodology

In this paper, we combine actual observation data and a meteorologically and statistically motivated estimate of the future to overcome the issue of delay and causality (see Section 1). The estimate about the future is composed of climatological information about the seasonal as well as diurnal cycle, whereby the latter is also allowed to vary over the year. For each observation point t_0, these two time series, the observation for time steps with $t_i \leq t_0$ and the statistical estimation for $t_i > t_0$, are concatenated. By doing this, noncausal filters can be applied to the composite time series in order to separate the oscillation components of the time series such as the dominant seasonal and diurnal cycle.

The decomposition of the time series is obtained by the iterative application of several low-pass filters with different cutoff frequencies. The signal resulting from a first filter run, which only has frequencies below a given cutoff frequency, is then subtracted from the original composite signal. The next filter iteration with a higher cutoff frequency then starts on this residual, the result of which is again subtracted. By applying this cycle several times, a time series with the long-term components, multiple series covering certain frequency ranges, and a last residual time series containing all remaining short-term components are generated. Here, we test filter combinations with four and two frequency bands. The exact cycle of filtering is described in Section 2.1.

Each filtered component is finally used as an input branch of a so-called multibranch NN (MB-NN), which first processes the information of each input branch separately and then combines it in a subsequent layer. In Section 2.2, we go into more detail about the architecture of the MB-NN.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

2.1. Time series filter

For each time step t_0, a composite time series $\breve{x}_i^{(0)}$,

$$\breve{x}_i^{(0)}(t_0) = \begin{cases} x_i^{(0)}, & t_i \le t_0, \\ a_i^{(0)}, & t_i > t_0, \end{cases} \tag{1}$$

can be created that is composed of the true observation $x_i^{(0)}$ for past time steps and a climatological estimate

$$a_i^{(0)} = \overline{x}_{\text{month}}^{(0)}(t_i) + \Delta_{\text{hour}}^{(0)}(t_i) \tag{2}$$

for future values. The composite time series $\breve{x}_i^{(0)}$ is always a function of the current observation time t_0. The climatological estimate is derived from a monthly mean value $\overline{x}_{\text{month}}^{(0)}(t_i)$ with

$$\overline{x}_{\text{month}}^{(0)}(t_i) = f^{(0)}\left(x_i^{(0)}\right) \tag{3}$$

and a daily anomaly $\Delta_{\text{hour}}^{(0)}(t_i)$ of it with

$$\Delta_{\text{hour}}^{(0)}(t_i) = g^{(0)}\left(x_i^{(0)} - \overline{x}_{\text{month}}^{(0)}(t_i)\right) \tag{4}$$

that may vary over the year. $f^{(0)}$ and $g^{(0)}$ are arbitrary functions used to calculate these estimates. The composite time series $\breve{x}_i^{(0)}(t_0)$ can then be convolved with an FIR filter with given properties $b_i^{(0)}$. The result of this convolution is a low-pass filtered time series:

$$\tilde{x}_n^{(0)}(t_0) = \sum_{i=t_0-N/2}^{t_0+N/2} b_i^{(0)} \cdot \breve{x}_{n-i}^{(0)}(t_0). \tag{5}$$

It should be noted again that $\tilde{x}_i^{(0)}$ is still a function of the current observation time t_0. From the composite time series and its filtered result, a residual

$$x_i^{(1)}(t_0) = x_i^{(0)} - \tilde{x}_i^{(0)}(t_0) \tag{6}$$

can be calculated, which represents the equivalent high-pass signal.

A new filtering step can now be applied to the residual $x_i^{(1)}(t_0)$. For this, the a priori information, which is used to estimate the future, is first newly calculated. Ideally, if the first filter application in equation (5) has already completely removed the seasonal cycle, the climatological mean $\overline{x}_{\text{month}}^{(1)}(t_i)$ is zero, and based on our assumption in equation (2), only an estimate of the hourly daily anomaly $\Delta_{\text{hour}}^{(1)}(t_i)$ remains. With this information, a composite time series $\breve{x}_i^{(1)}(t_0)$ can now be formed, which can separate higher frequency oscillation components using another low-pass filter with a higher cutoff frequency. A time series $\tilde{x}_n^{(1)}(t_0)$ created in this way corresponds to the application of a band-pass filter. On the residual $x_i^{(2)}(t_0)$, the next filter iteration with corresponding a priori information can be carried out. Generalized, equations (1)–(6) result in

$$\overline{x}_{\text{month}}^{(j)}(t_i) = f^{(j)}\left(x_i^{(j)}\right), \tag{7}$$

$$\Delta_{\text{hour}}^{(j)}(t_i) = g^{(j)}\left(x_i^{(j)} - \overline{x}_{\text{month}}^{(j)}(t_i)\right), \tag{8}$$

$$a_i^{(j)} = \overline{x}_{\text{month}}^{(j)}(t_i) + \Delta_{\text{hour}}^{(j)}(t_i), \tag{9}$$

$$\breve{x}_i^{(j)}(t_0) = \begin{cases} x_i^{(j)}, & t_i \le t_0, \\ a_i^{(j)}, & t_i > t_0, \end{cases} \tag{10}$$

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

$$\tilde{x}_n(j)(t_0) = \sum_{i=t_0-N/2}^{t_0+N/2} b_i^{(j)} \cdot \breve{x}_{n-i}^{(j)}(t_0), \tag{11}$$

$$x_i^{(j+1)}(t_0) = x_i^{(j)} - \tilde{x}_i^{(j)}(t_0). \tag{12}$$

If a time series was decomposed according to this procedure using equations (7)–(12) with J filters, it now consists of a component $\tilde{x}^{(0)}$, that contains all low-frequency components, $J-1$ components $\tilde{x}^{(j)}$ with oscillations on different frequency intervals, and a residual term $x^{(J)}$ that only covers the high-frequency components. The original signal can be completely reconstructed at any time t_i by summing up the individual components.

In this study, oscillation patterns that have a periodicity of months or years are separated from the series in the first filter iteration by using a cutoff period of 21 days, which is motivated by the work of Kang et al. (2013). We also consider a cutoff period of around 75 days, as used, for example, in Rao et al. (1997) and Wise and Comrie (2005), and evaluate the impact of this low-frequency cutoff. For the further decomposition of the time series, we first follow the cutoff frequencies proposed in Kang et al. (2013) and divide the time series into the four components baseline (BL, period >21 days), synoptic (SY, period >2.7 days), diurnal (DU, period >11 hr), and intraday (ID, residuum). Since Kang et al. (2013) found that a clear separation of the individual components is not possible for the short-term components, but can be achieved between the long-term and short-term components, we conduct a second series of experiments in which the input data are only divided into long term (LT, period >21 days) and short term (ST, residuum).

Figure 1 shows the result of such a decomposition into four components. It can be seen that the BL component decreases with time. The SY component fluctuates around zero with a moderate oscillation between August 16 and 20. In the DU component, the day-to-day variability and diurnal oscillation patterns are visible, and in the ID series, several positive and negative peaks are apparent. Overall, it can be seen that the climatological statistical estimation of the future provides a reliable prediction. However, since a slightly higher ozone episode from August 25 onward cannot be covered by the climatology, the long-term component BL is slightly underestimated, but for time points up to t_0, this has hardly any effect. This small difference of a few parts per billion (ppb) is covered by the SY and DU components, so that the residual component ID no longer contains any deviations.

2.2. Multibranch NN

The time series divided into individual components according to Section 2.1 serves as the input of an MB-NN. In this work, we investigate three different types of MB-NNs based on fully connected, convolutional, or recurrent layers. We therefore refer to the corresponding NNs in the following as MB-FCN, MB-CNN, and MB-RNN. The respective filter components of all input variables are presented together to one branch each. Thereby, each filter component leads to a distinct input branch in the NN. A branch first learns the local characteristics of the oscillation patterns and can therefore also be understood as its own subnetwork. Afterward, the MB-NN can learn global links, that is, the interaction of the different scales, by a learned (nonlinear) combination of the individual branches in a subsequent network. However, the individual branches are not trained separately, but the error signal propagates from the very last layer backward through the entire network and then splits up between the individual branches.

The sample MB-NN shown on the left in Figure 2 consists of four input branches, each receiving a component from the long-term $\tilde{x}^{(0)}$ (BL) to the residual $x^{(3)}$ (ID) of the filter decomposition. Here, the data presented as example input are the same as in Figure 1, but each component has already been scaled to a mean of zero and a standard deviation of 1, taking into account several years of data. In addition to the characteristics of the example already discussed in Section 2.1, it can be seen from the scaling that the BL component is above the mean, indicating a slightly increased long-term ozone concentration. The SY component, on the other hand, shows only a weak fluctuation. The data are fed into four different branches, each of which consists of an arbitrary architecture based on fully connected, convolutional, or

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

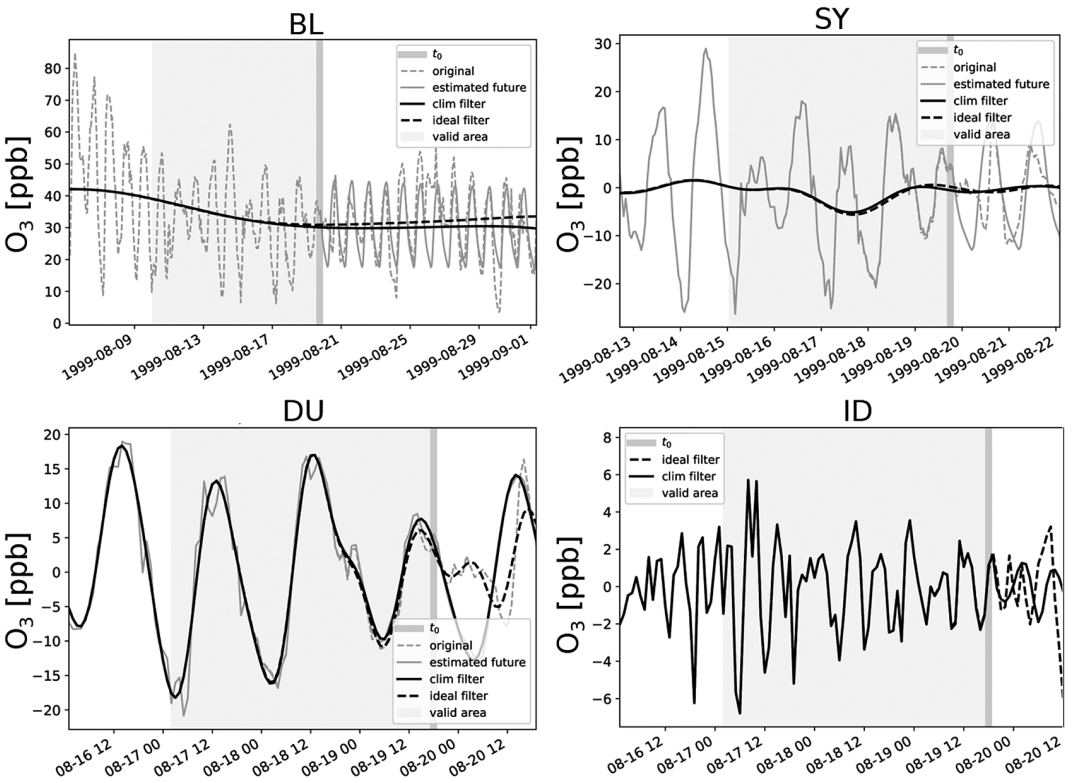

Figure 1. *Decomposition of an ozone time series into baseline (BL), synoptic (SY), diurnal (DU), and intraday (ID) components at t_0 = August 19, 1999 (dark gray background) at an arbitrary sample site (here DEMV004). Shown are the true observations $x_i^{(j)}$ (dashed light gray), the a priori estimation $a_i^{(j)}$ about the future (solid light gray), the filtering of the time series composed of observation and a priori information $x^{(j)}(t_0)$ (solid black), and the response of a noncausal filter with access to future values (dashed black) as a reference for a perfect filtering. Because of boundary effects, only values inside the marked area (light gray background) are valid.*

recurrent layers. Subsequently, the information of these four subnetworks is concatenated and parsed in the tail to a concluding neural block, which finally results in the output layer.

On the right side in Figure 2, only the decomposition into the two components LT and ST is applied. Since the cutoff frequency is the same for LT and BL, the LT input is equal to the BL input. All short-term components are combined and fed to the NN in the form of the ST component. This arbitrary MB-NN again uses a specified type of neural layers in each branch before the information is interconnected in the concatenate layer and then processed in the subsequent neural block, which finally leads to the output again. Figure 2 shows a generic view of the four-branch and two-branch NNs. The specific architectures employed in this study are depicted in Section 3.3, Tables B2 and B3 in Appendix B, and Figures D1–D5 in Appendix D.

3. Experiment Setup

For data preprocessing and model training and evaluation, we employ the software MLAir (version 2.0.0; Leufen et al., 2022). MLAir is a tool written in Python that was developed especially for the application of ML to meteorological time series. The program executes a complete life cycle of an ML training, from preprocessing to training and evaluation. A detailed description of MLAir can be found in Leufen et al. (2021).

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

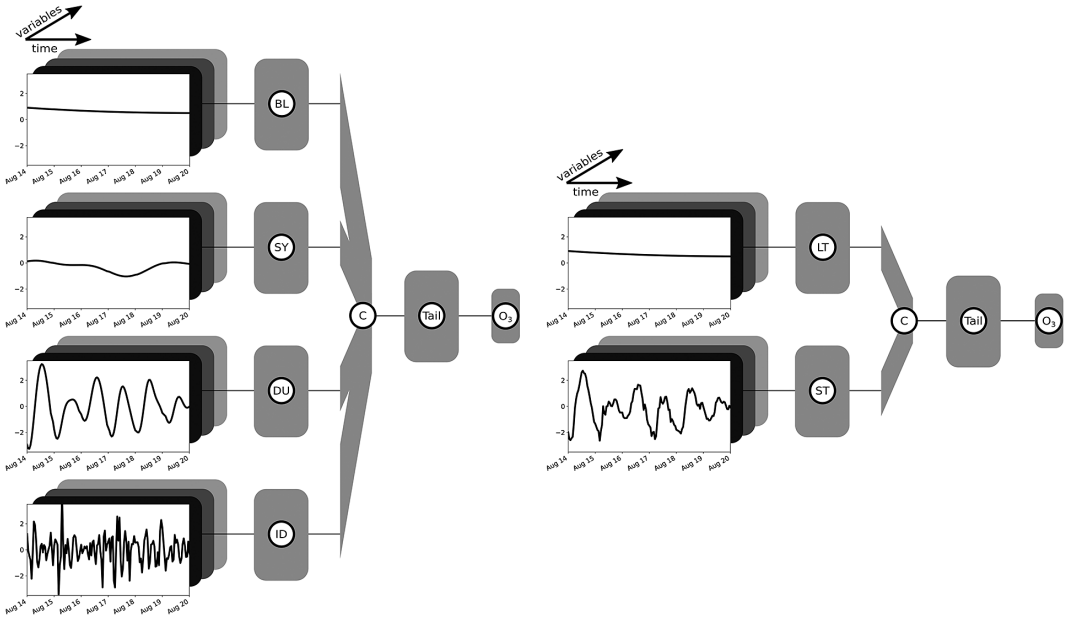

Figure 2. *Sketching of two arbitrary MB-NNs with inputs divided into four components (BL, SY, DU, and ID) on the left and two components (LT and ST) on the right. The input example shown here corresponds to the data shown in Figure 1, whereby the components SY, DU, and ID on the right-hand side have not been decomposed, but rather grouped together as the short-term component ST. Moreover, the data have already been scaled. Each input component of a branch consists of several variables, indicated schematically by the boxes in different shades of gray. The boxes identified by the branch name, also in gray, each represent an independent neuronal block with user-defined layer types such as fully connected, convolutional, or recurrent layers and any number of layers. Subsequently, the branches are then combined via a concatenation layer marked as "C." This is followed by a final neural block labeled as "Tail," which can also have any configuration and finally ends in the output layer of the NN indicated by the tag "O_3." The sketches are based on a visualization with the Net2Vis tool (Bauerle et al., 2021).*

3.1. Data

In this study, data from the Tropospheric Ozone Assessment Report database (TOAR DB; Schultz et al., 2017) are used. This database collects global in situ observation data with a special focus on air quality and in particular on ozone. As part of the Tropospheric Ozone Assessment Report (TOAR, 2021), over 10,000 air quality measuring stations worldwide were inserted into the database. For the area over Central Europe, these observations are supplemented by model reanalysis data interpolated to the measuring stations, which originate from the consortium for small-scale modeling (COSMO) reanalysis with 6-km horizontal resolution (COSMO-REA6; Bollmeyer et al., 2015). The measured data provided by the German Environment Agency (Umweltbundesamt) are available in hourly resolution.

By following Kleinert et al. (2021), we choose a set of nine input variables. As regards chemistry, we use the observation of O_3 as well as the measured values of NO and NO_2, which are important precursors for ozone formation. In this context, it would be desirable to include other chemical variables and especially volatile organic compounds (VOCs), such as isoprene and acetaldehyde, which have a crucial influence on the ozone production regime (Kumar and Sinha, 2021). However, the measurement coverage of VOCs is very low, so that only very sporadic recordings are available, which would result in a rather small dataset. Concerning meteorology, in addition to the wind in its individual components as well as the height of the planetary boundary layer as an indicator for advection and mixing, we use temperature and the cloud cover as a proxy for solar irradiance, and the relative humidity. All meteorological variables are

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Table 1. Input and target variables with respective temporal resolution and origin. Data labeled with UBA originate from measurement sites provided by the German Environment Agency, and data with flag COSMO-REA6 have been taken from reanalysis.

	Variable	Origin	Temporal resolution
Input	NO	UBA	1 hr
	NO_2	UBA	1 hr
	O_3	UBA	1 hr
	Cloud cover	COSMO-REA6	1 hr
	Planetary boundary layer height	COSMO-REA6	1 hr
	Relative humidity	COSMO-REA6	1 hr
	Temperature	COSMO-REA6	1 hr
	Wind's u-component	COSMO-REA6	1 hr
	Wind's v-component	COSMO-REA6	1 hr
Target	dma8eu O_3	UBA	1 day

Abbreviations: COSMO-REA6, consortium for small-scale modeling reanalysis with 6-km horizontal resolution; UBA, Umweltbundesamt.

extracted from COSMO-REA6, but are treated in the following as if they were observations at the measuring stations. Table 1 provides an overview of the observation and model variables used. The target variable ozone is also obtained directly from the TOAR DB. Rather than using the hourly values, however, the daily aggregation to the daily maximum 8-hr average value according to the European Union definition (dma8eu) is performed by the TOAR DB and extracted directly in daily resolution. It is important to note that the calculation of dma8eu includes observations from 5 p.m. of the previous day (cf. European Parliament and Council of the European Union, 2008). Care must therefore be taken that ozone values from 5 p.m. on the day of t_0 may no longer be used as inputs to ensure a clear separation, as they are already included in the calculation of the target value.

This study is based on a relatively homogeneous dataset, so that the NNs can learn better and thus the effect due to time series filtering becomes clearer. In order to obtain such a dataset of observations, we restrict our investigations to the area of the North German Plain, which includes all areas in Germany north of 52.5°N. We choose this area because of the rather flat terrain; no station is located higher than 150 m above sea level. In addition, we restrict ourselves to measurement stations that are classified as background according to the European Environmental Agency AirBase classification (European Parliament and Council of the European Union, 2008), which means that no industry or major road is located in the direct proximity of the stations and consequently the pollution level of this station is not dominated by a single source. All stations are located in a rural or suburban environment. These restrictions result in a total number of 55 stations distributed over the entire area of the North German Plain. A geographical overview can be found in Figure A2 in Appendix A. It should be noted that no measuring station provides complete time series, so that gaps within the data occur. However, since the filter approach requires continuous data, gaps of up to 24 consecutive hours on the input side and gaps of 2 days on the target side are filled by linear interpolation along time.

3.2. Preparing of input and target data

The entire dataset is split along the temporal axis into training, validation, and test data. For this purpose, all data in the period from January 1, 1997 to December 31, 2007 are used for training. The a priori information of the time series filter about seasonal and diurnal cycles is calculated based on this set. The following 2 years, January 1, 2008 to December 31, 2009, are used for the validation of the training, and

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Table 2. Number of measurement stations and resulting number of samples used in this study. All stations are classified as background and situated either in a rural or suburban surrounding in the area of the North German Plain. Data are split along the temporal axes into three subsequent blocks for training, validation, and testing.

	Training	Validation	Testing
Stations			
Rural	31	17	17
Suburban	24	15	13
Total	55	32	30
Samples			
Rural	54,544	10,927	16,858
Suburban	40,968	10,405	13,622
Total	95,512	21,332	30,480

all data from January 1, 2010 onward are used for the final evaluation and testing of the trained model. For the meteorological data, there are no updates in the TOAR DB since January 1, 2015, so more recent air quality measurements cannot be used in this study.

For each time step t_0, the time series is decomposed using the filter approach as defined in Section 2.1. The a priori information is obtained from the training dataset alone so that validation and test datasets remain truly independent. Afterward, the input variables are standardized so that each filter component of each variable has a mean of zero and a standard deviation of 1 (Z-score normalization). For the target variable dma8eu ozone, we choose the Z-score normalization as well. All transformation properties for both inputs and targets are calculated exclusively on the training data and applied to the remaining subsets. Moreover, these properties are not determined individually per station, but jointly across all measuring stations.

In this work, we choose the number of past time steps for the input data as 65 hr. This corresponds to the three preceding days minus the measurements starting at 5 p.m. on the current day of t_0 due to the calculation procedure of dma8eu as already mentioned. The number of time steps to be predicted is set to the next 4 days for the target. All in all, we use almost 100,000 training samples and 20,000 and 30,000 samples for validation and testing, respectively (see Table 2 for exact numbers). The data availability at individual stations, as well as the total number of different stations at each point in time, is shown in Figures A1 and A3 in Appendix A, respectively. The visible larger data gaps are caused by a series of missing values that exceed the maximum interpolation length.

3.3. Training setup and hyperparameter search

First, we search for an optimal decomposition of the input time series for the NNs by optimizing the hyperparameters for the MB-FCN. Second, we use the most suitable decomposition and train different MB-CNNs and MB-RNNs on these data. Finally, we train equivalent network architectures without decomposition of the input time series to obtain a direct comparison of the decomposition approach as outlined in Section 3.4. All experiments are assessed based on the mean square error (MSE), as presented in Section 3.5. Since we are testing a variety of different models, we have summarized the most relevant abbreviations in Table 3.

The experiments to find an optimal decomposition of the inputs and best hyperparameters for the MB-FCN start with the same cutoff frequencies for decomposition as used in Kang et al. (2013), who divide their data into the four components BL, SY, DU, and ID, as explained in Section 2.1. Since there is generally no optimal a priori choice for a filter (Oppenheim and Schafer, 1975) and furthermore this is

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Table 3. Summary of model acronyms used in this study depending on their architecture and the number of input branches. The abbreviations for the branch types refer to the unfiltered original raw data and either to the temporal decomposition into the four components baseline (BL, period >21 days), synoptic (SY, period >2.7 days), diurnal (DU, period >11 hr), and intraday (ID, residuum), or to the decomposition into two components long term (LT, period >21 days) and short term (ST, residuum). When multiple input components are used, as indicated in the column labeled Count, the NNs are constructed with multiple input branches, each receiving a single component, and are therefore referred to as multibranch (MB). For technical reasons, this MB approach is not applicable to the OLS model, which instead uses a flattened version of the decomposed inputs and is therefore not specified as MB.

Input branches		Model name			
Branch type(s)	Count	FCN	CNN	RNN	OLS
Raw	1	FCN	CNN	RNN	OLS
LT and ST	2	MB-FCN-LT/ST	MB-CNN-LT/ST	MB-RNN-LT/ST	OLS-LT/ST
LT, ST, and raw	3	MB-FCN-LT/ST+raw	–	–	–
BL, SY, DU, and ID	4	MB-FCN-BL/SY/DU/ID	–	–	–
BL, SY, DU, ID, and raw	5	MB-FCN-BL/SY/DU/ID+raw	–	–	–

Abbreviations: CNN, convolutional neural network; FCN; fully connected network; OLS, least squares regression.

likely to vary from one application to another, we choose a Kaiser filter (Kaiser, 1966) with a beta parameter of $\beta = 5$ for the decomposition of the time series. We prefer this filter for practical considerations, as a filter with a Kaiser window features a sharper gain reduction in the transition area at the cutoff frequency in comparison to the KZ filter. Based on this, we test a large number of combinations of the hyperparameters (see Table B1 in Appendix B for details). The trained MB-FCN with the lowest MSE on the validation in this experiment is referred to as MB-FCN-BL/SY/DU/ID in the following. Since, as already mentioned in Section 2.1, a clear decomposition in individual components is not always possible, we start a second series of experiments in which the input data are only divided into long term (LT) and short term (ST). We tested cutoff periods of 75 (Rao et al., 1997; Wise and Comrie, 2005) and 21 days (Kang et al., 2013), and found no difference with respect to the MSE of the trained networks. Hence, we selected the cutoff period of 21 days and refer to the trained network as MB-FCN-LT/ST in the following. After finding an optimal set of hyperparameters for both experiments, we vary the input data and study the resulting effect on the prediction skill. In two extra experiments, we add an additional branch with the unfiltered raw data to the inputs. According to the previous labels, these experiments result in the NNs labeled MB-FCN-BL/SY/DU/ID+raw and MB-FCN-LT/ST+raw. We have summarized the optimal hyperparameters for each of the MB-FCN architectures in Table B2 in Appendix B.

Based on the findings with the MB-FCNs, we choose the best MB-FCN and the corresponding preprocessing and temporal decomposition of the input time series for the second part of the experiments, in which we test more sophisticated network architectures. With the data remaining the same, we investigate to what extent using MB-CNN or MB-RNN leads to an improvement compared to MB-FCN and also in relation to their counterparts without temporal decomposition (CNN and RNN). For this purpose, we test different architectures for CNN and RNN with and without temporal decomposition separately and compare the best representative found by the experiment, respectively. The optimal hyperparameters given by this experiments are outlined in Table B3 in Appendix B, and a visualization of the best NNs can be found in Figures D1–D5 in Appendix D. Regarding the CNN architecture, we varied the total number of layers and filters in each layer, the filter size, the use of pooling layers, as well as the application of convolutional blocks after the concatenate layer and the layout of the final dense layers. For the RNNs, during hyperparameter search, we used different numbers of LSTM cells per layer and tried stacked LSTM layers. Furthermore, we added recurrent layers after the concatenate layer in some experiments. In general, we tested different dropout rates, learning rates, a decay of the learning rate, and several activation functions.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

3.4. Reference forecasts

We compare the results of the trained NNs with a persistence forecast, which generally performs well on short-term predictions (Murphy, 1992; Wilks, 2006). The persistence consists of the last observation, in this case the value of dma8eu ozone on the day of t_0, which serves as a prediction for all future days. We also compare the results with climatological reference forecasts following Murphy (1988). Details are given in Section 3.5. Furthermore, we compare the MB-NNs to an ordinary least squares regression (OLS), an FCN, a CNN, and an RNN. The basis for these competitors is hourly data without special preparation, that is, without prior decomposition into the individual components. The parameters of the OLS are created analogously to the NNs on the training data only. For the FCN, CNN, and RNN, sets of optimal parameters were determined experimentally in preliminary experiments also on training data. Only the NNs with the lowest MSE on the validation data are shown here. Furthermore, as with MB-NNs, we apply an OLS method to the temporally decomposed input data. For technical reasons, the OLS approach is not able to work with branched data and therefore uses flattened inputs instead. Finally, we draw a comparison with the IntelliO3-ts model from Kleinert et al. (2021). IntelliO3-ts is a CNN based on inception blocks (see Section 1). In contrast to the study here, IntelliO3-ts was trained for the entire area of Germany. It should be noted that IntelliO3-ts is based on daily aggregated input data, whereas all NNs trained in this study use an hourly resolution of input data. For all models, the temporal resolution of the targets is daily, so that the NNs of this study have to deal with different temporal resolutions, which does not apply for IntelliO3-ts.

3.5. Evaluation methods

The evaluation of the NNs takes place exclusively on the test data that are unknown to the models. To assess the performance of the NNs, we examine both absolute and relative measures of accuracy. Accuracy measures generally represent the relationship between a prediction and the value to be predicted. Typically, for an absolute measure of the predictive quality on continuous values, the MSE is used. The MSE is a good choice as a measure because it takes into account the bias as well as the variances and the correlation between prediction and observation. To determine the uncertainty of the MSE, we choose a resampling test procedure (cf. Wilks, 2006). Due to the large amount of data, a bootstrap approach is suitable. Synthetic datasets are generated from the test data by repeated blockwise resampling with replacement. For each set, the error, in our case the MSE, is calculated. With a sufficiently large number of repetitions (here $n = 1,000$), we can access an estimate of the error uncertainty. To reduce misleading effects caused by autocorrelation, we divide the test data along the time axis into monthly blocks and draw from these instead of the individual values.

To compare individual models directly with each other, we derive a skill score from the MSE as a relative measure of accuracy. In this study, the skill score always consists of the MSE of the actual forecast as well as the MSE of the reference forecast and is given by

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{ref}}}. \tag{13}$$

Accordingly, a value around zero means that no improvement over a reference could be achieved. If the skill score is positive, an improvement can generally be assumed, and if it is negative, the prediction accuracy is below the reference.

For the climatological analysis of the NN, we refer to Murphy (1988), who determines the climatological quality of a model by breaking down the information into four cases. In Case 1, the forecast is compared with an annual mean calculated on data that are known to the model. For this study, we consider both the training and validation data to be internal data, since the NN used these data during training and hyperparameter search. Case 2 extends a climatological consideration by differentiating into 12 individual monthly averages. Cases 3 and 4, respectively, are the corresponding transfers of the aforementioned analyses, but on test data that are unknown to the model.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Another helpful method for the verification of predictions is the consideration of the joint distribution $p(y_i, o_j)$ of prediction y_i and observation o_j (Murphy and Winkler, 1987). The joint distribution can be factorized to shed light on particular facets. With the calibration-refinement factorization

$$p(y_i, o_j) = p(o_j|y_i) \cdot p(y_i), \tag{14}$$

the conditional probability $p(o_j|y_i)$ and the marginal distribution of the prediction $p(y_i)$ are considered. $p(o_j|y_i)$ provides information on the probability of each possible event o_j occurring when a value y_i is predicted, and thus how well the forecast is calibrated. $p(y_i)$ indicates the relative frequency of the predicted value. It is desirable to have a distribution of y with a width equal to that for o.

3.6. Feature importance analysis

Due to the NN's nonlinearity, the influence of individual inputs or variables on the model is not always directly obvious. Therefore, we again use a bootstrap approach to gain insight into the feature importance. In general, we remove a certain piece of information and examine the skill score in comparison to the original prediction to see whether the prediction quality of the NN decreases or increases as a result. If the skill score remains constant, this is an indication that the examined information does not provide any additional benefit for the NN. The more negative the skill score becomes in the feature importance analysis, the more likely it is that the examined variable contains important information for the NN. In the unlikely case of a positive skill score, it can be inferred that the context of this variable was learned incorrectly and thus disturbs the prediction.

For the feature importance analysis, we take a look at three different cases. First, we analyze the influence of the temporal decomposition by destroying the information of an entire input branch, for example, all low-frequent components (BL resp. LT). This yields information about the effect of the different time scales from long term to short term and the residuum. In the second step, we adopt a different perspective and look at complete variables with all temporal components (e.g., both LT and ST components of temperature). In the third step, we drive down one tier and consider each input separately to get information about whether a single input has a very strong influence on the prediction (e.g., BL component of NO_2).

To break down the information for the feature importance analysis, we randomly draw the quantity to be examined from its observations. Statistically, a test variable obtained in this way is sampled from the same distribution as the original variable. However, the test variable is detached from its temporal context as well as from the context of other variables. This procedure is repeated 100 times to reduce random effects.

The feature importance analysis considers only the influence of a single quantity and no pairwise or further correlations. However, the isolated approach already provides relevant information about the feature importance. It is important to note that this analysis can only show the importance of the inputs for the trained NN and that no physical or causal relationships can be deduced from this kind of analysis in general.

4. Results

Since a comparison of all models against each other would quickly become incomprehensible, we first look at the results of the resampling in order to obtain a ranking of MB-FCNs (see Table 3 for a summary of model acronyms). The results of the bootstrapping are shown in Figure 3 and listed also in Table C1 in Appendix C. With a block length of 1 month and 1,000 repetitions of the bootstrapping, it can be seen that the simple FCN cannot adequately represent the relationships between inputs and targets in comparison to the other models. Moreover, it is visible that the performance of the MB-FCN-BL/SY/DU/ID falls behind in comparison to the other MB-FCNs with an average MSE $> 70\,\text{ppb}^2$. The smallest resampling errors could be achieved with the models MB-FCN-BL/SY/DU/ID+raw, MB-FCN-LT/ST, and MB-FCN-LT/ST+raw.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

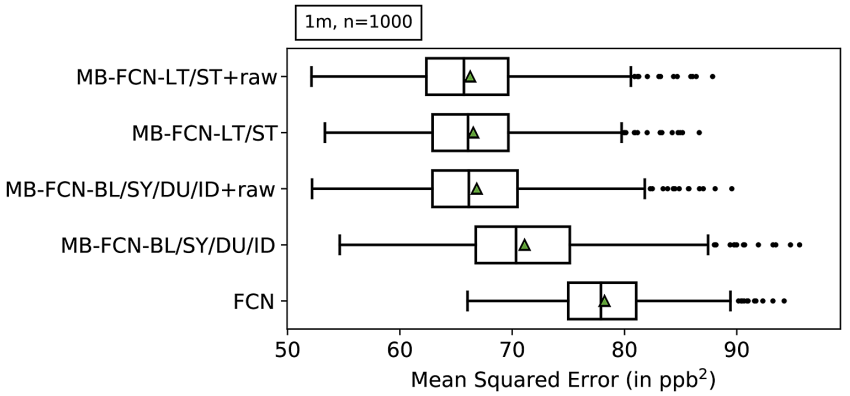

Figure 3. *Results of the uncertainty estimation of the MSE using a bootstrap approach represented as box-and-whiskers. For each model, the median is shown as a black vertical line, the mean as a green triangle, the upper and lower quartiles in the form of the box, the upper and lower whiskers, which correspond to 1.5 times the interquartile range, and outliers beyond the whiskers as individual data points. The models are ordered from top to bottom with ascending average MSE. A total of 1,000 bootstrap samples were created by resampling with the replacement of single-month blocks.*

When comparing the decomposition into BL/SY/DU/ID and the decomposition into LT/ST components, the latter decomposition tends to yield a lower error. Alternatively, it is possible to achieve comparative performance by adding the raw data to both variants of decomposition. For the LT/ST decomposition, however, this improvement is marginal.

Since the forecast accuracy of the top three NNs is nearly indistinguishable, especially for the two models with the LT/ST split, we choose the MB-FCN-LT/ST network and so the LT/ST decomposition for further analysis, since, of the three winning candidates, this is the network with the smallest number of trainable parameters (see Table B2 in Appendix B).

So far, we have shown the advantages of an LT/ST decomposition during preprocessing for FCNs. Therefore, in the following, we apply our proposed decomposition to more elaborated network architectures, namely a CNN and an RNN architecture. We again consider the uncertainty estimation of the MSE using the bootstrap approach and calculate the skill score with respect to the MSE in pairs for an NN type that was trained once as an MB-NN with temporally decomposed inputs and once with the raw hourly data. Similarly, we consider the skill score of OLS on decomposed and raw data, respectively. The results are shown in Figure 4. It can be seen that the skill score is always positive for all models. This in turn means that using our proposed time decomposition of the input time series improves all the models analyzed here. When looking at the individual models, it can be differentiated that the FCN architecture in particular benefits from the decomposition, whereas the improvement is smaller for RNN and smallest but still significant for OLS and CNN.

Based on the uncertainty estimation of the MSE shown in Figure 5 and also listed in Table C2 in Appendix C, the models can be roughly divided into three groups according to their average MSE. The last group consists solely of the persistence prediction, which delivers a significantly worse prediction than all other methods and lies at an MSE of 107 ppb^2 on average. In the intermediate group with an MSE between 70 and 80 ppb^2, only approaches that do not use temporally decomposed inputs are found, including the IntelliO3-ts-v1 model. Overall, the FCN performs worst with a mean MSE of 78 ppb^2, and the best results in this group are achieved with the CNN. In the leading group are exclusively methods that rely on the decomposition of the input time series. The OLS with the LT/ST decomposition has the highest error within this group with 68 ppb^2. The lowest errors can be obtained with the MB-FCN and the MB-RNN, whereby the MSE for both NNs is around 66 ppb^2.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

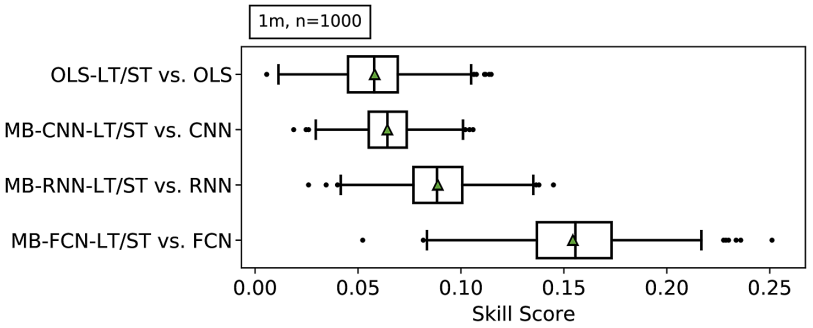

Figure 4. *Pairwise comparison of different models running with temporal decomposed or raw data by calculating the skill score on the results from the uncertainty estimation of the mean square error using a bootstrap approach represented as box-and-whiskers. For each model, the median is shown as a black vertical line, the mean as a green triangle, the upper and lower quartiles in the form of the box, the upper and lower whiskers, which correspond to 1.5 times the interquartile range, and outliers beyond the whiskers as individual data points. A total of 1,000 bootstrap samples were created by resampling with the replacement of single-month blocks.*

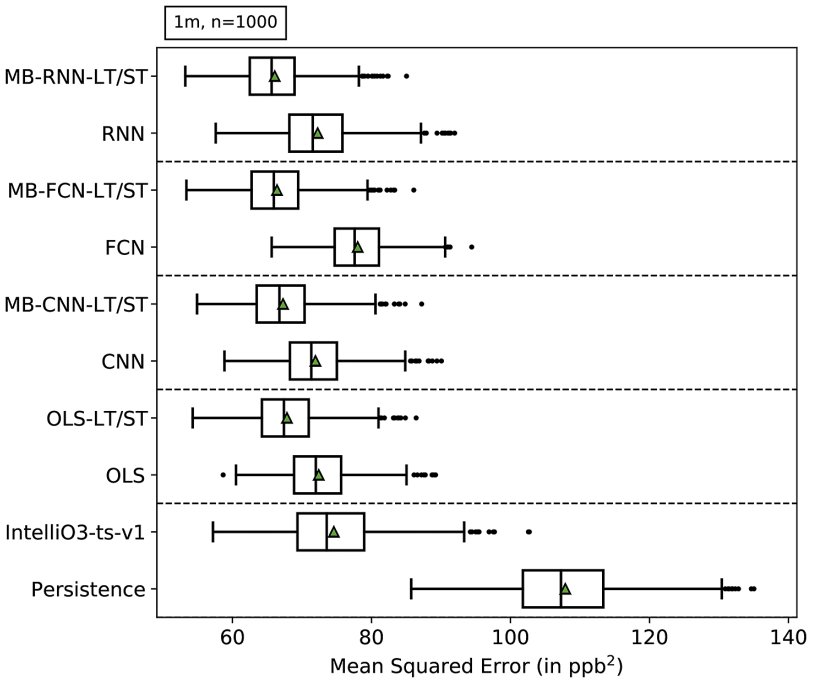

Figure 5. *The same as Figure 3, but for a different set of models. Results of the uncertainty estimation of the MSE using a bootstrap approach represented as box-and-whiskers. For each model, the median is shown as a black vertical line, the mean as a green triangle, the upper and lower quartiles in the form of the box, the upper and lower whiskers, which correspond to 1.5 times the interquartile range, and outliers beyond the whiskers as individual data points. The models are ordered from top to bottom with ascending average MSE. A total of 1,000 bootstrap samples were created by resampling with the replacement of single-month blocks. Note that the uncertainty estimation shown here is independent of the results shown in Figure 3, and therefore numbers may vary for statistical reasons.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

In order to understand why the decomposition consistently brings about an improvement for all methods considered here, we look exemplarily at the MB-FCN-LT/ST in more detail in the following. However, it should be mentioned that the discussed aspects are also basically valid for the other NN types.

First, we have a look at the calibration-refinement factorization of the joint distribution (Figure 6) according to equation (14). It can be seen that the distribution of the forecasted concentration of ozone becomes narrower toward the mean with increasing lead time. While the MB-FCN-LT/ST is still able to predict values of >70 ppb for the 1-day forecast, it is limited to values below 60 ppb for the 4-day forecast and tends to underestimate larger concentrations with increasing lead time. According to the conditional probability of observing an issued forecast, the MB-FCN-LT/ST is best calibrated for the first forecast day and especially in the value range from 20 to 60 ppb. However, observations of high ozone concentrations, starting from values above 60 ppb, are generally underestimated by the NN. Coupled with the already mentioned narrowing of the forecast's distribution, the underestimation of high ozone concentrations increases with lead time.

The shortcomings with the prediction of the tails of the distribution of observations are also evident when looking at the seasonal behavior of the MB-FCN-LT/ST. Figure 7 summarizes the distribution of observations and predictions of the NN for each month. The narrowing toward the mean with increasing lead time is also clearly visible here in the whiskers and the interquartile range in the form of the box. However, it can already be observed that, from a climatological perspective, the forecasts are in the range

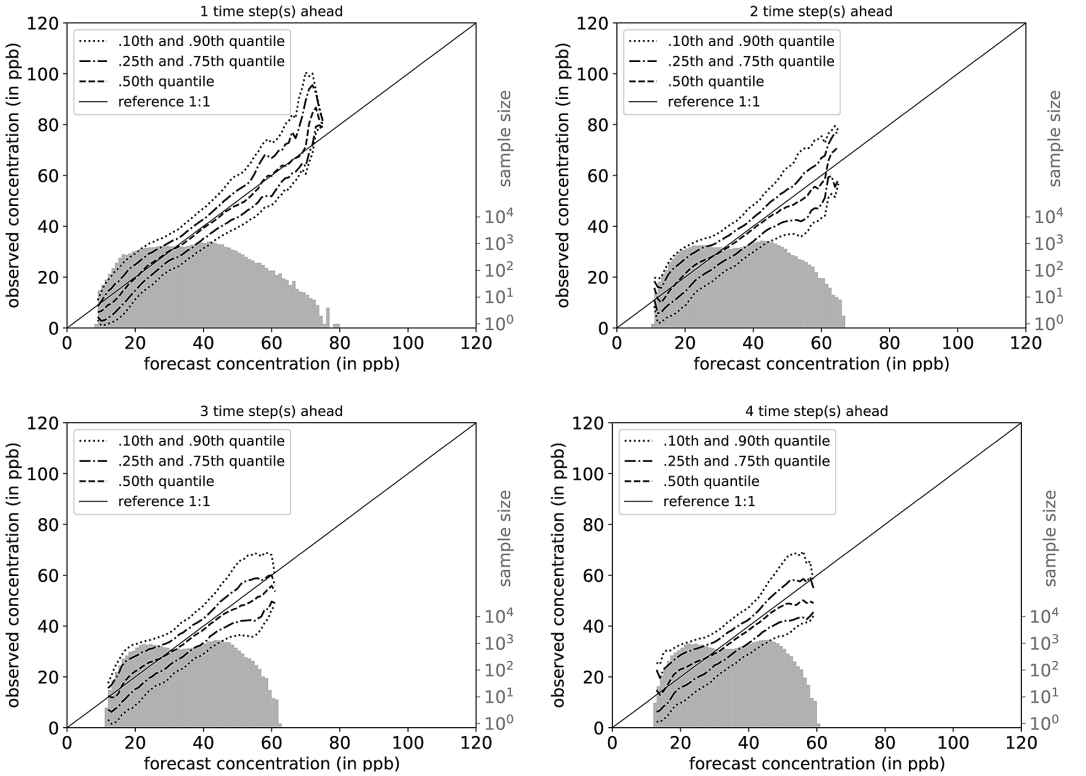

Figure 6. *Joint distribution of prediction and observation in the calibration-refinement factorization $p(y_i, o_j)$ for the MB-FCN-LT/ST for all four lead times. On the one hand, the marginal distribution $p(y_i)$ of the prediction is shown as a histogram in gray colors with the axis on the right, and on the other hand, the conditional probability $p(o_j|y_i)$ is expressed by quantiles in the form of differently dashed lines. The reference line of a perfectly calibrated forecast is also shown as a solid line.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

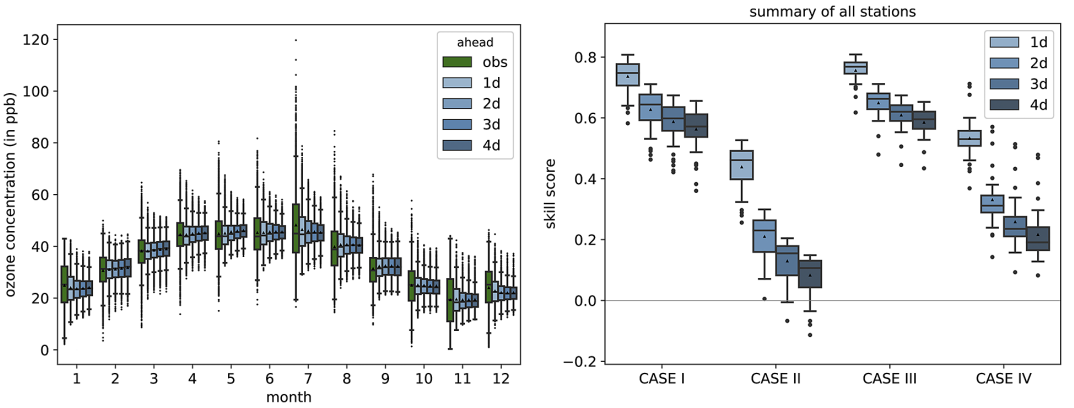

Figure 7. *Overview of the climatological behavior of the MB-FCN-LT/ST forecast shown as a monthly distribution of the observation and forecasts on the left and the analysis of the climatological skill score according to Murphy (1988) differentiated into four cases on the right. The observations are highlighted in green and the forecasts in blue. As in Figure 3, the data are presented as box-and-whiskers, with the black triangle representing the mean.*

of the observations, and the annual cycle of the ozone concentration can be modeled. Yet the month of July stands out in particular, where it is clearly recognizable that the NN is not able to represent the large variability of values from 20 ppb to values over 100 ppb that occur during summer.

The direct comparison according to Murphy (1988) between the climatological annual mean of the observation and the forecast of the NN for the training and validation data (Case 1) as well as for the test data (Case 3) shows a high skill score in favor of the NN compared to the single-valued climatological reference as the NN captures the seasonal cycle (Figure 7). Furthermore, in direct comparison with the climatological monthly means (Cases 2 and 4), the MB-FCN-LT/ST can achieve an added value in terms of information. However, the skill score on all datasets decreases gradually with longer lead times. Nonetheless, a nearly continuously positive skill score shows that the seasonal pattern of the observations can be simulated by the NN.

The feature importance analysis provides insight on which variables the MB-FCN-LT/ST generally relies upon. An examination of the importance of the individual branches, as shown in Figure 8, shows that, for the first forecast day, both LT and ST have a significant influence on the forecast accuracy. For longer forecast horizons, this influence decreases visibly, especially for ST. It is worth noting here that the influence of LT decreases less for Days 2–4, remaining at an almost constant level. Consequently, the long-term components of the decomposed time series have an important influence on all forecasts.

Looking at the importance of each variable with its components shows first of all that the NN is strongly dependent on the input ozone concentration. This dependence decreases continuously with lead time. Important meteorological drivers are temperature, relative humidity, and planetary boundary layer height. All these variables diminish in importance with increasing forecast horizon, analogously to the importance of the ozone concentration. On the chemical side, NO_2 also has an influence. Here, it must be emphasized that, in contrast to the other variables, the influence does not decrease with lead time, but remains constant over all forecast days. From the feature importance, we can see that the trained model does not make extensive use of information from wind, NO, or cloud cover.

Isolating the effects of the individual inputs in the LT branch shows that the NN is hardly dependent on the long-term components of the input variables apart from ozone (see Figure 9). The importance of ozone is higher on Day 1 than on the following days, but then remains at a constant level. For the short-term components, the concentration of ozone is also decisive. However, its influence decreases rapidly from the 1-day to the 2-day forecast. The individual importance of the ST components of the other input variables behaves in the same way as the overall importance of these variables.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

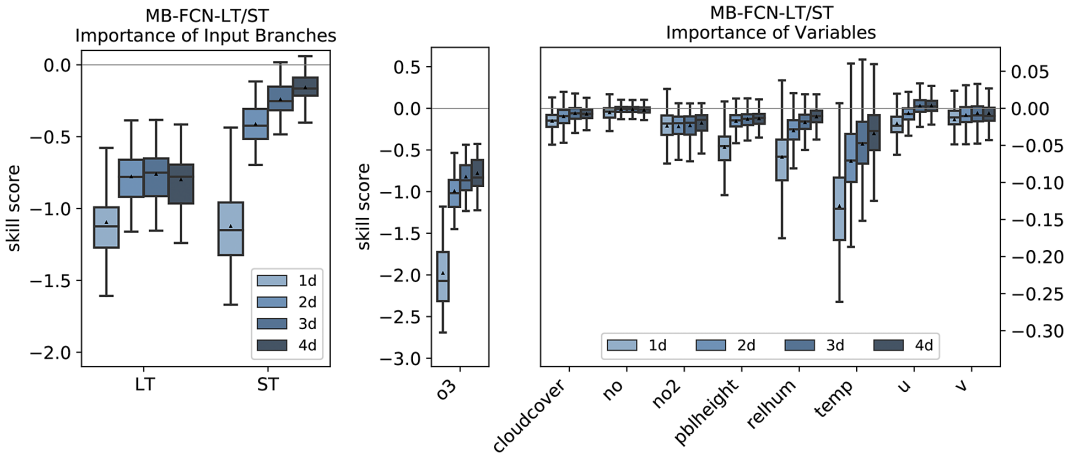

Figure 8. *Importance of single branches (left) and single variables (right) for the MB-FCN-LT/ST using bootstrapping. In blue colors, the skill score for lead times from 1 day (light blue) to 4 days (dark blue) is shown. A negative skill score indicates a strong influence on the forecast performance. The skill score is calculated with the original undisturbed prediction of the same NN as reference. Note that due to the significantly stronger dependence, ozone is visualized on a separate scale.*

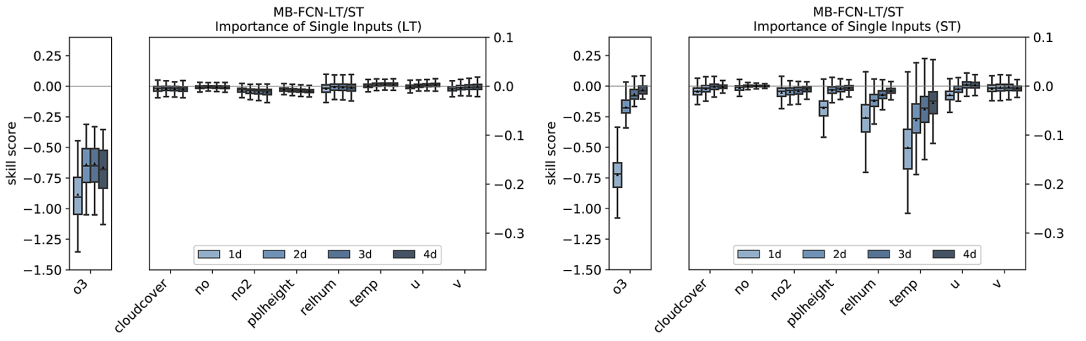

Figure 9. *Importance of single inputs for the LT branch (left) and the ST branch (right) for the MB-FCN-LT/ST using a bootstrap approach. In blue colors, the skill score for lead times from 1 day (light blue) to 4 days (dark blue) is shown. A negative skill score indicates a dependence. The skill score is calculated with the original undisturbed prediction of the same NN as reference.*

As previously mentioned, the points discussed before can be more or less transferred to the other NN architectures. The feature importance analysis of the branches and the individual variables for MB-CNN and MB-RNN is shown in Figures E1–E3 in Appendix E. In particular, the LT for all forecast days and the ST for the first day contain important information, with the ST branch being less relevant for the MB-RNN. Moreover, MB-CNN and MB-RNN also show a narrowing of the distribution of issued forecasts with increasing lead time, as was also observed for MB-FCN.

5. Discussion

The experimental results described in the previous section indicate that the NNs learn oscillation patterns on different time scales, and in particular climatological properties, better when the input time series are explicitly decomposed into different temporal components. The MB-NNs outperform all reference

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

models, such as simple statistical regression methods as well as the naïve persistence forecasts and climatological references. The MB-NNs are also preferable to their corresponding counterparts without temporal decomposition, considered individually but also as a collective.

The uncertainty estimate of the MSE of the forecast shows that FCNs that either use a decomposition into a long-term and a short-term component or access unfiltered raw data as a supplementary source of information have the highest forecast accuracy. Separating the input signals into more than two components without adding the unfiltered raw data cannot improve the performance of the FCNs, with respect to the architectures chosen in this study. This recognition coincides with the findings of Kang et al. (2013), who show that a clear separation of the short-term components is generally not possible due to the superposition of multiple oscillation patterns.

With regard to the network architecture, several key points can be identified in this study. Without special processing of the input data, the best results were achieved with a CNN architecture. This could be explained by the fact that the convolutional layers of the CNN already filter the time series. However, it must also be mentioned that with a filter size of only 5 hr, there is no chance to extract an annual cycle, so that the explicit decomposition into LT and ST components also offers added value for the CNN. However, due to the higher baseline level, the MB-CNN cannot benefit as much from the data processing compared to the MB-FCN and MB-RNN and is moreover behind the other two MB-NNs in terms of prediction quality in absolute terms. The RNN also achieves better results on the unfiltered data than the FCN, for example, because it can benefit from a more specific understanding of time. The FCN is therefore inferior to the CNN and RNN due to its comparatively simple architecture and the lack of possibility to relate neighboring data points explicitly. However, it benefits most from the temporal decomposition of the inputs, so that these disadvantages disappear, and overall, the smallest errors can be achieved with MB-FCN and MB-RNN. These finding therefore highlight the importance of jointly optimizing data preprocessing and NN model architecture, which is taught in many ML courses, but not always followed in practical applications.

The difficulties of NNs to recognize annual patterns in daily resolved data noted by Cho et al. (2014) did not apply to the MB-NNs. However, the networks still encounter difficulties in anticipating very low and very high ozone concentrations. As the lead time increases, the NN's forecast strategy becomes more cautious about extremes, leading to a narrowing of the distribution of issued forecasts. Despite this circumstance, the NNs always retain within an optimal range from a climatological point of view, so that the forecast has higher accuracy than a climatological forecast. The analysis of the feature importance can provide an explanation for this. For the first day of the forecast, both long-term and short-term components have an equally strong influence on the MB-FCN forecast, but for a longer forecast horizon, the long-term components are given more weight. Accordingly, the LT branch in particular enables the NN to generate a climatologically meaningful forecast. In addition, the NN remains strongly dependent on the ozone concentrations from the inputs. Learning a form of autocorrelation is advantageous for climatological accuracy, but at the same time leads to a poorer representation of scarcer events such as sudden and strong increases in the daily maximum concentration from one day to the next.

In addition, it must be mentioned that strong deviations from climatological norm states also have an impact on the filter decomposition of the inputs, since climatology can only be an estimate of a long-term mean state, which can deviate strongly from the actual weather in individual cases. For example, the long-term signal of temperature in the case of a very warm summer would be weakened by the added climatology, since such a deviation represents an exceptional case from a climatological point of view. In this case, the second filter component, which should actually be free of an annual variation, also contains a proportion of an annual oscillation. However, as discussed in Section 2, this combination allows to apply noncausal filters in a forecasting situation, where generally only causal filters are applicable, which lead to phase shifts in the data and show larger errors.

A look at the importance of the individual inputs for the MB-FCN yields two views. First, it becomes apparent that the dependency of the LT and ST components are each strongly based on the corresponding component of the ozone concentration and that the MB-FCN accordingly learns the connection between observed hourly ozone values and the target ozone statistic. Second, all other variables seem to have an

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

influence only on the short-term scale. Since the ST component by definition represents the deviation from the climatological normal state, it can be seen that the MB-FCN relies on the deviation from normal states as a forecasting strategy.

Finally, we would like to discuss the filters used for the decomposition. Since there are many different types of filters with various advantages and disadvantages, we have limited this work to the use of a Kaiser window and have not carried out any further experiments with different types of filters, such as the KZ filter, which could possibly lead to an improved separation of individual components as stated in Rao and Zurbenko (1994) in the presence of a weather forecast. Furthermore, we have not undertaken any in-depth investigations into which separation frequencies lead to an optimal decomposition of the time series.

6. Conclusion

In this work, we explored the potential of training different NNs, namely FCN, CNN, and RNN, for dma8eu ozone forecasting using inputs decomposed into different frequency components from long term to short term with noncausal filters in order to improve the forecast accuracy of the NNs. The temporal decomposition of the inputs not only improves the different NN architectures and the linear OLS model, but also offers an overall added value for the prediction of ozone compared to all reference models using raw hourly inputs and naïve approaches based on persistence and climatology. As exemplary shown with the MB-FCN, the MB-NNs work better with a decomposition into two components compared to four and they rely on both long-term and short-term components for their prediction, with a strong dependence on past ozone observations and a decreasing importance of the short-term components with lead time. In order to realize a valid decomposition in a forecast setup without time delay of the signal introduced by the filter itself, a combination of observations and a priori information in the form of climatology was chosen.

Acknowledgements. The authors gratefully acknowledge the Jülich Supercomputing Centre for providing the infrastructure and resources to run all experiments. We are also thankful to all air quality data providers that made their data available in the TOAR DB.

Author Contributions. Conceptualization: All authors; Data curation: L.H.L. and F.K.; Formal analysis: L.H.L.; Funding acquisition: M.G.S.; Investigation: L.H.L.; Methodology: L.H.L.; Project administration: M.G.S.; Resources: M.G.S.; Software: L.H.L. and F.K.; Supervision: M.G.S.; Validation: All authors; Visualization: L.H.L. and F.K.; Writing—original draft: L.H.L.; Writing—review and editing: All authors.

Data Availability Statement. Replication data and code can be found on b2share: https://doi.org/10.34730/decca0f4bbfc42c693812f7a648b2d6f.

Funding Statement. This research was supported by grants from the European Research Council, H2020 Research Infrastructures (Grant No. 787576 [IntelliAQ]).

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Competing Interests. The authors declare none.

References

Bagnall A, **Lines J**, **Bostrom A**, **Large J and Keogh E** (2017) The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery 31*(3), 606–660.

Bauerle A, **van Onzenoodt C and Ropinski T** (2021) Net2Vis—a visual grammar for automatically generating publication-tailored CNN architecture visualizations. *IEEE Transactions on Visualization and Computer Graphics 27*(6), 2980–2991.

Bessagnet B, **Pirovano G**, **Mircea M**, **Cuvelier C**, **Aulinger A**, **Calori G**, **Ciarelli G**, **Manders A**, **Stern R**, **Tsyro S**, **García Vivanco M**, **Thunis P**, **Pay M-T**, **Colette A**, **Couvidat F**, **Meleux F**, **Rouíl L**, **Ung A**, **Aksoyoglu S**, **Baldasano JM**, **Bieser J**, **Briganti G**, **Cappelletti A**, **D'Isidoro M**, **Finardi S**, **Kranenburg R**, **Silibello C**, **Carnevale C**, **Aas W**, **Dupont J-C**, **Fagerli H**, **Gonzalez L**, **Menut L**, **Prévôt ASH**, **Roberts P and White L** (2016) Presentation of the EURODELTA III intercomparison exercise—evaluation of the chemistry transport models' performance on criteria pollutants and joint analysis with meteorology. *Atmospheric Chemistry and Physics 16*(19), 12667–12701.

Bollmeyer C, **Keller JD**, **Ohlwein C**, **Wahl S**, **Crewell S**, **Friederichs P**, **Hense A**, **Keune J**, **Kneifel S**, **Pscheidt I**, **Redl S and Steinke S** (2015) Towards a high-resolution regional reanalysis for the European cordex domain. *Quarterly Journal of the Royal Meteorological Society 141*(686), 1–15.

Chandra R, **Goyal S and Gupta R** (2021) Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access 9*, 83105–83123.

Cho K, **Van Merriënboer B**, **Bahdanau D and Bengio Y** (2014) On the properties of neural machine translation: Encoder–decoder approaches. *Preprint*, arXiv:1409.1259.

Chung J, **Gulcehre C**, **Cho K and Bengio Y** (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *Preprint*, arXiv:1412.3555.

Clevert D-A, **Unterthiner T and Hochreiter S** (2016) Fast and accurate deep network learning by exponential linear units (ELUs). *Preprint*, arXiv:1511.07289.

Cohen, A. J., **Brauer, M.**, **Burnett, R.**, **Anderson, H. R.**, **Frostad, J.**, **Estep, K.**, **Balakrishnan, K.**, **Brunekreef, B.**, **Dandona, L.**, **Dandona, R.**, **Feigin, V.**, **Freedman, G.**, **Hubbell, B.**, **Jobling, A.**, **Kan, H.**, **Knibbs, L.**, **Liu, Y.**, **Martin, R.**, **Morawska, L.**, … **Forouzanfar, M. H.** (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907-1918. https://doi.org/10.1016/s0140-6736(17)30505-6

Collins WJ, **Stevenson DS**, **Johnson CE and Derwent RG** (1997) Tropospheric ozone in a global-scale three-dimensional lagrangian model and its response to nox emission controls. *Journal of Atmospheric Chemistry 26*(3), 223–274.

Comrie AC (1997) Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association 47*(6), 653–663.

Cui Z, **Chen W and Chen Y** (2016) Multi-scale convolutional neural networks for time series classification. Preprint, arXiv: 1603.06995.

De Gooijer JG and Hyndman RJ (2006) 25 years of time series forecasting. *International Journal of Forecasting 22*(3), 443–473.

Di Q, **Dai L**, **Wang Y**, **Zanobetti A**, **Choirat C**, **Schwartz JD and Dominici F** (2017) Association of short-term exposure to air pollution with mortality in older adults. *JAMA 318*(24), 2446–2456.

Donner LJ, **Wyman BL**, **Hemler RS**, **Horowitz LW**, **Ming Y**, **Zhao M**, **Golaz J-C**, **Ginoux P**, **Lin S-J**, **Schwarzkopf MD**, **Austin J**, **Alaka G**, **Cooke WF**, **Delworth TL**, **Freidenreich SM**, **Gordon CT**, **Griffies SM**, **Held IM**, **Hurlin WJ**, **Klein SA**, **Knutson TR**, **Langenhorst AR**, **Lee H-C**, **Lin Y**, **Magi BI**, **Malyshev SL**, **Milly PCD**, **Naik V**, **Nath MJ**, **Pincus R**, **Ploshay JJ**, **Ramaswamy V**, **Seman CJ**, **Shevliakova E**, **Sirutis JJ**, **Stern WF**, **Stouffer RJ**, **Wilson RJ**, **Winton M**, **Wittenberg AT and Zeng F** (2011) The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *Journal of Climate 24*(13), 3484–3519.

Elkamel A, **Abdul-Wahab S**, **Bouhamra W and Alper E** (2001) Measurement and prediction of ozone levels around a heavily industrialized area: A neural network approach. *Advances in Environmental Research 5*(1), 47–59.

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. *Official Journal of the European Union 29*, 169–212.

Fawaz HI, **Lucas B**, **Forestier G**, **Pelletier C**, **Schmidt DF**, **Weber J**, **Webb GI**, **Idoumghar L**, **Muller P-A and Petitjean F** (2020) InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery 34*(6), 1936–1962.

Fleming ZL, **Doherty RM**, **von Schneidemesser E**, **Malley CS**, **Cooper OR**, **Pinto JP**, **Colette A**, **Xu X**, **Simpson D**, **Schultz MG**, **Lefohn AS**, **Hamad S**, **Moolla R**, **Solberg S and Feng Z** (2018) Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. *Elementa: Science of the Anthropocene 6*, 12.

Fuentes M and Raftery AE (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics 61*(1), 36–45.

Gardner M and Dorling S (1999) Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment 33*(5), 709–719.

Granier C, **Bessagnet B**, **Bond T**, **D'Angiola A**, **van Der Gon HD**, **Frost GJ**, **Heil A**, **Kaiser JW**, **Kinne S**, **Klimont Z**, *et al.* (2011) Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Climatic Change 109*(1), 163–190.

Grell GA, **Peckham SE**, **Schmitz R**, **McKeen SA**, **Frost G**, **Skamarock WC and Eder B** (2005) Fully coupled "online" chemistry within the WRF model. *Atmospheric Environment 39*(37), 6957–6975.

He K, **Zhang X**, **Ren S and Sun J** (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, pp. 1026–1034.

Hilboll A, **Richter A and Burrows J** (2013) Long-term changes of tropospheric NO_2 over megacities derived from multiple satellite instruments. *Atmospheric Chemistry and Physics 13*(8), 4145–4169.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Hornik K, **Stinchcombe M and White H** (1989) Multilayer feedforward networks are universal approximators. *Neural Networks 2*(5), 359–366.

Horowitz LW, **Stacy W**, **Mauzerall DL**, **Emmons LK**, **Rasch PJ**, **Granier C**, **Tie X**, **Lamarque J**, **Schultz MG**, **Tyndall GS**, **Orlando JJ and Brasseur GP** (2003) A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2. *Journal of Geophysical Research: Atmospheres 108*(D24), 2–6.

Jiang G, **He H**, **Yan J and Xie P** (2019) Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. *IEEE Transactions on Industrial Electronics 66*(4), 3196–3207.

Kaiser JF (1966) Chapter 7: Digital filters. In Kuo FF and Kaiser JF (eds), *System Analysis by Digital Computer*. New York: Wiley, pp. 218–285.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Kang D, **Hogrefe C**, **Foley KL**, **Napelenok SL**, **Mathur R and Rao ST** (2013) Application of the Kolmogorov–Zurbenko filter and the decoupled direct 3D method for the dynamic evaluation of a regional air quality model. *Atmospheric Environment 80*, 58–69.

Keren G and Schuller B (2016) Convolutional RNN: An enhanced model for extracting features from sequential data. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 3412–3419.

Klambauer G, **Unterthiner T**, **Mayr A and Hochreiter S** (2017) Self-normalizing neural networks. In *Advances in Neural Information Processing Systems 30*. Curran Associates Inc.

Kleinert F, **Leufen LH and Schultz MG** (2021) IntelliO3-ts v1.0: A neural network approach to predict near-surface ozone concentrations in Germany. *Geoscientific Model Development 14*(1), 1–25.

Kolehmainen M, **Martikainen H and Ruuskanen J** (2001) Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment 35*(5), 815–825.

Kumar V and Sinha V (2021) Season-wise analyses of VOCs, hydroxyl radicals and ozone formation chemistry over north-west India reveal isoprene and acetaldehyde as the most potent ozone precursors throughout the year. *Chemosphere 283*, 131184.

LeCun Y, **Haffner P**, **Bottou L and Bengio Y** (1999) Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*. Springer, Berlin, Heidelberg, pp. 319–345.

Lefohn AS, **Malley CS**, **Simon H**, **Wells B**, **Xu X**, **Zhang L and Wang T** (2017) Responses of human health and vegetation exposure metrics to changes in ozone concentration distributions in the European Union, United States, and China. *Atmospheric Environment 152*, 123–145.

Leufen LH, **Kleinert F and Schultz MG** (2021) MLAir (v1.0)—a tool to enable fast and flexible machine learning on air data time series. *Geoscientific Model Development 14*(3), 1553–1574.

Leufen LH, **Kleinert F**, **Weichselbaum F**, **Gramlich V and Schultz MG** (2022) MLAir—a tool to enable fast and flexible machine learning on air data time series, version 2.0.0, source code. Available at https://gitlab.jsc.fz-juelich.de/esde/machine-learning/mlair/-/tags/v2.0.0 (accessed 21 June 2022).

Liang M and Hu X (2015) Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3367–3375.

Lou Thompson M, **Reynolds J**, **Cox LH**, **Guttorp P and Sampson PD** (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment 35*(3), 617–630.

Maas AL, **Hannun AY**, **Ng AY**, *et al.* (2013) Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. *30*. Atlanta, Georgia, USA, p. 3.

Maas R and Grennfelt P (eds) (2016) *Towards Cleaner Air. Scientific Assessment Report 2016.* EMEP Steering Body and Working Group on Effects of the Convention on Long-Range Transboundary Air Pollution, Oslo.

Manders AMM, **van Meijgaard E**, **Mues AC**, **Kranenburg R**, **van Ulft LH and Schaap M** (2012) The impact of differences in large-scale circulation output from climate models on the regional modeling of ozone and PM. *Atmospheric Chemistry and Physics 12*(20), 9441–9458.

Monks PS, **Archibald A**, **Colette A**, **Cooper O**, **Coyle M**, **Derwent R**, **Fowler D**, **Granier C**, **Law KS**, **Mills G**, *et al.* (2015) Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric Chemistry and Physics 15*(15), 8889–8973.

Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review 116*(12), 2417–2424.

Murphy AH (1992) Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and Forecasting 7*(4), 692–698.

Murphy AH and Winkler RL (1987) A general framework for forecast verification. *Monthly Weather Review 115*(7), 1330–1338.

Oppenheim AV and Schafer RW (1975) *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Ortiz A and Guerreiro C (2020) *Air Quality in Europe—2020 Report.* European Environment Agency, Publications Office, Copenhagen, Denmark.

Rao S, **Zurbenko I**, **Neagu R**, **Porter P**, **Ku J and Henry R** (1997) Space and time scales in ambient ozone data. *Bulletin of the American Meteorological Society 78*(10), 2153–2166.

Rao ST and Zurbenko IG (1994) Detecting and tracking changes in ozone air quality. *Air & Waste 44*(9), 1089–1092.

Reichstein M, **Camps-Valls G**, **Stevens B**, **Jung M**, **Denzler J**, **Carvalhais N**, *et al.* (2019) Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204.

REVIHAAP (2013) *Review of Evidence on Health Aspects of Air Pollution—REVIHAAP Project Technical Report.* Bonn: World Health Organization (WHO) Regional Office for Europe.

Richter A, **Burrows JP**, **Nüß H**, **Granier C and Niemeier U** (2005) Increase in tropospheric nitrogen dioxide over China observed from space. *Nature 437*(7055), 129–132.

Russell A, **Valin L and Cohen R** (2012) Trends in OMI NO_2 observations over the United States: Effects of emission control technology and the economic recession. *Atmospheric Chemistry and Physics 12*(24), 12197–12209.

Schultz MG, **Betancourt C**, **Gong B**, **Kleinert F**, **Langguth M**, **Leufen LH**, **Mozaffari A and Stadtler S** (2021) Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A 379*(2194), 20200097.

Schultz MG, **Schröder S**, **Lyapina O**, **Cooper OR**, **Galbally I**, **Petropavlovskikh I**, **Von Schneidemesser E**, **Tanimoto H**, **Elshorbany Y**, **Naja M**, *et al.* (2017). Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations. *Elementa: Science of the Anthropocene 5*, 58.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Seltzer KM, **Shindell DT**, **Faluvegi G**, & **Murray LT** (2017). Evaluating Modeled Impact Metrics for Human Health, Agriculture Growth, and Near-Term Climate. *Journal of Geophysical Research: Atmospheres*, *122*(24), 13, 506–13, 524. Portico. https://doi.org/10.1002/2017jd026780

Seltzer KM, **Shindell DT**, **Kasibhatla P and Malley CS** (2020) Magnitude, trends, and impacts of ambient long-term ozone exposure in the United States from 2000 to 2015. *Atmospheric Chemistry and Physics 20*(3), 1757–1775.

Shih S-Y, **Sun F-K and Lee H** (2019) Temporal pattern attention for multivariate time series forecasting. *Machine Learning 108*(8), 1421–1441.

Shindell D, **Faluvegi G**, **Kasibhatla P and Van Dingenen R** (2019) Spatial patterns of crop yield change by emitted pollutant. *Earth's Future 7*(2), 101–112.

Simon H, **Reff A**, **Wells B**, **Xing J and Frank N** (2015) Ozone trends across the United States over a period of decreasing NO_x and VOC emissions. *Environmental Science & Technology 49*(1), 186–195.

Szegedy C, **Liu W**, **Jia Y**, **Sermanet P**, **Reed S**, **Anguelov D**, **Erhan D**, **Vanhoucke V and Rabinovich A** (2015) Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1–9.

TOAR (2019). *Tropospheric Ozone Assessment Report (TOAR): Global Metrics for Climate Change, Human Health and Crop/Ecosystem Research.* International Global Atmospheric Chemistry. Available at https://igacproject.org/activities/TOAR (accessed 29 January 2021).

US EPA (2013) *Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants (Final Report, February 2013)*. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-10/076F.

US EPA (2020) *Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants (Final Report, April 2020)*. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-20/012.

Vautard R, **Moran MD**, **Solazzo E**, **Gilliam RC**, **Matthias V**, **Bianconi R**, **Chemel C**, **Ferreira J**, **Geyer B**, **Hansen AB**, **Jericevic A**, **Prank M**, **Segers A**, **Silver JD**, **Werhahn J**, **Wolke R**, **Rao S and Galmarini S** (2012) Evaluation of the meteorological forcing used for the air quality model evaluation international initiative (AQMEII) air quality simulations. *Atmospheric Environment 53*, 15–37. AQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models—Phase 1.

von Kuhlmann R, **Lawrence MG**, **Crutzen PJ and Rasch PJ** (2003) A model for studies of tropospheric ozone and nonmethane hydrocarbons: Model description and ozone results. *Journal of Geophysical Research: Atmospheres 108*(D9), 4294.

Wang Y, **Jacob DJ and Logan JA** (1998a) Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry: 1. Model formulation. *Journal of Geophysical Research: Atmospheres 103*(D9), 10713–10725.

Wang Y, **Logan JA and Jacob DJ** (1998b) Global simulation of tropospheric O_3-NO_x-hydrocarbon chemistry: 2. Model evaluation and global ozone budget. *Journal of Geophysical Research: Atmospheres 103*(D9), 10727–10755.

Wilks DS (2006) *Statistical Methods in the Atmospheric Sciences*, 2nd Edn. London: Academic Press.

Wise EK and Comrie AC (2005) Extending the Kolmogorov–Zurbenko filter: Application to ozone, particulate matter, and meteorological trends. *Journal of the Air & Waste Management Association 55*(8), 1208–1216.

Young PJ, **Naik V**, **Fiore AM**, **Gaudel A**, **Guo J**, **Lin MY**, **Neu JL**, **Parrish DD**, **Rieder HE**, **Schnell JL**, **Tilmes S**, **Wild O**, **Zhang L**, **Ziemke J**, **Brandt J**, **Delcloo A**, **Doherty RM**, **Geels C**, **Hegglin MI**, **Hu L**, **Im U**, **Kumar R**, **Luhar A**, **Murray L**, **Plummer D**, **Rodriguez J**, **Saiz-Lopez A**, **Schultz MG**, **Woodhouse MT and Zeng G** (2018) 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.

Zhang Y, **Cooper OR**, **Gaudel A**, **Thompson AM**, **Nédélec P**, **Ogino S-Y and West JJ** (2016) Tropospheric ozone change from 1980 to 2010 dominated by equatorward redistribution of emissions. *Nature Geoscience 9*(12), 875–879.

Zhang Y, **West JJ**, **Mathur R**, **Xing J**, **Hogrefe C**, **Roselle SJ**, **Bash JO**, **Pleim JE**, **Gan C-M and Wong DC** (2018) Long-term trends in the ambient $PM_{2.5}$- and O_3-related mortality burdens in the United States under emission reductions from 1990 to 2010. *Atmospheric Chemistry and Physics 18*, 1–14.

Zhao J, **Huang F**, **Lv J**, **Duan Y**, **Qin Z**, **Li G and Tian G** (2020) Do RNN and ISTM have long memory? In *Proceedings of the 37th International Conference in Machine Learning*. Proceedings of Machine Learning Research (PMLR) *119*, online, pp. 11365–11375.

Zheng Y, **Liu Q**, **Chen E**, **Ge Y and Zhao JL** (2014) Time series classification using multi-channels deep convolutional neural networks. In Li F, Li G, Hwang S-w, Yao B and Zhang Z (eds), *Web-Age Information Management*. Cham: Springer International Publishing, pp. 298–310.

Ziyin L, **Hartwig T and Ueda M** (2020) Neural networks fail to learn periodic functions and how to fix it. In Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds), *Advances in Neural Information Processing Systems*, Vol. *33*. Curran Associates, Inc., pp. 1583–1594.

Žurbenko IG (1986) *The Spectral Analysis of Time Series*, Vol. *2*. North-Holland Series in Statistics and Probability. Elsevier, Amsterdam.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Appendix A: Details on Data

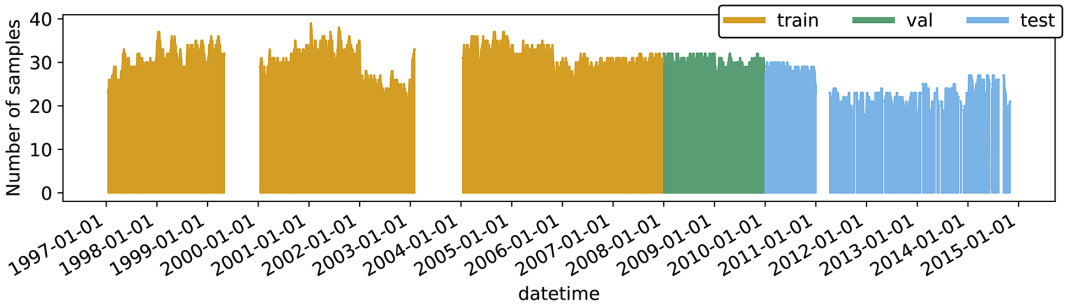

Figure A1. *Graphical representation of the number of samples available for training (orange), validation (green), and testing (blue) per time step. Apart from three periods in which the data cannot meet the requirements, more than 20 stations are available at each time step, and for training in particular, more than 30 stations for the most time. The graph does not show the available raw data, but indicates for which time steps t_0 a sample with fully processed input and target values is available.*

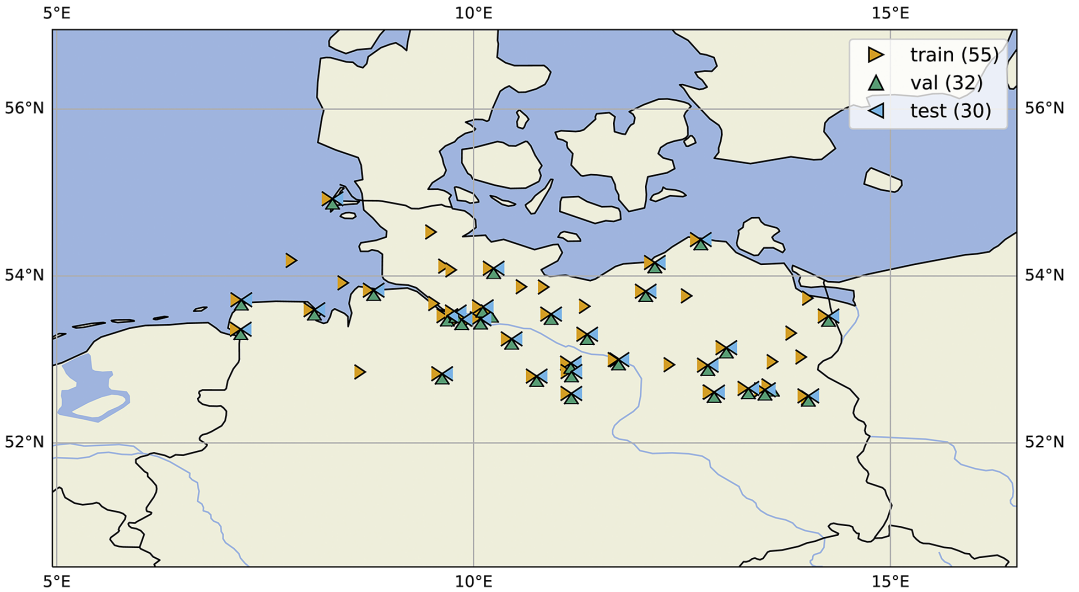

Figure A2. *Geographical location of all rural and suburban monitoring stations used in this study divided into training (orange), validation (green), and test (blue) data represented by triangles in the corresponding colors. The tip of the triangles points to the exact location of the station.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

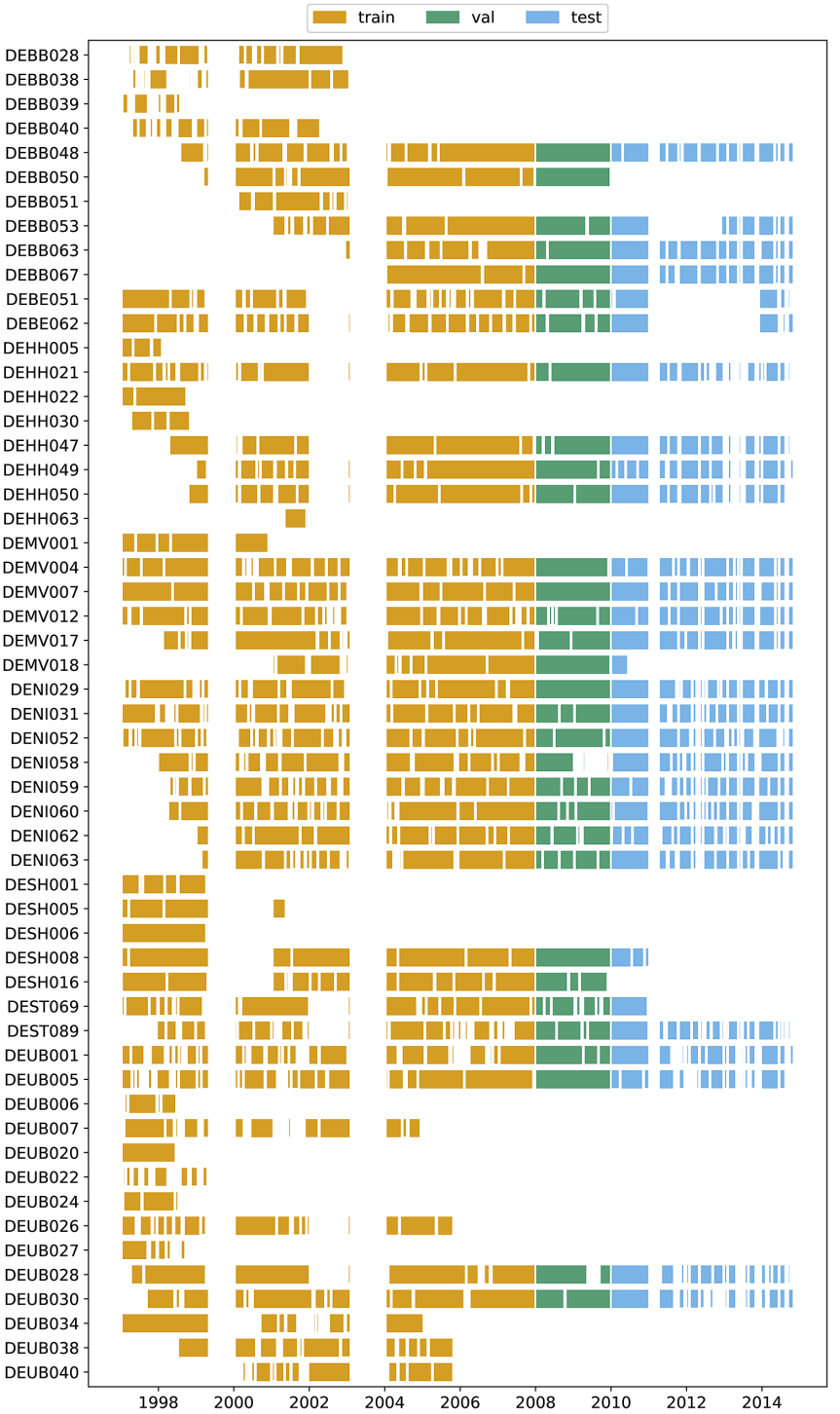

Figure A3. *Detailed overview of the availability of station data broken down for all individual stations as a timeline separated by color for training (orange), validation (green), and test (blue) data. Individual gaps are caused by missing observation data that exceed the interpolation limit of 24 hr for inputs or 2 days for targets.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Appendix B: Details on Hyperparameter Search

Table B1. Details on tested hyperparameters for the MB-FCNs. The square brackets indicate a continuous parameter range, and the curly brackets indicate a fixed set of parameters. Parameter spaces covering different orders of magnitude were sampled on a logarithmic scale. For details on the activation functions, we refer to rectified linear unit (ReLU) and leaky rectified linear unit (LeakyReLU, Maas et al., 2013), exponential linear unit (ELU, Clevert et al., 2016), scaled exponential linear unit (SELU, Klambauer et al., 2017), and parametric rectified linear unit (PReLU, He et al., 2015).

Parameter	Parameter range
Learning rate	[0.1, 0.0001]
Learning rate decay	[0, 0.0001]
Batch size	{64, 128, 256, 512}
Activation function	{relu, leakyrelu, elu, selu, prelu}
Dropout	[0, 0.5]
Batch normalization	{true, false}
Branch layers	{512/256/128, 512/128/32, 512/64, 512/32, 256/128/64/32, 256/64, 128/64, 128/32, 64/32}
Tail layers	{4, 32/4, 64/4}

Table B2. Summary of best hyperparameters and fixed parameters for different setups with MB-FCN. The entire parameter ranges of all hyperparameters are given in Table B1. Details on the activation functions can be found in He et al. (2015) for the parametric rectified linear unit (PReLU) and in Clevert et al. (2016) for the exponential linear unit (ELU). A visualization of MB-FCN-LT/ST can be found in Figure D1 in addition.

Parameter	MB-FCN-BL/ SY/DU/ID	MB-FCN- LT/ST	MB-FCN-BL/SY/ DU/ID+raw	MB-FCN-LT/ ST+raw
Hyperparameters				
Learning rate	0.00033	0.1	0.00027	0.0002
Learning rate decay	0.001	0.007	0.0001	0.0002
Batch size	512	512	512	256
Activation function	PReLU	ELU	ELU	ELU
Dropout	0.3	0.56	0.28	0.43
Batch normalization	True	True	True	True
Branch layers	64/32	128/64	64/32	128/64
Tail layers	64/4	4	4	4
Layers summary	4x(585/64/32)-64/4	2x(585/128/64)-4	5x(585/64/32)-4	3x(585/128/64)-4
Trainable parameters	168,196	167,812	199,524	251,716
Fixed				
Cutoff period(s)	21 days, 2.7 days, 11 hr	21 days	21 days, 2.7 days, 11 hr	21 days
Filter order(s)	42 days, 7 days, 2 days	42 days	42 days, 7 days, 2 days	42 days
filter window	Kaiser ($\beta = 5$)	Kaiser ($\beta = 5$)	Kaiser ($\beta = 5$)	Kaiser ($\beta = 5$)
Use unfiltered raw inputs	False	False	True	True
Number of epochs[a]	150	150	150	150

Abbreviation: FCN, fully connected network.
[a]With early stopping.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Table B3. Summary of best hyperparameters and fixed parameters for experiments with the CNN, MB-CNN, RNN, and MB-RNN. The entire parameter ranges of all hyperparameters are not listed. Details on the activation functions can be found in Maas et al. (2013) for the rectified linear unit (ReLU) and the leaky rectified linear unit (LeakyReLU) and in He et al. (2015) for the parametric rectified linear unit (PReLU).

Parameter	CNN	MB-CNN	RNN	MB-RNN
Hyperparameters				
Learning rate	0.057	0.1668	0.0009	0.0123
Learning rate decay	0.006	0.009	0.0006	0.015
Activation function	PReLU	PReLU	ReLU	LeakyReLU
Dropout	0.43	0.42	0.5 & 0.17 (recurrent)	0.23 & 0 (recurrent)
Batch normalization	Conv and FC	Conv and FC	only LSTM	only LSTM
Filter size	5×1	5×1	–	–
(Branch) layers[a]	C16/MP/C32/MP/C64	C16/MP/C32/MP/C64	LSTM64	LSTM32
Tail/dense layers	256/4	256/4	128/4	128/4
Trainable parameters	281,140	560,228	27,908	19,716
Fixed				
Number of epochs[b]	250	250	100	100
Batch size	512	512	512	512

Abbreviations: CNN, convolutional neural network; LSTM, long short-term memory.
[a]C<*n*>: Conv2D with *n* filters; LSTM<*n*>: LSTM layer with *n* LSTM cells; MP: MaxPooling.
[b]With early stopping.

Appendix C: Tabular Results

Table C1. Key numbers of the uncertainty estimation of the MSE for all MB-FCNs as an average over all prediction days using the bootstrap approach visualized in Figure 3. All reported numbers are in the unit of square parts per billion. Numbers in percentage point to the corresponding percentile of the error distribution.

	MB-FCN-BL/SY/DU/ID	MB-FCN-BL/SY/DU/ID+raw	MB-FCN-LT/ST	MB-FCN-LT/ST+raw	FCN
Mean	71.83	67.88	67.12	**66.72**	77.51
Min	56.56	**55.15**	57.38	56.17	67.01
Lower whisker	59.15	56.41	57.38	**56.17**	68.22
25%	68.53	64.92	64.25	**63.70**	75.25
50%	71.67	67.72	66.99	**66.54**	77.42
75%	74.78	70.59	69.69	**69.40**	79.94
Higher whisker	84.16	79.09	**77.86**	77.93	86.97
Max	87.52	82.17	80.89	**80.59**	91.75

Abbreviations: FCN, fully connected network; MSE, mean square error.

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

Table C2. Key numbers of the uncertainty estimation of the MSE as an average over all prediction days using the bootstrap approach visualized in Figure 5. All reported numbers are in the unit of square parts per billion. Numbers in percentage point to the corresponding percentile of the error distribution. Note that the uncertainty estimation reported here is independent of the results shown in Table C1, and therefore numbers may vary for statistical reasons.

	CNN	FCN	IntelliO3	MB-CNN-LT/ST	MB-FCN-LT/ST	MB-RNN-LT/ST	OLS-LT/ST	OLS	Persistence	RNN
Mean	71.94	78.02	74.59	67.28	66.41	**66.08**	67.84	72.41	107.89	72.26
Min	41.93	50.74	40.75	39.85	38.52	**38.12**	40.03	40.87	52.66	42.49
25%	60.52	69.90	62.47	57.43	57.20	**55.55**	58.85	59.99	83.80	61.88
50%	75.66	80.53	77.36	70.27	**69.33**	69.56	71.09	76.93	115.17	75.38
75%	82.32	86.49	85.63	76.75	**75.44**	75.77	76.95	83.40	130.46	82.02
Max	105.92	107.01	121.41	101.99	98.55	99.63	**98.38**	104.31	168.47	105.20

Abbreviations: CNN, convolutional neural network; FCN, fully connected network; MSE, mean square error; OLS, least squares regression.

Appendix D: Model Architecture

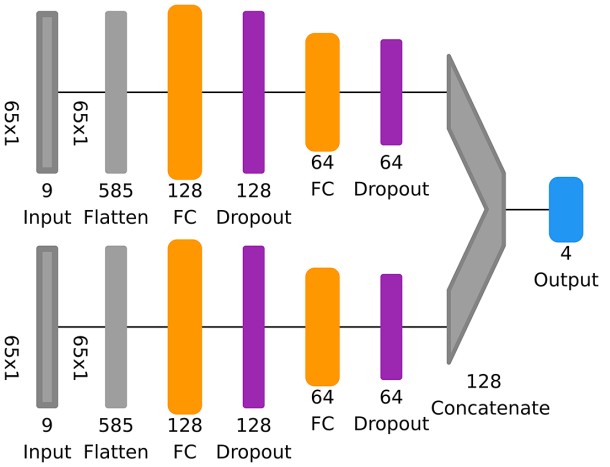

Figure D1. *Visualization of MB-FCN-LT/ST using the tool Net2Vis (Bauerle et al., 2021). Shown from left to right are the input data, followed by the flattened layer and two fully connected layers (FC) with 128 and 64 neurons. In total, the neural network has two such branches, whose weights can be trained independently of each other. All branches are concatenated and bounded by the output layer with four neurons. The orange FC block consists of a fully connected layer, a batch normalization layer, and an exponential linear unit activation. The output layer contains only a fully connected layer followed by a linear activation. The dropout layers are highlighted in purple, and all other remaining layers with nontrainable parameters are shown in gray.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

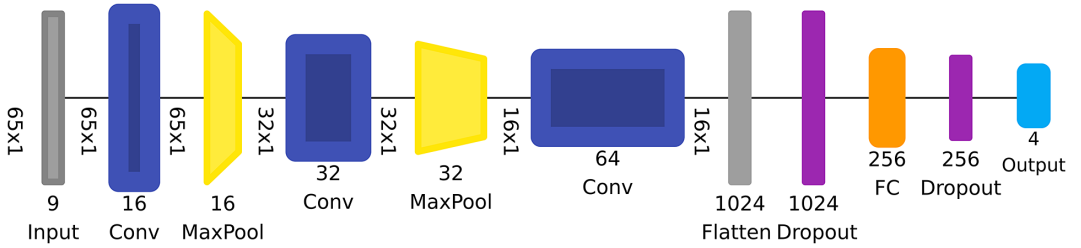

Figure D2. *Visualization of a convolutional neural network as in Figure D1. In addition, this neural network consists of convolutional blocks highlighted in blue and MaxPooling layers shown in yellow. Each convolutional block consists of a convolutional layer with a kernel size of 5 × 1 and the same padding, followed by a batch normalization layer and a parametric rectified linear unit (PreLU) activation. The MaxPooling layers use a pooling size of 2 × 1 and strides with 2 × 1. The FC blocks in this model consist of the fully connected layer, batch normalization, and a PReLU activation.*

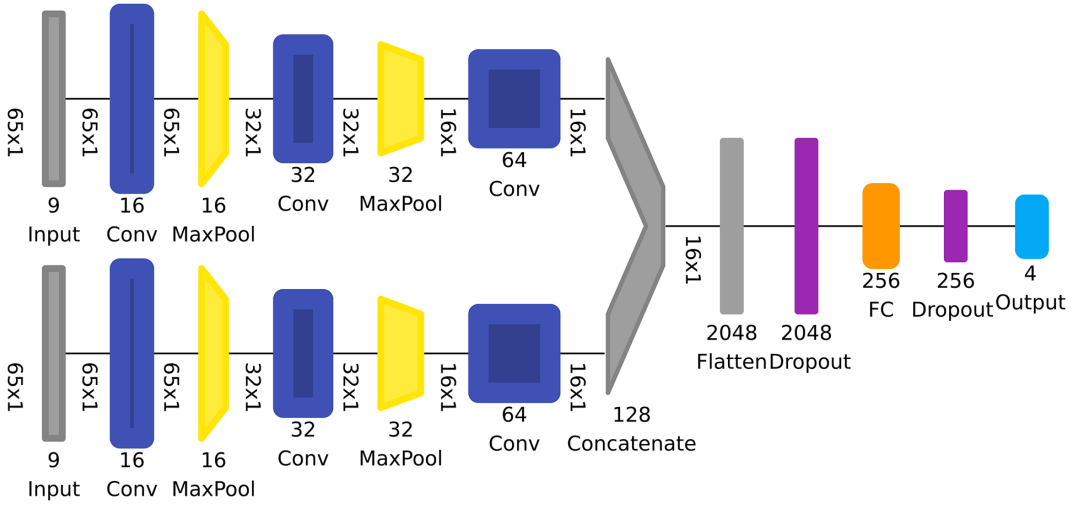

Figure D3. *Visualization of a multibranch convolutional neural network as in Figure D2.*

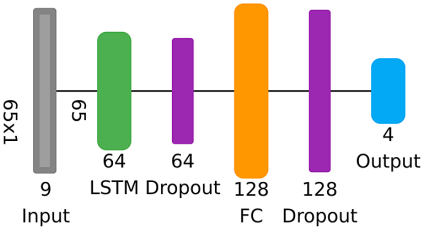

Figure D4. *Visualization of RNN as in Figure D1. In addition, the neural network shown here consists of long short-term memory layer (LSTM) blocks indicated in green. Each LSTM block includes an LSTM layer with a given number of LSTM cells within followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. Note that the dropout shown here is not the recurrent dropout, but the regular dropout that is applied on the activation of a layer. The FC block also uses a ReLU activation function, but no batch normalization.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

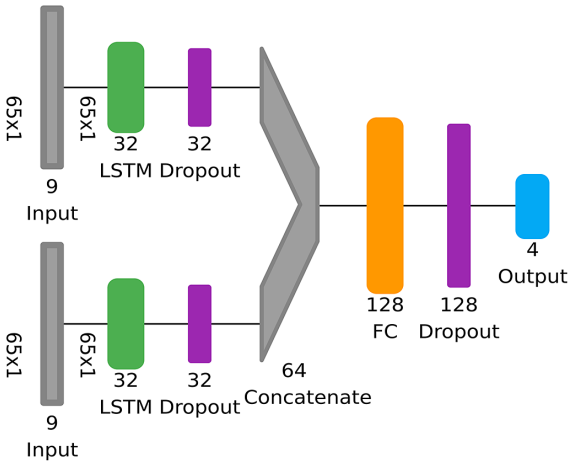

Figure D5. *Visualization of MB-RNN as in Figure D4. Deviating here, the activation is LeakyReLU both for the long short-term memory layer and the FC layer.*

Appendix E: Feature Importance of MB-CNN and MB-RNN

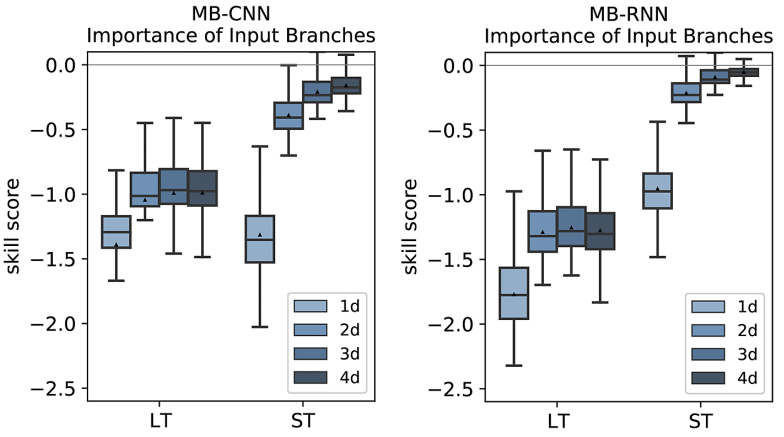

Figure E1. *Importance of single branches for multibranch convolutional neural network (left) and multibranch recurrent neural network (right) as in Figure 8.*

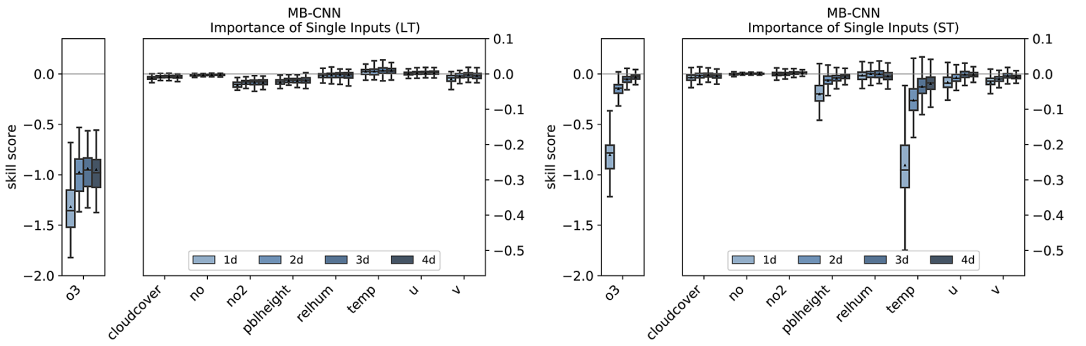

Figure E2. *Importance of single inputs for the LT branch (left) and the ST branch (right) for the multibranch convolutional neural network.*

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press

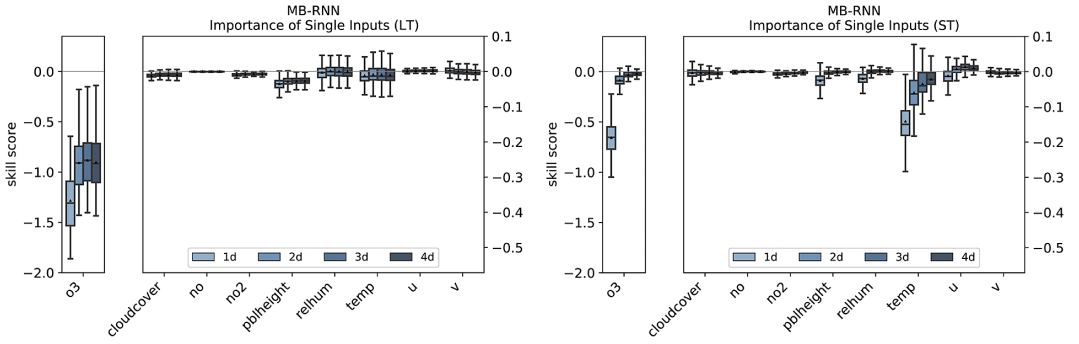

Figure E3. *Importance of single inputs for the LT branch (left) and the ST branch (right) for the multibranch recurrent neural network.*

Cite this article: Leufen LH, Kleinert F. and Schultz MG (2022). Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-hour average ozone prediction. *Environmental Data Science*, 1: e10. doi:10.1017/eds.2022.9

https://doi.org/10.1017/eds.2022.9 Published online by Cambridge University Press