https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

AQ-Bench: A Benchmark Dataset for Machine Learning on Global Air Quality Metrics

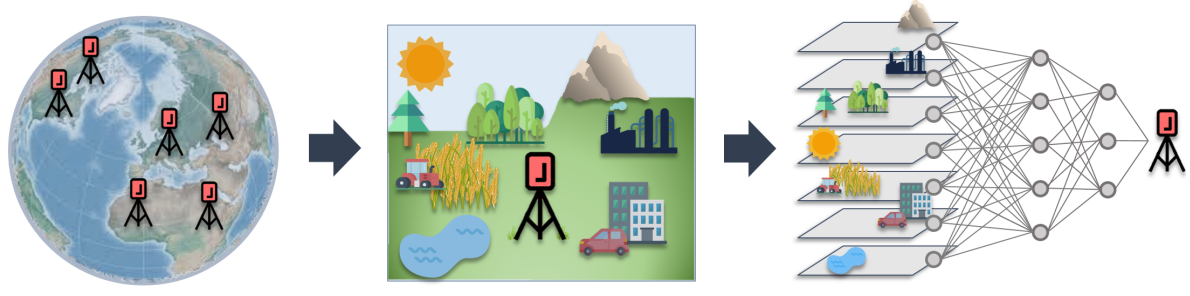
Clara Betancourt[1], Timo Stomberg[1], Scarlet Stadtler[1], Ribana Roscher[2], and Martin G. Schultz[1]

[1]Jülich Supercomputing Centre, Jülich Research Centre, Wilhelm-Johnen-Straße, 52425 Jülich, Germany
[2]Institute of Geodesy and Geoinformation, University of Bonn, Nußallee 17, 53115 Bonn, Germany

Correspondence: Martin G. Schultz (m.schultz@fz-juelich.de)

Abstract. With the AQ-Bench dataset, we contribute to the recent developments towards shared data usage and machine learning methods in the field of environmental science. The dataset presented here enables researchers to relate global air quality metrics to easy-access metadata and to explore different machine learning methods for obtaining estimates of air quality based on this metadata. AQ-Bench contains a unique collection of aggregated air quality data from the years 2010-2014 and metadata at more than 5500 air quality monitoring stations all over the world, provided by the first Tropospheric Ozone Assessment Report (TOAR). It focuses in particular on metrics of tropospheric ozone, which has a detrimental effect on climate, human morbidity and mortality, as well as crop yields. We validate these data as a machine learning benchmark by providing a well-defined task together with a suitable evaluation metric. Baseline scores obtained from a linear regression method, a fully connected neural network and random forest are provided for reference. AQ-Bench offers a low-threshold entrance for all machine learners with an interest in environmental science and for atmospheric scientists who are interested in applying machine learning techniques. It enables them to start with a real-world problem relevant to humans and nature. The dataset and introductory machine learning code are available at https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f (Betancourt et al., 2020) and https://gitlab.version.fz-juelich.de/toar/ozone-mapping. AQ-Bench thus provides a blueprint for environmental benchmark datasets as well as an example for data re-use according to the FAIR principles.

Graphical abstract



The AQ-Bench dataset contains air quality metrics and metadata at sites around the globe.
Map © 2019-20, The GMT Developers.

The air quality at a site is influenced by its surroundings.

The proposed machine learning task is to train a machine learning algorithm, which maps from metadata to aggregated air quality metrics at measurement sites.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

1 Introduction

In recent years, machine learning has achieved remarkable success in areas such as pattern-, image- and speech- recognition by usage of increasing computing power, innovative algorithms and high data availability (Krizhevsky et al., 2012; Amodei et al., 2016; Silver et al., 2016). This aroused the interest of environmental scientists to explore the application of machine learning and data-driven methods in their fields. The strength to be exploited is the ability of machine learning algorithms to find complex relationships in large multivariate, inhomogeneous datasets (as described in Wise and Comrie, 2005; Porter et al., 2015, e.g.).

In air quality research, there is one pollutant which is especially challenging to track: tropospheric ozone, a toxic trace gas which harms human health, vegetation and also impacts the climate (Cooper et al., 2014; Monks et al., 2015). Tropospheric ozone is difficult to track because it has no direct emission sources, but is produced as a secondary air-borne pollutant by several chemical reaction chains involving a large variety of precursors and photochemistry. With a lifetime of days to weeks (Wallace and Hobbs, 2006), the ozone concentration is affected by various physical and chemical processes which produce and destroy ozone. Therefore, ozone is a scientifically interesting candidate for machine learning applications: It is influenced by many interconnected environmental factors - and it is interesting to see if machine learning algorithms can learn these.

Data-driven atmospheric chemistry research was combined with machine learning from the late 1990s, to model and predict surface ozone concentrations in an alternative way to multivariate regression from (Yi and Prybutok, 1996; Comrie, 1997; Elkamel et al., 2001; Caselli et al., 2009). These data driven approaches take ground based measurements as input and predict the pollutant concentrations for the next days at individual locations. The principle behind recent machine learning applications in O_3 research is often a similar principle as Schultz et al. (2020) described for weather data: The input data are directly mapped to a specific data product, e.g. from meteorological and past O_3 measurements to the next day's maximum O_3 ozone value. In recent studies, Sayeed et al. (2020) and Kleinert et al. (2020) predicted regional ozone time series with convolutional neural networks and meteorological input data. Furthermore, Silva et al. (2019) trained a feed forward neural network to output ozone dry deposition at two forest measurement sites. Nevertheless, to our knowledge there are currently no machine learning projects that attempt analyzing and predicting ozone on the global scale, for longer time periods and with many kinds of metadata.

Developments in machine learning are accelerated by the existence of precompiled benchmark datasets, that allow machine learners to try out specific tasks, exchange solutions and compete with each other (LeCun et al., 2010; Deng et al., 2009; Rasp et al., 2020). Benchmarks can also be used for the development of explainable artificial intelligence approaches (Kierdorf et al., 2020; Roscher et al., 2020). So far, few of such benchmark datasets exist in the field of environmental science, especially related to air quality. While air quality data are in principle easily accessible from a variety of archives, there is often incomplete information and insufficient metadata to develop useful machine learning applications from this data. Furthermore, harmonization of such data from different sources, which is needed to achieve a global picture of ozone air pollution, is a difficult and time-consuming task.

With the AQ-Bench dataset, we aim to fill this gap and provide a dataset of global air quality metrics and metadata compiled from the TOAR database (Tropospheric Ozone Assessment Report, Schultz et al. (2017)). To make these data usable for

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

machine learning developments, this paper also describes the specific task of mapping between the metadata and the air quality metrics. Our ready-to-use, fully documented dataset is freely available under the DOI https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f (Betancourt et al., 2020). We also provide our baseline machine learning code at https://gitlab.version.fz-juelich.de/toar/ozone-mapping, offering a low-threshold entrance to machine learning in environmental science within a relevant research topic. In Sect. 2 of this paper we present the main factors affecting tropospheric ozone as the scientific background for the design of the AQ-Bench data set. Section 3 introduces the TOAR data products from which AQ-Bench was constructed. In Sect. 4, we describe the dataset itself. Section 5 contains the machine learning task for AQ-Bench and three baseline experiments to evaluate the applicability of these data in the machine learning context. We discuss opportunities and challenges of AQ-Bench, and give problem-related expected difficulties in Sect. 6. Information on data and code availability is given in Sect. 7, followed by a conclusion in Sect. 8.

2 What factors influence ozone?

Ozone (O_3) is a toxic greenhouse gas. While stratospheric ozone protects life on the planet's surface from ultraviolet radiation, tropospheric ozone is detrimental to human health, vegetation and climate. The AQ-Bench dataset and this paper focus exclusively on tropospheric ozone, more precisely the near-surface ozone to which humans, animals and plants are exposed. Ozone is a secondary pollutant that is formed from emissions of precursor substances and undergoes a variety of physical and chemical processes during its atmospheric lifetime. Figure 1 summarizes these processes, and is further elaborated in the following Subsections. How the described processes translate into the data in AQ-Bench, is described in the dataset description (Sect. 4).

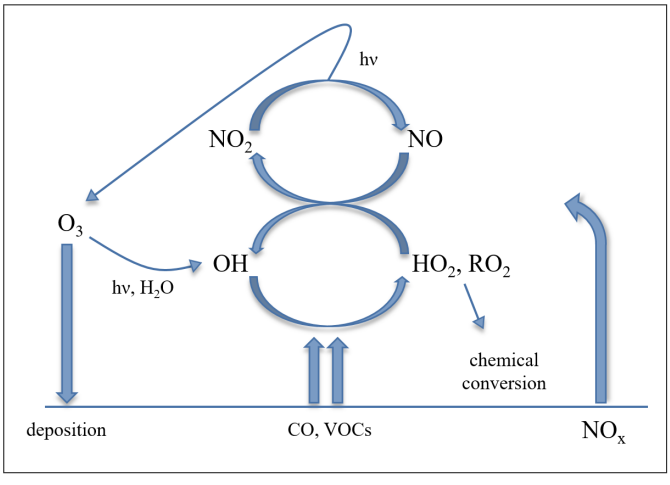


Figure 1. Simplified scheme describing the ozone chemical cycle. Figure adapted and modified from Jacob (2000). See text for elaboration.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

2.1 Precursor emissions

The most important ozone precursors are nitrogen oxides, carbon monoxide and volatile organic compounds (denoted as NO_x, CO and VOCs in Figure 1; note that $NO_x = NO_2 + NO$). Many of these precursors are emitted by human activities, e.g. from traffic, industry and agriculture (Benkovitz et al., 1996; Field et al., 1992). NO_x concentrations resulting primarily from combustion processes are especially high at very heavily polluted sites such as in city centers or near power plants. Industrial and traffic pollution are closely related to energy consumption depending on population density and economical activities. Agriculture machinery emits similar trace gases as traffic or industry. Moreover, agricultural plants are often fertilized, which adds more trace gas emissions (Veldkamp and Keller, 1997). In addition to emissions from human activities, several processes in nature also lead to emissions, especially of VOC compounds. For example, plants emit VOCs which are often more reactive (and could therefore produce more ozone) than VOCs emitted from human activities. The exact emission patterns vary among the types of plants and are thus related to land cover. I.e. agriculture fields, forests and grasslands will yield different magnitudes and seasonal cycles of VOC emissions (Simpson et al., 1999). Emissions can also occur from oceans, barren land and snow or ice covered surfaces. For example, the latter emit substantial quantities of NO_x in Arctic regions (Wang et al., 2007).

2.2 Ozone chemistry

The daily average ozone volume mixing ratios vary in the orders of magnitude from 10 to 100 ppbv (parts per billion by volume), with a lifetime of days to weeks (Wallace and Hobbs, 2006). Ozone has practically no direct emissions but is exclusively formed through atmospheric chemical reactions. The chemical processes leading to ozone formation are driven by ultraviolet radiation (denoted with $h\nu$ in Fig. 1). At wavelengths < 0.43 nm, photons convey enough energy to release chemical bonds in nitrogen dioxide (NO_2) molecules. This process (photo dissociation) leads to the formation of nitrogen oxide (NO) and a free oxygen radical (O). NO is also a radical and thus recombines quickly, while O collides with a high probability with O_2 and forms O_3. The produced O_3 is removed rapidly when it reacts with NO to $NO_2 + O_2$. The reactions form a null cycle, because O_3 is both created and destroyed. The cycle stabilizes at a certain O_3 concentration, depending on the available NO_2, ultraviolet light intensity and temperature. Up to a certain point, the ozone concentration rises with increasing NO_2 concentrations.

The dynamic equilibrium of this cycle can be altered by the presence of VOCs and CO (denoted as primary emissions in Fig. 1), which provide chemical pathways to convert NO to NO_2 without the destruction of O_3 by oxidation (oxidized pollutants denoted as HO_2 and RO_2 in Fig. 1). This leads to a non-linear system, where O_3 concentrations depend on the ratio of VOCs + CO and NO_x ($= NO + NO_2$) concentrations. During daytime, O_3 can photo dissociate and recombine with water vapor (H_2O in Fig. 1), thereby forming hydroxy radicals (OH in Fig. 2) which fuel a large share of atmospheric oxidation. There are several thousand chemical reactions occurring in the atmosphere, which need to be considered for an adequate description of ozone formation and loss processes and Fig. 1 only provides a very small glimpse on this rather complex system. For more details on ozone chemistry we refer to Brasseur et al. (1999).

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021


Open Access Earth System Science Data Discussions

2.3 Transport and loss processes

During its atmospheric lifetime, O_3 can be transported on spatial scales of hundreds or even thousands of kilometers (Schultz et al., 1999), until it is removed via atmospheric chemical reactions and deposition (indicated with downwards pointing arrows in Fig. 1). Primary chemical loss of O_3 is rather indirect via removal of NO_2 in polluted regimes and radical-radical reactions in clean environments with low NO_2 concentrations. Besides the chemical loss, O_3 can be removed by deposition on surfaces, especially on the leaves of natural or agricultural plants (Emberson et al., 2000). Ozone irreversibly damages plant tissue, when the plant leaves take it up, this leads to reduced crop yields (Mills et al., 2011) and tissue damage (Schraudner et al., 1997). Ozone deposition on water surfaces is relatively slow, but due to the large extent of them, this process also matters in the context of the global ozone budget (Luhar et al., 2018).

2.4 Interconnected factors

In the following, we describe exemplary how the influences of ozone precursor emission, chemistry, transport and loss (described in Sects. 2.1 - 2.3) can come together. The combination of chemistry and transport of air pollutants favors ozone formation downwind of sites with high precursor exhaust. A typical example are summertime rural areas downwind of larger city centers, where peak ozone values can often be observed (Xu et al., 2011). In the close vicinity of power plants or in city centers, NO_x is often very high and low ozone levels are observed (Sillman, 1999).

There are several geographical factors which determine the rates of chemical formation and loss of ozone. These factors can result in different mixes of ozone precursor emissions, varying reaction rates and varying rates of deposition. For example, the climate in a certain location determines the vegetation cover and the local weather. Since temperatures near the equator are high and more intense sunlight is available, ozone levels are generally higher there than near the poles. Moreover, at higher altitudes the air is generally cooler and drier, which leads to changes in reaction rates. Local flow patterns can also influence the ozone concentration, for example through the transport of air masses from valley to mountain tops (Kaiser et al., 2007).

Besides natural geographic factors, political decisions can also influence ozone formation. Many governments and decision makers worldwide strive to reduce air pollution by emission regulation, but these regulations differ between countries and may be implemented with more or less rigor. Ozone regulation is more difficult than that of primary air pollutants as one has to limit both VOC and NO_x emissions in order to control ozone, because of the chemical cycles described in Sect. 2.2.

Although ozone has a rather long lifetime, the local ozone concentration can change substantially in a matter of minutes and on scales of meters (e.g. in a street canyon), but it can also remain stable across hundreds of kilometers and for several weeks (e.g. at higher altitudes over the oceans). The "radius of influence" where ozone determined by near-by precursor emissions and deposition surfaces is typically around 25 km (European Union, 2008) in mid-latitude areas. All in all, ozone concentrations measured at a station are determined by many interconnected influences from precursor emissions, land use / land cover and the local weather conditions. Many of these factors are poorly quantified and often the interconnections are not understood well yet (Schultz et al., 2017). With AQ-Bench and the machine learning task described below we want to explore a novel way of

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021


Open Access Earth System Science Data Discussions

using a multitude of geographical features to predict ground-level ozone levels around the world. The details of data selection are described in Sect. 4, while the machine learning task is provided in Sect. 5.1.

3 TOAR data products

The TOAR database (Schultz et al., 2017) was created in context of the Tropospheric Ozone Assessment Report (TOAR). It contains one of the world's largest collections of near-surface ozone measurements, gathered from public bodies, research institutions and air quality networks all over the world. TOAR data products enabled the first comprehensive global assessment of the tropospheric ozone distribution and trends (Schultz et al., 2017; Fleming et al., 2018; Gaudel et al., 2018; Lefohn et al., 2018; Chang et al., 2017; Young et al., 2018; Mills et al., 2018; Tarasick et al., 2019; Xu et al., 2020). In the spirit of FAIR data usage (Wilkinson et al., 2016), these data products are openly available via the JOIN graphical interface[1], a REST interface[2], and through the PANGAEA repository[3].

For the AQ-Bench dataset we selected and harmonized air quality metrics and metadata from TOAR (see Sect. 4 and Appendix A). This section therefore contains a description of these selected data products, introducing the concepts of metrics and metadata.

3.1 Air quality metrics

The TOAR database contains hourly ozone measurements, transmitted from air quality observation sites. These hourly data are usually aggregated into statistics or "metrics" for further analysis. Ozone metrics consolidate air quality properties of longer time series (e.g. a season or a year) in a single figure, which can then be directly used for a scientific assessment and in decision making. Longer aggregation periods also average out short term weather fluctuations. There are specific metrics for different areas of ozone impact assessments (respiratory and cardiovascular disease, vegetation damage, climate impacts) and control.

The JOIN web-service is connected to the TOAR database and provides more than 30 of the most frequently used metrics as data products, calculated on-demand from hourly data. Besides these specialized metrics, also basic statistics such as mean, median and percentiles are available in JOIN. In the context of evaluating air quality, the validity of reported ozone metrics hinges on the data capture. Typically, statistical aggregations (i.e. metrics) of air quality data can only be used for decisions on attainment or non-attainment of air quality standards, if at least 75 % of the (hourly) samples in a dataset were reported. In this sense, the validity of ozone metrics is tied to the data completeness and we will use the term "valid data" to indicate samples with sufficient coverage of accurate data. All metrics which are part of AQ-Bench, are listed in Table 2 of the next section. Documentation and further information on all available metrics including data capture criteria are available in Schultz et al. (2017) and Lefohn et al. (2018).

[1] https://join.fz-juelich.de/
[2] https://join.fz-juelich.de/services/rest/surfacedata/
[3] https://doi.org/10.1594/PANGAEA.876108

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

3.2 Station metadata

The TOAR database also contains geographical information on air quality measurement station locations, i.e. station metadata. Metadata gives background information on the measurement site, where the data was retrieved from, and thus enables to characterize the location. These metadata are collected from different sources and were primarily derived from satellite Earth observations. For a complete list of the available metadata attributes see Schultz et al. (2017) and the REST interface[2].

For the AQ-Bench dataset described in this paper, we selected metadata from the TOAR database which characterize measurement locations and their surroundings with respect to pollution-relevant properties as introduced in Sect. 2. They are listed in Table 1 of the next Section.

4 AQ-Bench dataset description

The AQ-Bench dataset consists of metadata and mean ozone metrics from the years 2010-2014 at 5577 measurement stations all over the world, compiled from the TOAR database. The point of interest is to determine the resulting ozone metrics (see Sect. 3.1) given all environmental influences (Sect. 2) represented by metadata (Sect. 3.2). Our contribution in data preparation is to pick metadata with expert knowledge, relate it to processes, and aggregate air quality data to metrics in a way that it is representative for long time periods and meaningful in a machine learning context.

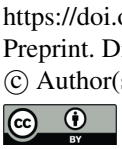
Three key points in the conception of this benchmark dataset are: 1) As targets, we use aggregated air quality metrics over five years. These are not influenced by short-term weather and emission forcings, but by site conditions on the climatological time scale. 2) Many known environmental influences on ozone are on short time scales (see Sect. 2), but we aim to predict long-term air quality conditions at the sites. Thus, we have identified which station metadata are the climatological representations of these short forcings. 3) We use a - to our knowledge unprecedented - variety of metadata that contains diverse information about environmental influences on the climatological scale. These metadata are sometimes not directly descriptive of the influences, but rather proxies for them. The benefits of machine learning must be taken to relate these proxies to air quality metrics.

This aggregated, climatological approach makes it possible to cover air quality data over a long period of time on the global scale with a relatively small and compact dataset. Yet, aggregated data accounts for long-term air quality conditions at a site, and daily or hourly influence on ozone variations not considered. Figure 2 gives an overview of all TOAR air quality monitoring stations included in AQ-Bench.

4.1 Station metadata

A summary of metadata in AQ-Bench is given in Table 1. The data originates from the TOAR database (Sect. 3), yet some data was harmonized, as documented in Appendix A. The metadata contains proxies for environmental influences on ozone on the climatological scale. In the following, we give two examples.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021



Open Access Earth System Science Data Discussions

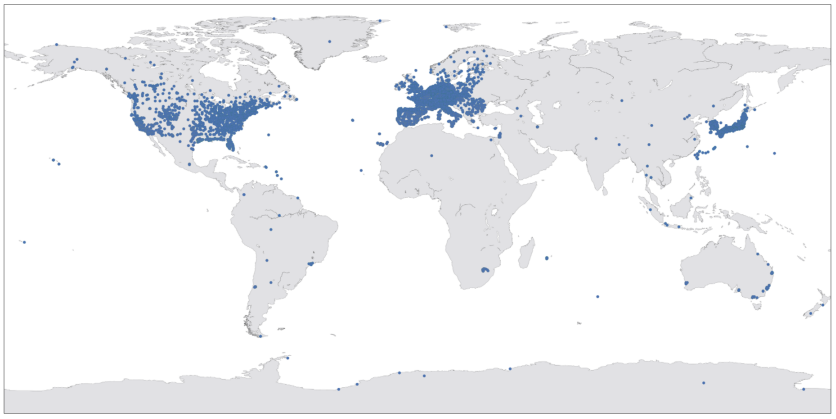


Figure 2. Worldwide measurement stations which are part of AQ-Bench, selected from the TOAR database. Map © 2019 - 20, The GMT Developers.

As mentioned in Sect. 2, ozone is influenced by weather. Likewise, ozone on longer time scales is influenced by climate. One variable in the AQ-Bench dataset is the *climatic zone* in which the site is located. The climate zone provides simplified information about climatic conditions at a location, for example, whether it is hot or cold, humid or dry, or of tropical climate.

A second example are ozone precursor emissions. In Sect. 2.1 we outlined that they are emitted by, for example, traffic and human activities. This means that the *population density* at a site is a good proxy for these activities. A second - more subtle - proxy is the *stable nightlight* at a location. It is the average intensity of light during night as seen from space, an indicator for industrial activity. In Sect. 2.2, we pointed out that ozone is often formed downwind of sites with high human and industrial activity. Therefore, in the AQ-Bench dataset, we do not only give *population density* and *stable night lights* at a site, but also related statistics of the closer surroundings. One example is the maximum *population density* in a radius of 5 km around the station.

All variables of the AQ-Bench dataset can be related to environmental impacts on the climatological time scale. We indicate the proxies in the right column of Table 1. Machine learning can make use of these proxies, even if they are not directly related to ozone concentrations.

Table 1: The station metadata of AQ-Bench.

Variable	Unit	Type	Proxy for...
Country	-	categorical	Emission regulation
HTAP region	-	categorical	World region set by the Task Force on Hemispheric Transport of Air Pollution http://htap.org

(continued on next page)

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

(Table 1 continued from previous page)

Variable	Unit	Type	Proxy for...
Climatic zone	-	categorical	Temperature, humidity, radiation
Longitude	deg	circular	-
Latitude	deg	continuous	Radiation, temperature
Altitude	m	continuous	Sinks, temperature
Relative altitude	m	continuous	Local flow patterns
Type	-	categorical	Industry / traffic emissions
Type of area	-	categorical	Proximity to human settlement
Water in 25km area	%	continuous	Deposition
Evergreen needle leaf forest in 25 km area	%	continuous	VOC Emissions, deposition
Evergreen broadleaf forest in 25 km area	%	continuous	VOC Emissions, deposition
Deciduous needle leaf forest in 25 km area	%	continuous	VOC Emissions, deposition
Deciduous broadleaf forest in 25 km area	%	continuous	VOC Emissions, deposition
Mixed forest in 25 km area	%	continuous	VOC Emissions, deposition
Closed shrub lands in 25 km area	%	continuous	VOC Emissions, deposition
Open shrub lands in 25 km area	%	continuous	VOC Emissions, deposition
Woody savannas in 25 km area	%	continuous	VOC Emissions, deposition
Savannas in 25 km area	%	continuous	VOC Emissions, deposition
Grasslands in 25 km area	%	continuous	VOC Emissions, deposition
Permanent wetlands in 25 km area	%	continuous	VOC Emissions, deposition
Croplands in 25 km area	%	continuous	Agricultural emissions
Urban And Built-Up in 25 km area	%	continuous	Human settlement
Cropland / Natural vegetation mosaic in 25 km area	%	continuous	Emissions, agriculture, deposition
Snow and ice in 25 km area	%	continuous	Factor on ozone formation
Barren or sparsely vegetated in 25 km area	%	continuous	Emissions, deposition
Wheat production	1000 Tons	continuous	Agricultural emissions
Rice production	1000 Tons	continuous	Agricultural emissions
NO_x emissions	$g\,m^{-2}\,y^{-1}$	continuous	NO_x emissions
NO_2 full column	10^5 molec cm^{-2}	continuous	NO_2
Population density	person km^{-2}	continuous	Human emissions
Max population density 5km	person km^{-2}	continuous	Human emissions near-by
Max population density 25km	person km^{-2}	continuous	Human emissions in area of influence
Nightlight 1 km	brightness index	continuous	Industrial activity

(continued on next page)

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

(Table 1 continued from previous page)

Variable	Unit	Type	Proxy for...
Nightlight 5 km	brightness index	continuous	Industrial activity near-by
Max nightlight 25 km	brightness index	continuous	Industrial activity in area of influence

4.2 Ozone metrics

The AQ-Bench dataset contains annually aggregated, averaged (years 2010-2014) ozone metrics as introduced in Sect. 3.1. There are therefore two steps involved in obtaining the metrics: 1) Getting up to five yearly metrics between 2010-2014 from hourly measurements, including data cover criteria to validate the metrics 2) averaging over these five years. If less than two yearly values are available, the value is considered missing. Missing values are denoted with -999 in the dataset. Some suspiciously high values were sorted out, as documented in Appendix A. The summary is given in Table 2.

Table 2: The ozone metrics of AQ-Bench

Metric	Description	Relevant field
Average values	Annual average value. No data capture criterion is applied, i.e. an average is valid if at least one hourly value is present.	Basic statistics
Daytime average	"Daytime average" is defined as average of hourly values for the 12-h period from 08:00 to 19:59 solar time. All hourly values in the aggregation period are averaged, and the resulting value is valid if at least 75 % of hourly values are present.	Basic statistics
Nighttime average	Same as "Daytime average" but accumulated over the daily interval from 20:00 to 07:59 solar time.	Basic statistics
Median	Median daily mixing ratio over one year. At least 10 hourly values must be present to accept a daily median value as valid.	Basic statistics
25 % percentile	25th-percentile of daily values in one year. At least 10 hourly values must be present to accept a daily percentile value as valid.	Basic statistics
75 % percentile	As "25 % percentile", but for the 75th-percentile.	Basic statistics
90 % percentile	As "25 % percentile", but for the 90th-percentile.	Basic statistics
98 % percentile	As "25 % percentile", but for the 98th-percentile.	Basic statistics

(continued on next page)

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

(Table 2 continued from previous page)

Metric	Description	Relevant field
dma8eu	Daily maximum 8-hour average statistics according to the EU definition. 8-hour averages are calculated for 24 bins starting at 17:00 local time of the previous day. The 8-h running mean for a particular hour is calculated on the concentration for that hour plus the following 7 hours. If less than 75% of the data are present (i.e. less than 6 hours), the average is considered missing. For annual aggregation, the 26th highest daily 8-hour maximum of the aggregation period will be computed. Note that in contrast to the official EU definition, a daily value is considered valid if at least one 8-hour average is present.	Human health
avgdma8epax	Average value of the daily "dma8epax" statistics during the aggregation period. "dma8epax" is the same as "dma8eu", but hourly bins start at 00:00 instead of 17:00.	Human health
drmdmax1h	Maximum of the 3-months running mean of daily maximum 1-hour mixing ratios during the aggregation period of one year.	Human health
W90	Daily maximum W90 5-h Experimental Exposure Index: EI = SUM($w_i C_i$) with weight $w_i = 1 / [1 + M \exp(-A\, C_i/1000)]$, where M = 1400, A = 90, and where C_i is the hourly average O_3 mixing ratio in units of ppb. For each day, 24 W90 indices are computed as 5-hour sums, requiring that at least 4 of the 5 hours are present (75%). If a sample consists of only 4 data points, a fifth value shall be constructed from averaging the 4 present mixing ratios. For annual aggregation, the 4th highest W90 value is computed, but only if at least 75% of days in this period have valid W90 values.	Vegetation
AOT40	Daily 12-h AOT40 values are accumulated using hourly values for the 12-h period from 08:00 until 19:59 solar time interval. AOT40 is defined as cumulative ozone above 40 ppb. If less than 75% of hourly values (i.e. less than 9 out of 12 hours) are present, the cumulative AOT40 is considered missing. When there exist 75% or greater data capture in the daily 12-h window, the scaling by fractional data capture ($n_{total}/n_{present}$) is utilized. For annual statistics, the daily AOT40 values are accumulated over the aggregation period and scaled by (n_{total}/n_{valid}) days. If less than 75% of days are valid, the value is considered missing.	Vegetation
nvgt70	Number of days with exceedance of the dma8epax value above 70 ppb. The value is marked as missing if less than 75% of days contain data.	Human health

(continued on next page)

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

(Table 2 continued from previous page)

Metric	Description	Relevant field
nvgt100	Number of days with exceedance of the daily max1h values above 100 ppb. The value is marked as missing if less than 75% of days contain data.	Human health

5 Validating AQ-Bench via machine learning

In this Section, we introduce the AQ-Bench dataset as a machine learning benchmark dataset. This means we combine the data documentation from the previous Section (Sect. 4) with the machine learning task for this dataset. We also provide an evaluation metric, a data split and baseline experiments.

5.1 Task description and evaluation metric

The task proposed for the AQ-Bench dataset is to train a machine learning model that maps from metadata in Table 1 to the ozone metric values in Table 2. This can be achieved with individual machine learning algorithms or in one multi-output algorithm.

The evaluation metric is the R^2, the coefficient of determination,

$$R^2 = 1 - \frac{\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}{\sum_{m=1}^{M}(y_m - \langle y \rangle)^2} \quad \text{with} \quad \langle y \rangle = \frac{1}{M}\sum_{m=1}^{M} y_m \tag{1}$$

where m denotes a sample index, M the total number of samples, $\hat{y}_m$ a predicted output value and y_m a reference target value.

R^2 measures the proportion of variance in the output values that the model predicts from the input values. A larger R^2 thus denotes a better model and the largest possible value is 1, or 100 %. We choose R^2 as it is comparable between all different targets, even if they cover different value ranges. The overall score of the solution is the mean of all scores achieved on the test set for all ozone metrics. We would like to challenge the machine learning and air pollution researchers to use this rather small dataset as efficiently as possible to extract all inherent information to accurately map onto the ozone metrics.

5.2 Data split

We provide a fixed data split within the AQ-Bench dataset to enable a comparison of our baseline results with future solutions, and to provide a suitable data setup for learning (see below). As it is good practice in machine learning, the dataset is split into three subsets for training, validation and hyperparameter tuning, and testing. The three data subsets are required to be independent, while having a similar statistical distribution to prevent the concealment of possible overfitting and an overestimation of accuracy. Because the dataset is relatively small, the split was chosen to be 60/20/20 %, as it is commonly used for datasets of this size. It is indicated in the dataset whether an example belongs to training, validation or test set.

In order to guarantee the spatial independence of the subsets, the data are divided into several spatial zones. The zones were created by spatial clustering, where stations are assigned to the same cluster if they are closer than 50 km. We chose 50 km, as

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

this is the distance in which the defined areas of influence of the stations do not overlap. Large station clusters were split again into smaller ones to ensure similar statistical distributions.

5.3 Baseline experiments

As baselines for machine learning approaches on the AQ-Bench dataset we present results obtained with three standard machine learning algorithms. For preprocessing, rows with missing values are dropped. Continuous metadata is scaled, each by a quantile range from 25 % to 75 % to avoid influence from outliers. Categorical metadata is one-hot encoded, resulting in 135 input features in total. We drop the *longitude* for our baseline experiments, since this is a circular variable and cannot be used without additional feature engineering. The preprocessed metadata is called input data in the following. Ozone metrics, which are the targets, are not scaled.

Methods:

- **Linear regression.** Linear regression models the simplest correlation between input and target values. It maps an input data example $\mathbf{x}_m$ with $\hat{y}_m = \mathbf{w}^\mathsf{T} \cdot \mathbf{x}_m + b$, where $\mathbf{w}$ and b are the regression parameters weights and bias. Vector $\mathbf{w} = [w_1, w_2, ..., w_N]^\mathsf{T}$ has the dimension of input vector $\mathbf{x}_m = [x_1, x_2, ..., x_N]^\mathsf{T}$.

- **Neural network.** We train a shallow fully connected neural network with two hidden layers of size 20 and 5 neurons, respectively. We use the Adam optimizer with MSE (mean squared error) loss function, L2 regularization and ReLU (rectified linear unit) as activation function (Goodfellow et al., 2016). Training is performed independently for each ozone metric. The hyperparameters for the individual targets are summarized in Appendix B. The model is written in Tensorflow/Keras (Chollet et al., 2015).

- **Random forest.** Our random forest model (Breiman, 2001) is built with a number of 100 trees for each target. We use the RandomForestRegressor of SciKit-learn (Pedregosa et al., 2011).

The baseline results are summarized in Table 3. Comparing the different models, random forest yields the best results for all targets except the *nvgt*-metrics, where the neural network performs best. The linear regression is the worst for all targets except e.g. *75 % percentile* where it is second best after the Random Forest. For some targets, e.g. *average values*, random forest is only slightly better than the neural network. However, there are targets, e.g. *AOT40*, where the gap between the two methods is almost 10 %. The neural network performs best for *nvgt070* and *nvgt100*. The baseline experiment results of *nvgt100* drops in comparison to other targets with partly negative R^2-scores. The results of *nvgt070* have the second least scores. These two targets count exceedances of a certain threshold, so that many values equal zero, which might be problematic to capture for standard machine learning algorithms. Except of those, R^2 is higher than 50 % for at least one of the three models per target. This shows that there is a quantitative relationship between input data and targets. Nevertheless, for our baseline experiments we used rather simple models, in order to proof the concept. Ozone, as a secondary pollutant with levels highly dependent on the environment and available precursors, is not captured perfectly by these simple baselines.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

Table 3: R^2-scores in %. Best results are marked in bold, second best results are underlined.

Target	Linear regression	Neural network	Random forest
Average values	53.69	58.25	**59.75**
Daytime average	55.93	56.26	**62.99**
Nighttime average	49.79	56.92	**59.00**
Median	52.21	56.67	**56.85**
25 % percentile	52.77	56.12	**62.75**
75 % percentile	51.75	45.92	**55.65**
90 % percentile	49.48	50.41	**58.54**
98 % percentile	47.68	54.89	**59.19**
dma8eu	49.32	54.95	**58.43**
avgdma8epax	54.76	58.23	**62.99**
drmdmax1h	40.21	50.12	**51.53**
W90	47.90	46.15	**51.29**
AOT40	45.88	50.91	**59.97**
nvgt70	26.38	**31.94**	30.53
nvgt100	-32.33	**12.51**	-66.57
Overall score	43.03	**49.35**	48.19
Overall score (excluding nvgt)	50.10	53.52	**58.38**

6 Discussion

6.1 Opportunities for machine learning in air quality research

With the AQ-Bench dataset, we used our knowledge on environmental influences on ozone, a toxic greenhouse gas, to bundle air quality data and metadata with machine learning approaches. By doing this, we enable a quick entry into machine learning in air quality research with reduced machine learning overhead. Our approach enables to use data from various sources that would otherwise be time consuming to acquire and prepare. We provide a ready to use dataset for the machine learning community, to support research on meaningful real-world applications (motivated by Wagstaff, 2012).

One great advantage of using machine learning for air quality research is the possibility to use data from various different sources, especially data which are not directly connected to air pollution via physical or biogeochemical models (e.g. stable nightlights). To explore this opportunity for ozone, we gathered an unprecedented variety of metadata to allow the machine learning approaches to obtain hints on the many interconnected, nonlinear influences, which determine ozone concentrations

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

(see Section 2). As the results from our baseline experiments show, the AQ-Bench dataset bears some potential to exploit these relations with machine learning methods.

Currently not many air pollution researchers use purely data-driven approaches for their studies. With AQ-Bench we offer a first data driven machine learning view on global tropospheric ozone. To achieve the global view, we use the JOIN web interface[4] of the TOAR data center, which provides customized data products from the TOAR database. As proposed by Schultz et al. (2020), our approach is to output the demanded metrics directly, and thus to obtain required data products directly from machine learning. Our dataset fits with the vision for benchmark datasets described by Ebert-Uphoff et al. (2017).

6.2 Limitations of AQ-Bench

AQ-Bench includes ozone metrics and metadata from 5577 stations and spans a time period of five years. The stations included in AQ-Bench are not distributed equally around the globe. The spatial coverage in most of the regions is low, except in USA, European countries and some regions of East Asia (Japan and South Korea). This raises the question of whether it is possible to generalize machine learning results to regions that are not included in the training data, even if they have similar input metadata. Possibly it may be necessary to use a combination of observational data and numerical models to achieve full global coverage (c.f. Chang et al. (2017)).

Measurement errors, interannual changes and drift result in noisy ozone metrics. Conversely, at least in the current version of AQ-Bench, the input metadata is fixed and has no temporal evolution, an assumption which we can make, because we average over five years of ozone metrics. It cannot be out ruled that within this time major environmental changes could have happened, e.g. settlements could grow or shrink during this time. This means, that metadata as given in AQ-Bench might not be valid for the whole time period of five years. The population density might have increased, the climate zone might have changed, and if a forest was cleared, for example, the land cover would have changed as well.

In Sect. 5.3 we gave some basic machine learning approaches to find a mapping between the metadata and the target ozone metrics. Since the complexity of the machine learning approach has to match the data complexity, we assume that the inaccuracies in our baselines arise from the complex relationships of ozone with the environment compared to the input dataset size and complexity of machine learning approaches. Through aggregation, we simplify the problem and make machine learning on the dataset easy - but this simplification also comes at the cost of introducing noise and uncertainties. A diversity of geographical data is available so additional variables could be identified to catch this complexity. Two examples are the distance to the next coast line or major roads in the close vicinity of a station. Furthermore, for a more complete description of ozone processes, time resolved data could be used.

6.3 Machine learning challenges arising from AQ-Bench

In order to provide some guidance on how the machine learning results could be improved compared to the standard machine learning methods applied in our baselines (Sect. 5.3), we briefly discuss some techniques here. One aspect to explore is feature engineering. Currently AQ-Bench includes for example the circular variable *longitude*, which cannot be accessed by the ma-

[4]https://join.fz-juelich.de/

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

chine learning algorithm without further feature engineering. Other variables could be accumulated, or transformed to improve machine learning results. See e.g. Duboue (2020) for an introduction to the topic. We hope that the research community will be creative in feature engineering.

Another aspect is multi-task learning. The baseline methods were performed independently for each ozone metric, but there may be a connection between them, as they all describe ozone pollution. Therefore, multi-task learning is a promising direction to exploit these connections. See Zhang and Yang (2017) for a review on this topic.

The baseline experiments show that extremes are sparse and thus difficult to catch. For example, the metric *nvgt070* which counts the days where maximum ozone exceeds 70 ppb (which happens at least once a year at approx. 75 % of the stations) gives acceptable results, but *nvgt100* of is not captured well. This is explained by the fact, that there are very few (< 25 %) stations which experience occasionally ozone values above 100 ppb. Extremes can be captured by imbalanced learning. See He and Garcia (2009) for a review on learning from imbalanced data.

7 Data and code availability

The AQ-Bench dataset is available in .csv format at https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f (Betancourt et al., 2020). To enable a machine learning quick start on the AQ-Bench dataset with reproduction of the baseline experiments, we also provide an introductory jupyter notebook on https://gitlab.version.fz-juelich.de/toar/ozone-mapping. To start it directly in your browser, click the button "launch on binder" in the readme of this repository.

8 Conclusions

In this paper, we introduced AQ-Bench as a benchmark dataset for machine learning on global air quality metrics. It allows to explore different machine learning methods on the real-world problem of air quality analyses. Specifically, the machine learning task is to map station metadata to air quality metrics at 5577 measurement stations around the globe. As data driven techniques for air quality research are emerging, we present a first benchmark dataset on the global scale. Following the vision of Ebert-Uphoff et al. (2017) to design benchmarks that bridge geoscience and data science, the key features of AQ-Bench are:

- **Active research area:** Ozone is a highly relevant and active field of research, as it harms living beings and the ecosystem. Ozone research benefits from making data available and developing data driven methods for ozone assessment.
- **Understandable context:** We introduced the complex mechanisms behind ozone formation as well as physical and chemical processes in Sect. 2, to make the scientific context of this dataset understandable to everyone, even without prior knowledge.
- **Impact on data science:** Since AQ-Bench is relatively small and thus easy to handle, it is suitable for beginners in programming. AQ-Bench can be trained in less than a minute on a common personal computer without GPUs, so one

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

can quickly iterate through different algorithms and configurations. Yet noise, the small size of the data set and the complicated underlying processes make it challenging to achieve satisfactory machine learning results on this dataset.

- **A means to evaluate success:** We propose R^2, the coefficient of determination, as an evaluation metric for AQ-Bench. It is a suitable metric because it measures the proportion of variance in the output values that the model predicts from the input values. It is comparable between all targets.
- **Quick start:** To start machine learning on AQ-Bench in a common browser, launch the "binder" in the following Git repository: https://gitlab.version.fz-juelich.de/toar/ozone-mapping. Running the introductory notebook on binder enables users to try out different training algorithms and hyperparameters directly in the browser.
- **Citability and reproducibility:** The dataset has a DOI, and the baseline experiments can be reproduced with the code that is openly available on Github (see Sect. 7).

We hope that the AQ-Bench dataset will help to advance data driven techniques in the field of air quality research, and form the basis for future experiments and research.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z.: Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin, in: Proceedings of The 33rd International Conference on Machine Learning, vol. 48 of *Proceedings of Machine Learning Research*, pp. 173–182, PMLR, New York, New York, USA, https://arxiv.org/abs/1512.02595, 2016.

Benkovitz, C. M., Scholtz, M. T., Pacyna, J., Tarrasón, L., Dignon, J., Voldner, E. C., Spiro, P. A., Logan, J. A., and Graedel, T.: Global gridded inventories of anthropogenic emissions of sulfur and nitrogen, Journal of Geophysical Research: Atmospheres, 101, 29 239–29 253, https://doi.org/10.1029/96JD00126, 1996.

Betancourt, C., Stomberg, T., Stadtler, S., Roscher, R., and Schultz, M. G.: AQ-Bench, B2SHARE, https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f, 2020.

Brasseur, G., Orlando, J. J., and Tyndall, G. S., eds.: Atmospheric chemistry and global change, Oxford University Press, Oxford, UK, 3 edn., 1999.

Breiman, L.: Random forests, Machine learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Caselli, M., Trizio, L., de Gennaro, G., and Ielpo, P.: A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model, Water, Air, and Soil Pollution, 201, 365–377, https://doi.org/10.1007/s11270-008-9950-2, 2009.

Chang, K.-L., Petropavlovskikh, I., Copper, O. R., Schultz, M. G., and Wang, T.: Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia, Elem Sci Anth, 5, 50, https://doi.org/10.1525/elementa.243, 2017.

Chollet, F. et al.: Keras, https://keras.io, 2015.

Comrie, A. C.: Comparing Neural Networks and Regression Models for Ozone Forecasting, Journal of the Air & Waste Management Association, 47, 653–663, https://doi.org/10.1080/10473289.1997.10463925, 1997.

Cooper, O. R., Parrish, D. D., Ziemke, J., Balashov, N. V., Cupeiro, M., Galbally, I. E., Gilge, S., Horowitz, L., Jensen, N. R., Lamarque, J.-F., Naik, V., Oltmans, S. J., Schwab, J., Shindell, D. T., Thompson, A. M., Thouret, V., Wang, Y., and Zbinden, R. M.: Global distribution and trends of tropospheric ozone: An observation-based review, Elementa: Science of the Anthropocene, 2, 29, https://doi.org/10.12952/journal.elementa.000029, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, Miami, FL, USA, https://doi.org/10.1109/CVPR.2009.5206848, 2009.

Duboue, P.: The Art of Feature Engineering: Essentials for Machine Learning, Cambridge University Press, 1 edn., https://doi.org/10.1017/9781108671682, 2020.

Ebert-Uphoff, I., Thompson, D. R., Demir, I., Gel, Y. R., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E., and Smyth, P.: A vision for the development of benchmarks to bridge geoscience and data science, in: Proceedings of the 7th International Workshop on Climate Informatics, Boulder, CL, USA, 2017.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Elkamel, A., Abdul-Wahab, S., Bouhamra, W., and Alper, E.: Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach, Advances in Environmental Research, 5, 47–59, https://doi.org/10.1016/S1093-0191(00)00042-3, 2001.

Emberson, L., Ashmore, M., Cambridge, H., Simpson, D., and Tuovinen, J.-P.: Modelling stomatal ozone flux across Europe, Environmental Pollution, 109, 403–413, https://doi.org/10.1016/S0269-7491(00)00043-9, 2000.

European Union: 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, OJ L, 1–44, http://data.europa.eu/eli/dir/2008/50/oj, 2008.

Field, R., Goldstone, M., Lester, J., and Perry, R.: The sources and behaviour of tropospheric anthropogenic volatile hydrocarbons, Atmospheric Environment. Part A. General Topics, 26, 2983–2996, https://doi.org/10.1016/0960-1686(92)90290-2, 1992.

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

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

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y.: Deep learning, MIT press Cambridge, Cambridge, UK, 1 edn., 2016.

He, H. and Garcia, E. A.: Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21, 1263–1284, https://doi.org/10.1109/TKDE.2008.239, 2009.

Jacob, D. J.: Heterogeneous chemistry and tropospheric ozone, Atmospheric Environment, 34, 2131 – 2159, https://doi.org/https://doi.org/10.1016/S1352-2310(99)00462-8, 2000.

Kaiser, A., Scheifinger, H., Spangl, W., Weiss, A., Gilge, S., Fricke, W., Ries, L., Cemas, D., and Jesenovec, B.: Transport of nitrogen oxides, carbon monoxide and ozone to the alpine global atmosphere watch stations Jungfraujoch (Switzerland), Zugspitze and Hohenpeißenberg (Germany), Sonnblick (Austria) and Mt. Krvavec (Slovenia), Atmospheric Environment, 41, 9273–9287, https://doi.org/10.1016/j.atmosenv.2007.09.027, 2007.

Kierdorf, J., Garcke, J., Behley, J., Cheeseman, T., and Roscher, R.: What Identifies a Whale by its Fluke? on the Benefit of Interpretable Machine Learning for Whale Identification, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2, 1005–1012, 2020.

Kleinert, F., Leufen, L. H., and Schultz, M. G.: IntelliO3-ts v1.0: A neural network approach to predict near-surface ozone concentrations in Germany, Geoscientific Model Development Discussions, 2020, 1–69, https://doi.org/10.5194/gmd-2020-169, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25, edited by Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., pp. 1097–1105, Curran Associates, Inc., https://doi.org/10.1145/3065386, 2012.

LeCun, Y., Cortes, C., and Burges, C. J.: MNIST handwritten digit database, http://yann.lecun.com/exdb/mnist/, 2010.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Lefohn, A. S., Malley, C. S., Smith, L., Wells, B., Hazucha, M., Simon, H., Naik, V., Mills, G., Schultz, M. G., Paoletti, E., De Marco, A., Xu, X., Zhang, L., Wang, T., Neufeld, H. S., Musselman, R. C., Tarasick, D., Brauer, M., Feng, Z., Tang, H., Kobayashi, K., Sicard, P., Solberg, S., and Gerosa, G.: Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research, Elem Sci Anth, 6, 28, https://doi.org/10.1525/elementa.279, 2018.

Luhar, A. K., Woodhouse, M. T., and Galbally, I. E.: A revised global ozone dry deposition estimate based on a new two-layer parameterisation for air–sea exchange and the multi-year MACC composition reanalysis, Atmospheric Chemistry and Physics, 18, 4329–4348, https://doi.org/10.5194/acp-18-4329-2018, 2018.

Mills, G., Hayes, F., Simpson, D., Emberson, L., Norris, D., Harmens, H., and Büker, P.: Evidence of widespread effects of ozone on crops and (semi-) natural vegetation in Europe (1990–2006) in relation to AOT40-and flux-based risk maps, Global Change Biology, 17, 592–613, 2011.

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

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., Stevenson, D. S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., and Williams, M. L.: 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, 8889–8973, https://doi.org/10.5194/acp-15-8889-2015, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf, 2011.

Porter, W. C., Heald, C. L., Cooley, D., and Russell, B.: Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression, Atmospheric Chemistry and Physics, 15, 10 349–10 366, https://doi.org/10.5194/acp-15-10349-2015, 2015.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A benchmark dataset for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, https://doi.org/10.1029/2020MS002203, 2020.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J.: Explainable machine learning for scientific insights and discoveries, IEEE Access, 8, 42 200–42 216, https://doi.org/10.1109/ACCESS.2020.2976199, 2020.

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

Schraudner, M., Langebartels, C., and Sandermann, H.: Changes in the biochemical status of plant cells induced by the environmental pollutant ozone, Physiologia Plantarum, 100, 274–280, https://doi.org/10.1111/j.1399-3054.1997.tb04783.x, 1997.

Schultz, M. G., Jacob, D. J., Wang, Y., Logan, J. A., Atlas, E. L., Blake, D. R., Blake, N. J., Bradshaw, J. D., Browell, E. V., Fenn, M. A., et al.: On the origin of tropospheric ozone and NOx over the tropical South Pacific, Journal of Geophysical Research: Atmospheres, 104, 5829–5843, https://doi.org/10.1029/98JD02309, 1999.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O., Galbally, I., Petropavlovskikh, I., Von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., Naja, M., Seguel, R., Dauert, U., Eckhardt, P., Feigenspahn, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Christian Kjeld, P., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolla, R., Wang, T., Sharps, K., Adame, J. A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M. Y., Gros,

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

V., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Huber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L., McClure-Begley, A., Mohamad, M., Murovic, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xu, X., Xue, L., and Zhiqiang, M.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem Sci Anth, 5, 58, https://doi.org/10.1525/elementa.244, 2017.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., and Stadtler, S.: Can Deep Learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A, (accepted article), 2020.

Sillman, S.: The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments, Atmospheric Environment, 33, 1821–1845, 1999.

Silva, S. J., Heald, C. L., Ravela, S., Mammarella, I., and Munger, J. W.: A Deep Learning Parameterization for Ozone Dry Deposition Velocities, Geophysical Research Letters, 46, 983–989, https://doi.org/10.1029/2018GL081049, 2019.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484, https://doi.org/10.1038/nature16961, 2016.

Simpson, D., Winiwarter, W., Börjesson, G., Cinderby, S., Ferreiro, A., Guenther, A., Hewitt, C. N., Janson, R., Khalil, M. A. K., Owen, S., et al.: Inventorying emissions from nature in Europe, Journal of Geophysical Research: Atmospheres, 104, 8113–8152, https://doi.org/10.1029/98JD02747, 1999.

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

Veldkamp, E. and Keller, M.: Fertilizer-induced nitric oxide emissions from agricultural soils, Nutrient Cycling in Agroecosystems, 48, 69–77, https://doi.org/10.1023/A:1009725319290, 1997.

Wagstaff, K.: Machine learning that matters, arXiv preprint, https://arxiv.org/abs/1206.4656, 2012.

Wallace, J. and Hobbs, P.: Atmospheric Science: An Introductory Survey: Second Edition, vol. 92 of *International Geophysics Series*, Elsevier Academic Press, Burlington, MA, USA, 2 edn., 2006.

Wang, Y., Choi, Y., Zeng, T., Davis, D., Buhr, M., Huey, L. G., and Neff, W.: Assessing the photochemical impact of snow NOx emissions over Antarctica during ANTCI 2003, Atmospheric Environment, 41, 3944–3958, https://doi.org/10.1016/j.atmosenv.2007.01.056, 2007.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Wise, E. K. and Comrie, A. C.: Extending the Kolmogorov–Zurbenko filter: application to ozone, particulate matter, and meteorological trends, Journal of the Air & Waste Management Association, 55, 1208–1216, https://doi.org/10.1080/10473289.2005.10464718, 2005.

Xu, J., Ma, J., Zhang, X., Xu, X., Xu, X., Lin, W., Wang, Y., Meng, W., and Ma, Z.: Measurements of ozone and its precursors in Beijing during summertime: impact of urban plumes on ozone pollution in downwind rural areas., Atmospheric Chemistry & Physics Discussions, 11, 2011.

Xu, X., Lin, W., Xu, W., Jin, J., Wang, Y., Zhang, G., Zhang, X., Ma, Z., Dong, Y., Ma, Q., Yu, D., Li, Z., Wang, D., and Zhao, H.: Tropospheric Ozone Assessment Report: Long-term changes of regional ozone in China: implications for human health and ecosystem impacts, Elem Sci Anth, 7, 13, https://doi.org/10.1525/elementa.409, 2020.

Yi, J. and Prybutok, V. R.: A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, Environmental pollution, 92, 349–357, 1996.

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

Zhang, Y. and Yang, Q.: A survey on multi-task learning, arXiv preprint, https://arxiv.org/abs/1707.08114, 2017.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021


Open Access Earth System Science Data Discussions

Appendix A: Data editing

Some data from TOAR/JOIN were modified in order to make it more understandable and user friendly.

- *HTAP region* was updated according to the number code:

#	Replaced with	Description
2	OCN	Non-arctic/Antarctic Ocean
3	NAM	US + Canada (up to 66° N; polar circle)
4	EUR	Western + Eastern EU + Turkey (up to 66° N; polar circle)
5	SAS	South Asia: India, Nepal, Pakistan, Afghanistan, Bangladesh, Sri Lanka
6	EAS	East Asia: China, Korea, Japan
7	SEA	South East Asia
8	PAN	Pacific, Australia + New Zealand
9	NAF	Northern Africa + Sahara + Sahel
10	SAF	Sub Saharan / sub Sahel Africa
11	MDE	Middle East: S. Arabia, Oman, Iran, Iraq etc.
12	MCA	Mexico, Central America, Caribbean, Guyana, Venezuela, Columbia
13	SAM	South America
14	RBU	Russia, Belarus, Ukraine
15	CAS	Central Asia
16	NPO	Arctic Circle (North of 66° N) + Greenland
17	SPO	Antarctic

- *Climatic zone* was updated according to the number code:

#	Replaced with
1	warm_moist
2	warm_dry
3	cool_moist
4	cool_dry
5	polar_moist
6	polar_dry
7	boreal_moist

(Table continued on next page)

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

(Table continued from previous page)

#	Replaced with
8	boreal_dry
9	tropical_montane
10	tropical_wet
11	tropical_moist
12	tropical_dry

– The variable *type* was harmonized, as there are some types which appear only once or twice. These types were replaced with the category they go best with:
- "agricultural", "commercial", "other-agricultural", "other-marine" were replaced with "other"
- "rural" was replaced with "background"
- "urban" was replaced with "unknown".
Five types remain: "background", "industrial", "traffic", "other" and "unknown".

– The variable *type_of_area* was harmonized in the same way as *type*:
- "alpine grasslands", "background", "forest" and "marine" were replaced with "unknown"
- "rural-nearcity" and "rural-regional" were replaced with "rural"
- "rural-remote" was replaced with "remote"
- "Urban" was replaced with "urban".
Five types of area remain: "rural", "urban", "suburban", "remote" and "unknown".

– The station with id 4587 was sorted out, because it was a remote background station in Romania which reported one of the highest *o3_average value* of all stations (65.5899 ppb), and had a low data coverage. We suspect these values are faulty.

– The station with id 4589 was sorted out because it reported a *max_population_density_5km* of ca. 1 million per square kilometer which we suspect is faulty.

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

Appendix B: Hyperparameters for baselines

Below are listed the hyperparameters for the neural network training in Sect. 5.3.

Target	Learning rate	L2 lambda	Batch size	Epochs
Average values	1.0E-04	1.0E-02	32	250
Daytime average	1.0E-04	1.0E-02	32	250
Nighttime average	1.0E-04	1.0E-02	32	250
Median	1.0E-04	1.0E-02	32	250
25 % percentile	1.0E-03	1.0E-02	64	100
75 % percentile	1.0E-03	1.0E-02	256	250
90 % percentile	1.0E-03	1.0E-02	256	250
98 % percentile	1.0E-03	1.0E-02	256	250
dma8eu	1.0E-03	1.0E-02	128	250
avgdma8epax	1.0E-04	1.0E-02	32	250
drmdmax1h	2.0E-04	1.0E-02	32	150
W90	1.0E-04	1.0E-02	32	250
AOT40	1.0E-02	1.0E-02	128	250
nvgt070	1.0E-04	1.0E-02	32	150
nvgt100	1.0E-05	1.0E-02	32	200

https://doi.org/10.5194/essd-2020-380
Preprint. Discussion started: 14 January 2021

Open Access Earth System Science Data Discussions

Author contributions. CB and TS prepared the dataset, developed the software and conducted the baseline experiments. CB, SS and TS prepared the initial manuscript draft. RR and MGS reviewed and edited the manuscript. All authors read and approved the manuscript. RR and MGS supervised the project. CB coordinated the project.

Competing interests. MGS is topic editor of the ESSD Journal.

Acknowledgements. CB, SS and MGS acknowledge funding from ERC-2017-ADG#787576 (IntelliAQ). TS and MGS are partly funded through the program Supercomputing & Big Data of the Helmholtz Association's research field Key Technologies. The authors gratefully acknowledge the computing resources granted by Jülich Supercomputing Centre (JSC). The Graphical Abstract of this paper was designed with icons from Flaticon (https://www.flaticon.com/).