Institut für Geodäsie und Geoinformation

Mapping and Interpolation of Tropospheric Ozone Data with Machine Learning Methods

Dissertation
zur Erlangung des Grades
Doktorin der Ingenieurwissenschaften (Dr.-Ing.)
der Landwirtschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Clara Betancourt
aus
Köln

Bonn 2023

Referent:
PD Dr. Martin Schultz, Forschungszentrum Jülich GmbH

1. Korreferentin:
Prof. Dr. Ribana Roscher, Forschungszentrum Jülich GmbH

2. Korreferent:
Prof. Dr. Jürgen Kusche, Rheinische Friedrich-Wilhelms-Universität Bonn

Tag der mündlichen Prüfung: 28. August 2023

Angefertigt mit Genehmigung der Landwirtschaftlichen Fakultät der Universität Bonn

Kurzfassung

Mapping und Interpolation von troposphärischen Ozondaten mit Machine Learning Methoden

Troposphärisches Ozon ist ein giftiges Spurengas in der Atmosphäre. Es schadet der menschlichen Gesundheit, Nutzpflanzen und Vegetation, und ist ein kurzlebiges Treibhausgas. Ozon ist ein sekundärer Luftschadstoff, der in der Atmosphäre zahlreiche physikalische und chemische Prozesse auf unterschiedlichen Zeitskalen durchläuft. Wie bei vielen anderen Umweltvariablen ist es daher schwierig, dort Ozonkonzentrationen zu quantifizieren, wenn keine Messungen verfügbar sind. Um dieses Problem zu lösen, ist das Ziel dieser Arbeit die Entwicklung von räumlich-zeitlichen Mapping- und Interpolationsmethoden unter Verwendung von Techniken des Machine Learning am Beispiel von Ozondaten. Wir trainieren die Machine Learning Modelle auf Ozonmesswerten, die in der Datenbank des Tropospheric Ozone Assessment Report (TOAR) verfügbar sind. Die wichtigsten Beiträge dieser Arbeit sind:

- **Mapping und Interpolation von Ozondaten, um hochauflösende, hochpräzise raum-zeitliche Datenprodukte zu liefern.** Die Datenprodukte decken räumliche Bereiche von der regionalen bis zur globalen Ebene ab, und ihre zeitliche Auflösung reicht von stündlichen Daten bis zu mehrjährigen Statistiken. Wir verwenden große Datensätze mit Ozonmesswerten, kombiniert mit Modelldaten und Geodaten, um die Datenprodukte zu erstellen.
- **Anpassung, Entwicklung und Erläuterung neuer Machine Learning Methoden, die wir zur Erstellung dieser Datenprodukte verwenden.** Die wichtigsten Algorithmen dieser Arbeit basieren auf Entscheidungsbäumen und Graphen. Zum Beispiel entwickeln wir eine Evaluierungstechnik auf unterschiedlichen Skalen für räumliche Machine Learning Modelle und überprüfen ihre physikalische Konsistenz mit Hilfe von Shapley-Werten.
- **Nutzung von raum-zeitlichen Mustern in Geodaten und Ozonmessungen in Machine Learning Modellen.** Wir verwenden aggregierte lokale bis regionale geospatiale Daten in Machine Learning Modellen. Außerdem wenden wir einen Machine Learning Algorithmus an, der Ozonmessungen an unregelmäßig angeordneten Stationen verarbeiten kann.

Mit dieser Arbeit veröffentlichen wir AQ-Bench, einen Benchmark-Datensatz für Machine Learning auf globalen Langzeit Ozonmetriken. Wir verknüpfen erklärbares Machine Learning auf AQ-Bench mit Unsicherheitsbewertungen, um die Grenzen des Datensatzes und die Anwendbarkeit der resultierenden Machine Learning Modelle aufzuzeigen. Mit den trainierten Modellen erstellen wir auch die erste vollständig datengetriebene, globale, hochauflösende Karte von Langzeit-Ozonmetriken (Auflösung: 0,1°, Jahre: 2010 bis 2014). Wir entwickeln auch eine graphbasierte Methode zur Interpolation fehlender Daten für Ozonmessungen. Die Methode hat einen Index of Agreement von 0,96 - 0,99 für die stündliche Interpolation fehlender Messdaten in Deutschland.

Die Synthese dieser Arbeit ist, dass ein Zusammenspiel von physikalisch fundierter Datenauswahl, Unsicherheitsquantifizierung und Erklärbarkeit beim Machine Learning zuverlässige Umweltdatenprodukte erzeugen kann. Wir haben auch festgestellt, dass die Genauigkeit der Datenprodukte in einer bestimmten Region hauptsächlich von einer guten Abdeckung mit Ozonmessungen in dieser Region abhängt. Daher trägt diese Arbeit nicht nur zur lückenlosen Quantifizierung von Ozonkonzentrationen bei, sondern auch zum Machine Learning in den Umweltwissenschaften im Allgemeinen.

Abstract

Mapping and Interpolation of Tropospheric Ozone Data with Machine Learning Methods

Tropospheric ozone is a toxic trace gas in the atmosphere. It threatens human health, damages crops and vegetation, and it is a short-lived climate forcer. Ozone is a secondary air pollutant that undergoes multiple physical and chemical processes on a wide range of timescales. Therefore, as with many environmental variables, it is difficult to quantify ozone concentrations where measurements are not available. To solve this problem, the goal of this work is to develop spatio-temporal mapping and interpolation methods using machine learning techniques with the example application of ozone data. We train the machine learning models on a large number of ozone measurements available in the Tropospheric Ozone Assessment Report (TOAR) database. The most important contributions of this work are:

- **Mapping and interpolating ozone data, providing high-resolution, high-accuracy, spatiotemporal data products.** The data products cover spatial domains from the regional to the global level, and their temporal resolution ranges from hourly data to multi-year statistics. We use large quantities of ozone measurements, combined with model data and geospatial data to generate the data products.
- **Adapting, developing, and explaining new state-of-the-art machine learning methods that we use to create these data products.** The most relevant algorithms of this work are tree-based and graph-based methods. For example, we develop a multi-scale evaluation technique for spatial machine learning models and verify their physical consistency by using Shapley additive explanations.
- **Utilizing spatiotemporal patterns in geospatial data and ozone measurements in machine learning models.** We use aggregated local to regional geospatial site conditions as input features for machine learning models. Furthermore, we adopt a graph machine learning algorithm to work on ozone measurements at irregularly placed air quality monitoring stations.

With this work, we publish AQ-Bench, a benchmark dataset for machine learning on global long-term ozone metrics. We link explainable machine learning on AQ-Bench with uncertainty assessments to point out limits in the dataset and the applicability of the resulting machine learning models. With the trained models, we also create the first completely data-driven, global, high-resolution map of long-term ozone metrics (resolution 0.1°$\times$ 0.1°, years 2010 - 2014). Finally, we develop a high-performance graph-based missing data interpolation method for ozone measurements. It has an index of agreement of 0.96 - 0.99 for hourly missing data interpolation in Germany.

The synthesis of this work is that an interplay of physically sound data selection, uncertainty quantification, and explainability in machine learning can produce trustworthy environmental data products. We also found that the accuracy of the data products in a specific region is mainly dependent on good coverage with ozone measurements in that region. Therefore, this work contributes not only to the gapless quantification of ozone concentrations but also to trustworthy machine learning in the environmental sciences.

Contents

1. Introduction

1.1 Atmospheric pollution and tropospheric ozone

The atmosphere is the gravitationally bound gas layer that surrounds the Earth. Its major constituents are nitrogen (N_2, $\sim$78 %), oxygen (O_2, $\sim$21 %), and argon (Ar, $\sim$1 %) (Wallace and Hobbs, 2006). The atmosphere also contains thousands of trace gases in smaller quantities of hundred parts per million (ppm) or less. Examples are carbon dioxide (CO_2), hydrogen (H_2), and ozone (O_3). Some of these trace gases, such as ozone, are highly reactive and have spatio-temporally variable atmospheric concentrations (Wallace and Hobbs, 2006). Since the start of the agro-industrial era, the "Anthropocene", humans have substantially altered the atmosphere's composition through air pollution with trace gases from industry, agriculture, transportation, fossil fuel combustion, and biomass burning (Houghton et al., 2001; Brasseur et al., 2003). Depending on their physical and chemical properties, these air pollutants can have a major impact on the Earth system and therefore on humanity, for example if they are toxic or alter the Earth's radiation budget. In more detail, air pollution harms agricultural productivity, human health, climate, and the environment (Sun et al., 2017; World Health Organization, 2022; Masson-Delmotte et al., 2021; Singh and Agrawal, 2007). The mitigation of air pollution is therefore of global importance and directly related to the sustainable development goals "zero hunger", "good health and well-being", "climate action" and "life on land" published by the United Nations (United Nations, 2022).

Air pollution affects the Earth system through various processes, and in the following, we give some examples. One environmental impact is acid rain formed through oxides of sulfur and nitrogen emitted by industry and traffic. When exposed to sunlight, the oxides react with atmospheric water vapors and form sulfuric and nitric acid mists, which then descend as precipitation and severely damage vegetation and soils. Acid rain events were especially severe in Scandinavia in the 1960s but occurred around industrialized areas all over Europe, America, and Asia as well (Singh and Agrawal, 2007). The most widely recognized impact of air pollution on the climate is global warming caused by greenhouse gases. Since the beginning of industrialization, human activity has increased the global concentrations of methane (CH_4), nitrogen oxide (N_2O), and ozone (O_3). Like CO_2, these air pollutants have the ability to trap heat in the atmosphere and therefore increase the Earth's temperature. The total anthropogenic radiative forcing compared to pre-industrial times is estimated to be 2.72 W m^{-2} (Masson-Delmotte et al., 2021), and strong anthropogenic emission reduction is needed to limit the resulting global temperature increase to 1.5 °C (IPCC, 2022). A well-known example of the effect of air pollution on human health is photochemical smog, the so-called "Los Angeles smog". The term "smog" is a portmanteau composed of the words "smoke" and "fog". It refers to urban air pollution that forms under stable meteorological conditions, limits visibility and affects human health (Wallace and Hobbs, 2006). Los Angeles smog results from the photochemical conversion of primary air pollutants exhausted from sources such as automobile engines. It forms through chemical reactions of these pollutants due to stable meteorological conditions under sunlight (Tiao et al., 1975). Los Angeles smog is associated with high concentra-

tions of nitrogen oxides ($NO + NO_2 = NO_x$), hydrocarbons, and ozone (O_3). It irritates the lungs and eyes (Tiao et al., 1975).

This work is about ozone, a toxic, odorless atmospheric trace gas and secondary air pollutant. Ozone molecules consist of three oxygen atoms (chemical formula O_3). About 90 % of the total atmospheric ozone is in the stratosphere. Stratospheric ozone is considered the "good" ozone as it protects life on Earth from harmful ultraviolet radiation (Seinfeld and Pandis, 2006). The remaining 10 % of the total atmospheric ozone is in the troposphere. This ozone is considered the "bad" ozone due to its negative impact on human health, vegetation, and climate, as in the examples given above. Tropospheric ozone concentrations are usually between 10 - 100 ppb (parts per billion) at sea level (Wallace and Hobbs, 2006). This work focuses on tropospheric, near-surface ozone, i.e., the harmful ozone to which humans, animals, and plants are exposed. Therefore, when ozone is mentioned in this work, tropospheric ozone is meant. The following sections provide an introduction to ozone processes, impacts and ozone research relevant to this work.

1.1.1 Ozone processes

Ozone is a secondary air pollutant, which means it is not emitted directly. Instead, it is formed by photochemical processes involving precursors emitted from sources such as fossil fuel combustion, agriculture, and vegetation (Monks et al., 2015). Figure 1.1 shows a simplified scheme of atmospheric ozone processes and is described below. The most important ozone precursors are carbon monoxide, volatile organic compounds, and nitrogen oxides (CO, VOC, and NO_x in Figure 1.1). The main ozone chemical cycle fuelled by these precursors is depicted by the blue arrows in Figure 1.1. Photodissociation of nitrogen dioxide ($h\nu$ and NO_2 in Figure 1.1) produces oxygen radicals which quickly recombine with molecular oxygen (O_2) to ozone (O_3). The ozone then recombines rapidly with nitrogen dioxide to nitrogen oxide and molecular oxygen ($O_3 + NO_2 \rightarrow O_2 + NO$). These two reactions combine into a null cycle in which ozone is both produced and destroyed. It stabilizes at a certain ozone concentration, depending on the available precursors, solar light intensity, and temperature. Ozone can also photo dissociate to $O_2 + O$ and recombine with water vapor (H_2O in Figure 1.1), forming hydroxy radicals (OH). OH is associated with many atmospheric oxidation processes and is often referred to as the "detergent" of the atmosphere (Comes, 1994). Ozone is therefore involved in the formation of oxidized compounds (HO_2 and RO_2 in Figure 1.1) and a major driver of atmospheric oxidation. There are thousands of atmospheric chemical reactions that contribute to ozone formation and destruction, so Figure 1.1 only provides a simplified scheme. Brasseur et al. (1999) gives a more detailed description of ozone chemistry, and Sander et al. (2006) and the Task Group on Atmospheric Chemical Kinetic Data Evaluation[1] provide a compendium of atmospheric chemical reaction rates.

The red arrows in Figure 1.1 indicate other atmospheric processes that alter the ozone concentration. Influx from the stratosphere is an important source of tropospheric ozone (Seinfeld and Pandis, 2006). Furthermore, ozone transport within the troposphere ranges from small-scale diffusion to long-range advection from formation sites to remote areas (Schultz et al., 1999). Besides chemical conversion, ozone can also be lost through deposition. The main deposition surfaces are

[1] `https://iupac-aeris.ipsl.fr/`, last access 15 March 2023

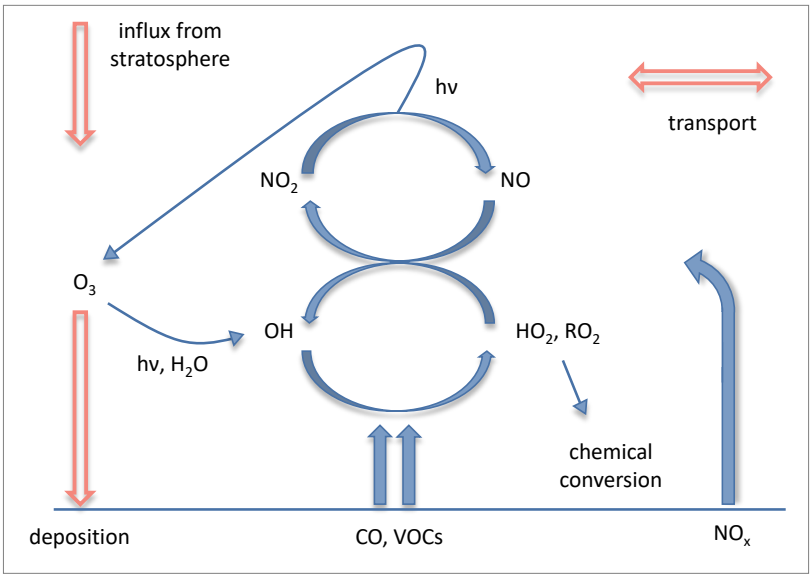


Figure 1.1: Ozone governing processes. Chemical processes are depicted by blue arrows and other processes red arrows. Figure adapted and modified from Betancourt et al. (2021). See text for elaboration.

soil, water, and the stomata of plants (Monks et al., 2015). While stomatal uptake is a major ozone sink on land, seawater is the largest dry deposition surface in absolute terms, with a loss of about 40 % of global ozone (Monks et al., 2015).

Ozone processes form a complex system, as many are nonlinear and interdependent. Below we describe some of the environmental spatio-temporal patterns that influence the ozone distribution. Local to regional ozone concentrations depend largely on precursor emissions, i.e., indirectly on land use and industrial or agricultural activity. Precursor emissions show daily, weekly, and annual cycles, for example, because traffic is heavier during the week. In addition, meteorological conditions such as radiation and wind are relevant. High ozone is frequently observed in suburban areas, downwind of city centers with high NO_x emissions, and under warmer temperatures (Xu et al., 2011). Ozone loss through deposition depends on the type of surface and land use. The resulting spatio-temporal ozone patterns are highly diverse. Ozone concentrations can sometimes be stable across spatiotemporal ranges of kilometers and days, for example, over the oceans. On the other hand, they can also change in a matter of meters or in a matter of seconds, for example, if a localized NO_x source triggers the fast reaction cycles described above. Many of the beforementioned ozone governing factors are poorly quantified, and their interconnection is still not understood well, making the ozone distribution especially hard to quantify and predict (Schultz et al., 2017; Archibald et al., 2020). This also makes ozone an ideal test case for machine learning methods like those we are developing in this work.

1.1.2 Ozone impacts

Ozone impacts human health, vegetation and the climate. Short- and long-term human exposure to elevated ozone concentrations damages lung tissue, causing inflammation and cardiovascular and respiratory disease (Henschel et al., 2013; World Health Organization, 2021). Fleming et al.

(2018) summarize clinical and epidemiological studies on these impacts. Public authorities such as the European Union (EU) and the World Health Organization (WHO) recognize this threat to human health and publish guidelines and enforce threshold values for ozone concentrations in living areas (European Union, 2008; World Health Organization, 2006). The WHO's guideline for 8 - hour mean ozone mass concentrations for human exposure is 100 μg m^{-3}.

Ozone also damages the vegetation when it enters the stomata of plants (Van Dingenen et al., 2009; Mills et al., 2018a; Mills et al., 2018b). How much damage the ozone does depends on the type of plant and the growing season. For example, Mills et al. (2018) show that soybeans are more vulnerable to ozone stress than rice. Meteorological factors also play a role, e.g., plants open their stomata when the air is humid and are therefore more sensitive to ozone when they are irrigated or under precipitation (Emberson, 2020). Crops that suffer from ozone injuries during their growing season exhibit reduced total biomass, yield, and flower number. Mills et al. (2018) estimate that ozone destroys 4 - 12 % of the world's annual corn, rice, soybean, and wheat yields, making it a risk factor for global food security.

Ozone is also a climate forcer as it absorbs long-wave (terrestrial) radiation. Its global surplus radiative forcing is estimated to be 0.39 Wm^{-2} (Skeie et al., 2020), which is about a quarter of the radiative forcing of carbon dioxide (CO_2). With a typical atmospheric lifetime of days to weeks (Wallace and Hobbs, 2006), ozone is a short-lived climate forcer. A reduction in ozone-producing factors can therefore attenuate global warming on much shorter time scales than mitigating the exhaust of long-lived climate forcers such as methane (CH_4) or CO_2 (Pachauri et al., 2014; Monks et al., 2015). It also has higher spatial variability than long-lived climate forcers.

1.1.3 Ozone monitoring

Long-term ozone monitoring is necessary to assess its oxidation cycles, short-term variability, climatological trends and impacts (Gaudel et al., 2018; Schultz et al., 2017; Tarasick et al., 2019). Public authorities evaluate ozone measurements to determine whether ozone guidelines are being met. Ozone measurements are usually reported as mass concentrations with the unit [μg m^{-3}] or mole fractions in parts per billion [ppb].

Schultz et al. (2015) list some challenges in measuring reactive trace gases, which also apply to ozone: First, its mole fractions are typically no higher than 100 ppb. This means that only 100 nanomoles of ozone are in one mole of air. Sophisticated instruments are therefore needed to measure ozone. Typically, ultraviolet absorption spectrometers are used. These instruments need to be maintained and calibrated, ideally on a daily basis, which is associated with high personnel expenditure. Second, as mentioned above, ozone is highly reactive, making sampling ozone and storing reference materials for calibration challenging. Third, ozone has a high spatiotemporal variability due to the inhomogeneity of precursor emissions and small-scale meteorology. This can limit the regional representativeness of measurement stations. In some situations, e.g. in a street canyon, a measuring station would be required approximately every meter to catch the ozone variability.

In many countries, near-surface ozone is measured by ground-based air quality observation stations operated by local environmental authorities. They usually provide hourly measurements,

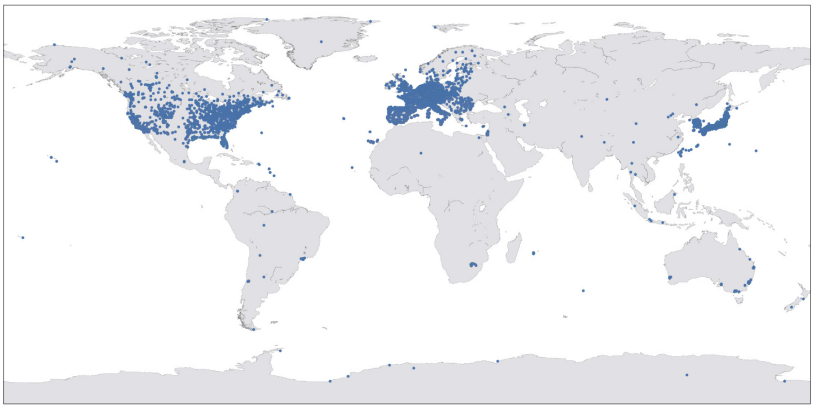


Figure 1.2: Location of ozone measurement stations in the TOAR database. The majority of the stations are located in Europe, the US, and East Asia. Figure adapted from Betancourt et al. (2021).

and the resulting time series often cover multiple years or even decades. The data are collected centrally, e.g., through the European Environmental Agency[2], or the US Air Quality System[3]. These networks comprise long-term data from thousands of stations on a continental scale. There are broader efforts to aggregate ozone data globally and make them openly available. Two examples are the data repositories of OpenAQ[4] and the Tropospheric Ozone Assessment Report (TOAR)[5]. The TOAR database (Schultz et al., 2017), which is the primary data source of this work, contains more than 12 000 hourly ozone time series from all over the globe, of which some are more than 40 years long. The TOAR data portal also provides ozone statistics and metrics. These are aggregations of hourly measurements that facilitate the analysis of ozone impacts. The most common metrics for health, vegetation, and climate impacts are available, as well as basic statistics such as averages and percentiles. Figure 1.2 shows the spatial distribution of the TOAR measurement stations. The database also provides geospatial metadata of these stations.

1.1.4 Machine learning for ozone research

Machine learning algorithms build models by fitting adaptive parameters based on sample data to make predictions without being explicitly programmed to do so (Samuel, 1959[6]). Machine learning can solve numerous tasks in image recognition, speech recognition, gaming, and video prediction (Krizhevsky et al., 2012; Silver et al., 2016; Mathieu et al., 2015; Zhou et al., 2023). The ability of machine learning to capture complex patterns in data has sparked motivation among environmental scientists to apply machine learning in their respective research areas (Hsieh, 2009; Liu et al., 2022; Haupt et al., 2022). This Section contains an introductory review of machine learning for ozone research. Section 2.2 introduces the technical aspects of machine learning which are relevant to this work.

[2] `https://www.eea.europa.eu/data-and-maps/data/aqereporting-9`, last access 29 August 2022

[3] `https://www.epa.gov/aqs`, last access 29 August 2022

[4] `https://openaq.org/`, last access 29 august 2022

[5] `https://toar-data.org/`, last access 29 August 2022

[6] This definition is based on verbal communication with Arthur Samuel, but the paper cited contains the most similar written statements.

Ozone concentrations are traditionally modeled by ODE-based (ODE = ordinary differential equation) atmospheric chemistry transport models (CTM, Rao et al., 2011; Schultz et al., 2018; Wagner et al., 2021). Given the multitude of influences on ozone concentrations (Figure 1.1) and the corresponding quantitative uncertainties, these atmospheric chemical models are complex, costly of coarse spatio-temporal resolution, and sometimes biased (Young et al., 2018). Recently, many machine learning methods for ozone research have been developed that complement traditional models and aspire to improve the accuracy and performance of ozone modeling. The machine learning methods are usually based on observations of ozone in combination with geospatial data that accounts for the ozone governing factors mentioned in Section 1.1.1.

The first attempts to forecast ozone with shallow neural networks were published in the 1990s (Yi and Prybutok, 1996; Comrie, 1997). As a further development, Kleinert et al. (2021) and Sayeed et al. (2020) use deep convolutional neural networks to forecast ozone at multiple locations. Their models were trained in parallel on thousands of ozone measurement samples and achieve high accuracy in predicting ozone concentrations of the next days. In the meantime, forecasting ozone or other air pollutants with machine learning is well established (Cabaneros et al., 2019). There are also efforts to improve the performance of CTM by emulating costly chemical reaction schemes with machine learning. For example, Kelp et al. (2020) achieve a performance increase by more than two orders of magnitude compared to the original numerical solver without losing accuracy. Ozone deposition processes in CTM are also improved through machine learning. E.g., Silva et al. (2019) achieve higher accuracy than the numerical dry deposition scheme of the GEOS-Chem model with their deep learning parameterization. Creating ozone maps completely independent of CTM, Ren et al. (2020) and Liu et al. (2020) map ozone concentrations across the US and China, respectively, using tree-based machine learning models. The resulting data products are gridded maps of ozone concentrations for the respective domains. In the field of ozone monitoring, machine learning is often used to calibrate modern low-cost sensors that can complement existing ozone measurements (Schmitz et al., 2021). Lastly, machine learning contributes to scientific insights into ozone-determining factors, such as precursors and meteorology (Balamurugan et al., 2022; Weng et al., 2022).

1.2 Research objectives

To allow for a thorough evaluation of ozone trends and impacts described in Section 1.1.2, it is essential to quantify ozone concentrations with high spatiotemporal coverage, and at sufficiently high spatial resolution. In an ideal setting, ozone concentrations at any point in space and time would be available, to evaluate for example, how much ozone individual humans are exposed to. It is evident that this information is not realistic to obtain, even though a large amount of ozone data is collected in some regions of the world (Figure 1.2). Observations are still lacking in some areas where no or few measurements are made, and measurement stations are heavily clustered (Figure 1.2). Measured ozone time series also have gaps, e.g. due to calibration processes or sensor malfunctions. This work aims to use the benefits of machine learning to produce data products that allow for ozone concentration evaluation where no measurements are available.

The main goal of this work is to map and interpolate ozone data using machine learning methods in conjunction with the high abundance of ozone measurements available.
The machine learning models are expected to learn the ozone patterns from the ozone measurement data. This is a complementary approach to classical numerical modeling, where known ozone processes (Section 1.1.1) are explicitly programmed. The expected advantages over numerical modeling are reduced costs, partly lower biases, and higher flexibility regarding the output resolution. The main data source of this work is the TOAR database (Schultz et al., 2017).

There are three main aspects to the research objective of this work:

1. **Ozone mapping and interpolation.** Mapping and interpolation are techniques that can use existing measurements to predict ozone concentrations where no measurements exist. The two terms are defined in more detail in Section 2.1. We aim to provide high-resolution, high-accuracy, spatio-temporal ozone data products wherever possible. The data products shall cover spatial domains from the regional to the global level, with a temporal resolution ranging from hourly data to multi-year statistics. We will use machine learning on large quantities of ozone measurements, combined with model and geospatial data to generate the data products.

   The corresponding research question is:
   "How can we use machine learning to map and interpolate ozone from existing measurements to data of any required spatiotemporal resolution?"

2. **Trustworthy machine learning.** To produce scientifically sound ozone data products, the machine learning models need to be suitable for ozone-related questions, trustworthy, and properly evaluated. In this work we adapt existing machine learning approaches to ozone research, develop new techniques, and open up black-box characteristics of existing machine learning models.

   The corresponding research question is:
   "What machine learning methods can we develop or adapt to create ozone data products, and how can we make these data products trustworthy?"

3. **Use of spatiotemporal patterns.** As described in Section 1.1, the ozone concentration at a time and location depends on its surroundings and environmental properties. Land use data, digital elevation models, or emission estimates of ozone precursors are commonly available in gridded format. These data exhibit spatiotemporal patterns that are crucial for estimating ozone concentrations. At the same time, ozone measurements are available from irregularly placed measurement stations. It is an objective of this work to use the spatiotemporal information inherent in both geospatial data and ozone measurements within the machine learning models.

   The corresponding research question is:
   "How can we use spatiotemporal patterns represented in geospatial data and ozone measurements effectively within machine learning models?"

We address the aspects 1 - 3 within this work as part of four papers (Chapter 3 and Appendix D). Each paper covers at least two of the three research aspects and examines them from different angles.

1.3 Thesis overview

This work is cumulative and based on four peer-reviewed papers. The further content is structured as follows:

- Chapter 2 “Methods” introduces the most important concepts and technical aspects of this work. It describes the statistical, mathematical, and machine learning methods that are used in this work but are not described in detail in the papers.
- Chapter 3 “Summary of papers” gives an overview of the papers of this work and puts them into the context of the overarching research objectives from Section 1.2.
- Chapter 4 “Synthesis”, links, elaborates and discusses the most important results and conclusions of the different papers. It details the insights relevant to the research objectives from Section 1.2 and therefore highlights the contributions of this work to ongoing research.
- Chapter 5 “Conclusion and new research directions” summarizes findings of this work and points to possible future research directions.
- The Appendices A – D contain lists of figures and tables, abbreviations, the bibliography, and original versions of the papers.

2. Methods

This chapter introduces the methods used in this work. Section 2.1 contains a taxonomy to describe the terms mapping and interpolation. The Sections 2.2 and 2.3 contain technical descriptions of the machine learning and statistical methods. This chapter only describes methods that are not described in detail in the papers to avoid duplicate content.

2.1 Taxonomy

The main goal of this work is mapping and interpolation of tropospheric ozone data. The two terms ozone "mapping" and "interpolation" are related because both methods use existing measurements and auxiliary geospatial data to predict ozone values where no measurements are available (Figure 2.1 (a)). Yet, they do not mean exactly the same. This section introduces the two terms and their differences.

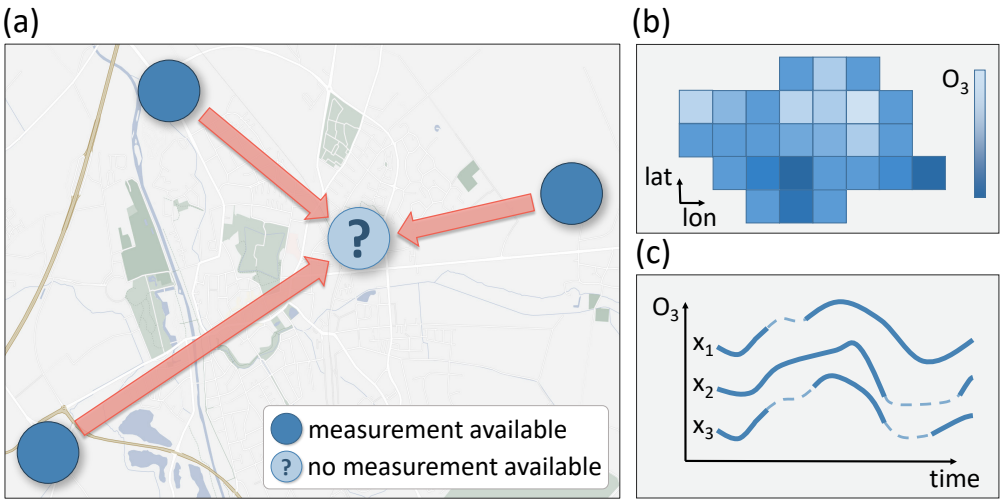


Figure 2.1: This figure illustrates the mapping and interpolation of ozone data in this work. (a) Both methods make use of existing measurements to predict ozone at time steps or locations with no measurements. (b) Mapping creates a spatial gridded field of ozone values. (c) Interpolation imputes the gaps in ozone time series. The underlying map in (a) shows the city of Jülich in Germany (Google Maps, 2022).

2.1.1 What is mapping?

The definition of mapping is "the process of making a map of an area" (*Oxford Learner's Dictionary* 2022), and the term was originally mentioned in the context of cartography. Meanwhile, it is also used when physical quantities are reported over a spatial domain, such as in weather mapping (Kington, 1990; Shermeyer et al., 2020) or satellite imagery (Goetz et al., 2009; Voigt et al., 2016). In the context of this work, mapping is the process of creating a gridded map of ozone values over a spatial domain, as indicated in Figure 2.1 (b). There exist various techniques that are suit-

able for ozone mapping. The most common traditional geostatistical technique is kriging, where ozone values at an unsampled location are estimated by weighted sums of ozone measurements at neighboring measurement stations (Cressie, 1988; Lefohn et al., 1987; Liu and Rossini, 1996). There are more recent efforts to map ozone by post-processing satellite data (Ziemke et al., 2005), or by multi-model fusion (DeLang et al., 2021).

In this work, we perform static regression mapping of global long-term average ozone values with a machine learning model (Betancourt et al., 2022, Section 3.3). For this, the ozone values are predicted for every grid cell of the mapped domain using a machine learning model:

$$\hat{y}_x = M(\vec{f}_x) \tag{2.1}$$

In this equation, x is the grid cell index and M is a machine learning model that takes geospatial features $\vec{f}$ as inputs and outputs ozone statistics $\hat{y}$. M is trained on available measurements and geospatial data. $\hat{y}$ are the predicted ozone statistics the resulting map contains for each grid point. Regression mapping with machine learning was shown to be more accurate than other (geo)-statistical techniques such as kriging in many studies (Nussbaum et al., 2018; Li et al., 2019, e.g.). In contrast to these methods which can only make predictions at locations between available measurements, it also has the advantage to allow mapping in areas with no or very sparse measurements. Regression mapping involves function fitting of the model M, which means that available measurements are not reproduced one-to-one in the final map (Von Storch and Zwiers, 2002). This is a major difference to interpolation which is described in the next section.

2.1.2 What is interpolation?

A mathematical definition of interpolation reads as follows: "The theory of interpolation may, with certain reservations, be said to occupy itself with that kind of information about a function which can be extracted from a table of the function." (Steffensen, 2013). Here, "table" means known discrete values of the function. In other words, interpolation derives a curve that connects known values of a quantity (ozone measurements, e.g.), if no underlying mathematical function is known to obtain these values. It then allows evaluating that quantity where no known values are available. Steffensen (2013) also notes that the process of interpolation is generally problematic if no additional information on the function or underlying process is given because no general assumption can be made about the behavior of the function in between the known values. In environmental applications, however, interpolation is a valid approach as environmental variables are usually smooth and continuous, and their statistical properties are well known. Consequently, the American Meteorological Society (2021) defines interpolation in a less rigid way than Steffensen (2013): "[Interpolation is] the estimation of unknown intermediate values from known discrete values of a dependent variable."

In this work, the interpolated quantity are ozone values (Figure 2.1 (c)). Traditional interpolation methods for ozone include nearest neighbor methods or spline interpolation (Junninen et al., 2004). These methods either make use of auxiliary data such as geospatial features, or of neighboring measurements. In this work, we combine both data sources with machine learning to inter-

polate gaps in ozone measurement time series (Betancourt et al., 2023, Section 3.4). We predict the missing ozone values of a time series at a location x at a time step t with machine learning:

$$\hat{y}_{x,t} = M(\vec{f}_{x,t}, \vec{y}) \tag{2.2}$$

In this equation, M is a machine learning method that is based on geospatial features $\vec{f}$ and available ozone measurements $\vec{y}$ at neighboring locations. It outputs ozone values $\hat{y}$. We use M only to predict ozone values of missing time steps. If a measurement is available, we report it. The differences to regression mapping described in the previous section are (1) we generate no static spatial ozone field, instead, we interpolate gaps in time series (2) we report true ozone values at time steps with available ozone measurements. This is a major difference to a regression map that contains predictions $\hat{y}$ for every grid point, and also the difference between regression and interpolation in a mathematical sense.

2.2 Machine learning

Machine learning models learn rules to make predictions directly from data, as opposed to traditional models where physical rules are programmed explicitly (Samuel, 1959). Training a machine learning model needs a model architecture with adaptive parameters, training data, and an algorithm that can update ("learn") the parameters iteratively. Machine learning has made significant progress in image recognition, speech recognition, gaming, and video prediction over the past decade (Krizhevsky et al., 2012; Mathieu et al., 2015; Amodei et al., 2016; Silver et al., 2016). The success of these works was possible due to the increased availability of computing capabilities, big training datasets, and effective learning algorithms.

One family of machine learning architectures is neural networks, which are loosely inspired by the functioning of the human brain (McCulloch and Pitts, 1943), and can technically approximate any function (Goodfellow et al., 2006). Feedforward neural networks (Figure 2.2) consist of multiple fully connected layers of nodes with adaptive parameters. Information propagates through the network from the input layer to the hidden layers and the output layer, causing different neural activations and solving regression or classification tasks. Recurrent neural networks such as long short-term memory networks (LSTM) are type of neural networks (Hochreiter and Schmidhuber, 1997) especially designed for time series data. They have connections to the nodes representing earlier time steps and are therefore able to catch temporal patterns. LSTM are used for time series-related tasks such as natural language processing and forecasting (Wang and Jiang, 2015; Zhao et al., 2017). In contrast, convolutional neural networks (CNN) are especially effective for image-related tasks (LeCun et al., 1998; Szegedy et al., 2015). They apply adaptive filters ("convolutions"), which are trained to recognize patterns and shapes in images. The transformer is a relatively new neural network architecture that makes less rigid assumptions about the structure of the input data and was proposed by Vaswani et al. (2017). Transformers have been shown to be suitable for tasks like natural language processing where they only rely on so called "attention" mechanisms and refrain from using recurrent elements. They are computationally cheaper, but have an accuracy that is comparable to that of neural networks with recurrent elements (Lin et al., 2021; Wen et al., 2022).

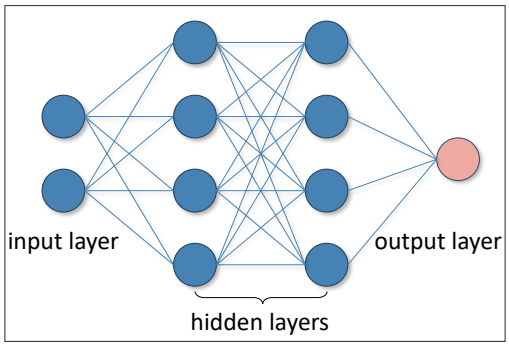


Figure 2.2: A shallow, fully connected neural network with two input nodes, two hidden layers of four nodes each, and one output node. The red node in the output layer represents a prediction.

There are several ways in which machine learning models can learn from data. One is supervised learning (Goodfellow et al., 2006), where the training data is labeled, and the model learns from these labeled data how to predict unlabeled data. There are also unsupervised algorithms, which do not require any labeled data (Duda and Hart, 1973). They can for example be used to group existing unlabeled data points with similar properties into clusters. An intermediate form is semi-supervised learning, where some data are labeled and some are not (Chapelle et al., 2006). Neural networks are usually trained by back propagation (Linnainmaa, 1970; LeCun, 1988). A cost function is defined, that measures the discrepancy between the current output of the neural network and the desired output. Gradients of that cost function are propagated backward through the network and the adaptive parameters are updated iteratively according to these gradients.

As an alternative to neural networks, tree-based machine learning models have been shown to be highly effective on tabular style data (Lundberg et al., 2020; Grinsztajn et al., 2022). Their architecture and training method are further described in Section 2.2.3. There also exist machine learning algorithms that are designed to work on graphs or networks with a more irregular structure than images or time series. Machine learning on graphs is a rapidly growing field because many real-world data have a non-Euclidean structure and can therefore be described using graph theory. For example, graph machine learning can forecast traffic, detect credit card fraud, or make predictions about customer shopping behavior or social networks (Nickel et al., 2016; Hamilton, 2020; Wu et al., 2021). Graph theory can be combined with neural networks (Wu et al., 2021), or other machine learning algorithms (Huang et al., 2020). Basic graph theory relevant to this work is introduced in Section 2.3.2.

2.2.1 Evaluation and generalizability in (semi-) supervised learning

The goal of (semi-) supervised machine learning is to create models that achieve a good accuracy in their given task, and that the accuracy is maintained when the model is applied to unseen data. It is therefore crucial to evaluate models with respect to their generalizability, i.e., that a regression or classification model is tested on samples with known labels that the model has not yet seen during training. The accuracy is measured with an evaluation score, that reflects the residual discrepancy between the known labels and predicted labels. Evaluation scores used in this work are the coefficient of determination R^2, the root mean square error RMSE and the index of agree-

ment d (Table 2.1). These scores are described in detail in the papers (Appendix D). It is custom in (semi-) supervised learning to fit the adaptive parameters only on a subset of the labeled data samples. This part of the data is called the "training set". A second subset is used to tune the hyperparameters of the model. Hyperparameters are parameters that control the training process, e.g. the learning rate. This set is called the "validation set". A third part of the labeled data is set aside for determining the final evaluation score of the model. This part is called the "test set". These three different sets ensure that the model generalizes well to unseen data and that the hyperparameters chosen work for these unseen data as well. Cross-validation can be applied to assess the robustness of the evaluation scores. Cross-validation is a resampling method that iteratively uses parts of the labeled data to test and train the model (Goodfellow et al., 2006).

Table 2.1: Machine learning evaluation scores used in this work.

Name	Abbreviation	Unit	Best score	Worst score
Coefficient of determination	R^2	–	1	$-\infty$
Root mean square error	RMSE	same as model output	0	∞
Index of agreement	d	–	1	0

It is a general requirement that the training, validation, and test sets are independent of each other while having an identical statistical distribution (IID, Goodfellow et al., 2006). Spurious correlations and interdependencies between training and testing data would make test evaluation scores unrealistically high (Meyer et al., 2018; Schultz et al., 2021). Creating an IID data split is trivial if e.g. images should be recognized (Krizhevsky et al., 2012) because it can be assumed that every image is completely independent of the other images. However, in Earth system research, data are often drawn from a continuous domain, on which Earth system variables are usually spatially and temporally correlated and thus dependent. For machine learning on Earth system data, the IID requirement therefore leads to problems, and independence may not even be possible at the same time with an identical distribution. Meyer et al. (2018) noted that there are differences in evaluation scores if models are evaluated on data within correlation distance and if they are not. To ensure independence, the easiest way would be to split the data into distant spatiotemporal regions, which would result in largely independent data. But distant regions often have different characteristics, violating the assumption of an identical statistical distribution. In the field of environmental research, therefore, a compromise has to be found between independence and an identical distribution.

When a machine learning model is used in production, any new input data has to be drawn from the same statistical distribution as the training data (Goodfellow et al., 2006). Production can mean an industrial application, or like in this work, using the model to generate a gridded ozone map. If the input data are drawn from a different statistical distribution as the training data, the model is not suitable for these data, and the model accuracy is not maintained (Meyer and Pebesma, 2021). This problem is also called "covariate shift".

2.2.2 Explainable machine learning

The common machine learning chain is to train a model on a dataset and use that model to produce output results, without the need to know how the model works. The models are called "black box" models because they are typically only used to make predictions and their functioning is not further explained. In scientific machine learning, however, the goal is not only to make accurate predictions but also to gain new scientific insights. Explaining machine learning models is a way of overcoming black box models and obtaining these new insights, combining machine learning with expert knowledge (Roscher et al., 2020; McGovern et al., 2020; Tuia et al., 2021).

Different levels of "explainable machine learning" comprise techniques that aim to achieve transparency, interpretability, and explainability of the models and their predictions. Roscher et al. (2020) elaborate these three terms and point out where the basic machine learning chain can be altered by incorporating domain knowledge and by ensuring that the model is consistent with previous research (Figure 2.3). They define transparency as follows: "[A machine learning] approach is transparent if the processes that extract model parameters from training data and generate labels from testing data can be described and motivated by the approach designer." This includes the model structure, its individual components, and also the learning algorithm. The decision for a transparent model is the beginning of a machine learning process and is shown in Figure 2.3, where an arrow points from the term transparency to the model. Roscher et al. (2020) furthermore define interpretability as follows: "[...] Interpretability pertains to the capability of making sense of an obtained [machine learning] model." This is the case when the way a model works and the basis on which it makes a decision or prediction is understandable to a human. Interpreting a model prediction can mean examining the internal state of the model, or "latent state". Examples of latent state are neural activations in a neural network or decision paths of a tree-based model (Stadtler et al., 2022). In Figure 2.3, arrows point from the term interpretability to both the model and its output results. Roscher et al. (2020) define explainability as follows: "[...] A collection of interpretations can be an explanation only with further contextual information, stemming from domain knowledge and related to the analysis goal." Explaining a model involves the understanding of the model user. The term explainability, depicted in blue color in Figure 2.3, points to the model and output results, and further from these two components to the scientific outcome. Scientific consistency, which according to Roscher et al. (2020) is given when "the result obtained is plausible and consistent with existing scientific principles", can be ensured by comparing the functioning of the model with known scientific principles and previous research. Domain knowledge can feed into this extended machine learning chain from Figure 2.3 on several occasions, such as in the choice of model architecture, in checking whether the model is scientifically consistent, and, along with explanations, in deriving scientific outcome.

Depending on the machine learning application, different explainable machine learning techniques are appropriate. For example, if a machine learning model should recognize images, then saliency maps (Lapuschkin et al., 2019) are valuable. For models that take tabular-style "structured" data as inputs, SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020) are a common ready-to-use-technique. SHAP values provide Shapley additive explanations, i.e, they explain the effect of the input features on the model result. Aggregating SHAP values is a model agnostic

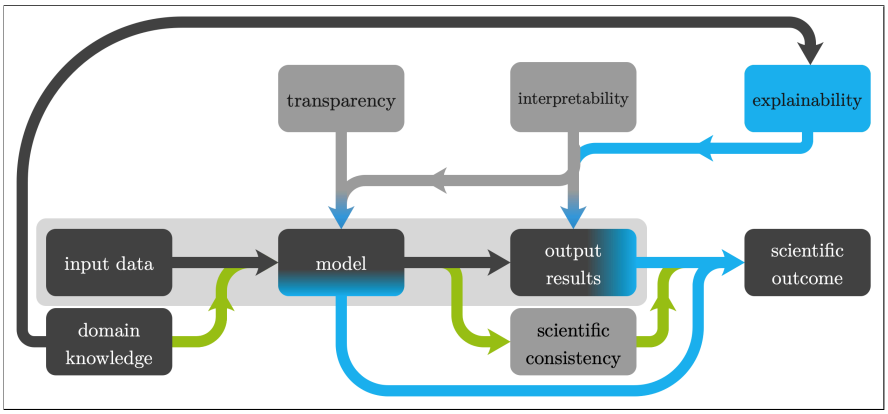


Figure 2.3: The common machine learning chain, extended with explainable machine learning. The light gray box contains the common black box machine learning workflow with a model that is trained on input data and then used to output results. Scientific outcome can be generated by explaining the model (blue arrows) and incorporating domain knowledge (green arrows). For more elaboration, see text. Figure from Roscher et al. (2020).

method to determine global input feature importances. As stated above, the goal of explainable machine learning is to gain new scientific insights and to make the results trustworthy. We define the term "trustworthy" as the model being explainable and interpretable (Sonnewald and Lguensat, 2021). Adadi and Berrada (2018) and Molnar (2020) provide extensive reviews on explainable machine learning.

2.2.3 Tree-based models

Tree-based models are a family of supervised machine learning architectures that consist of decision trees and work on tabular-style data. They take the features of a sample as inputs and can output categorical labels for classification and continuous values for regression. Early decision trees consist of single trees (Breiman et al., 1984, Figure 2.4 left side). Each node of the decision tree represents a logical rule, e.g. if a feature in the input data exceeds a certain threshold. Decision trees split the data at subsequent nodes until they reach leaf nodes. Each leaf represents an answer to the decision problem, e.g. a specific value predicted in regression or a class in classification problems. Figure 2.4, left side shows a decision tree for regression. The dark blue nodes denote an active decision path, while the red nodes denote an active leave, i.e., a prediction.

CART (classification and regression trees) by Breiman et al. (1984) is a frequently used algorithm to train decision trees. This greedy algorithm starts with a single root node and chooses a feature together with a logical rule based on that feature so that the resulting data split minimizes the cost function. The logical rule is usually whether a continuous feature exceeds a threshold, or which class a categorical feature belongs to. There are various suitable cost functions, depending on the problem at hand. For example, a regression tree can rely on the squared error. The feature and logical rule of a node are determined by an exhaustive search of all possible features and rules. After the first node of the decision tree is defined, child nodes are added iteratively to the root node until a stopping criterion is reached. Stopping criteria can be a fixed number of training samples in a leaf or a maximum depth of the tree.

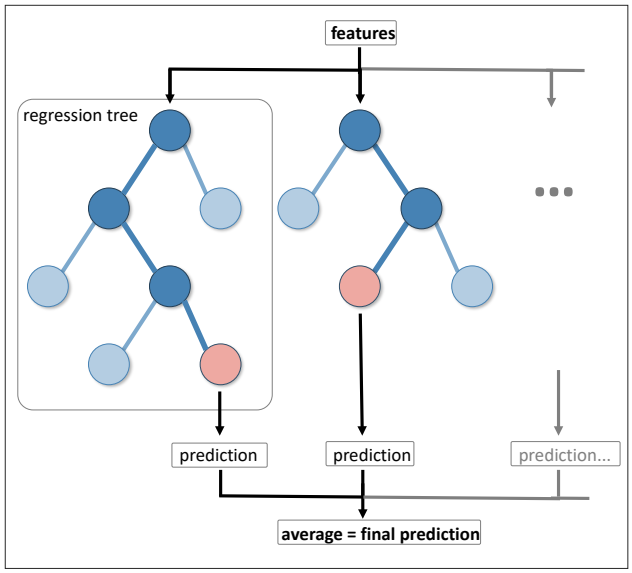


Figure 2.4: Example of a random forest. Active decision paths are indicated by dark blue nodes, and red nodes indicates an active leaf, i.e., a prediction.

Single decision trees have the advantage of being easy for humans to read, but they are sensitive to noise in the data, such as when there is an outlier and the model overfits on that data point (Bishop, 2006). Another problem that is crucial for regression problems, is that a single leaf usually predicts a specific, discrete value which leads to discontinuities in the predicted values. To overcome these problems, different ensemble tree algorithms were developed such as extreme gradient boosting (Friedman, 2001) and random forest (Breiman, 2001). Of these methods, random forest is more robust to noisy training data. The random forest is an ensemble of decision trees. Usually, the number of trees is in the order of several hundred. Figure 2.4 shows a scheme of a random forest. In a random forest for regression, every tree makes a prediction, and the average of all tree predictions is the final prediction. To grow the different trees, the training dataset is bootstrapped several times. Every tree is then trained on one subset of the training data drawn with replacement.

Currently, tree-based ensemble methods are the best choice for structured, tabular-style data (Lundberg et al., 2020; Grinsztajn et al., 2022). They are also insensitive to hyperparameters and have shorter training times compared to neural networks. Therefore, we rely on tree-based models for regression throughout all papers of this work. We use the implementation of Pedregosa et al. (2011). They use the classification and regression trees (CART) algorithm in slightly modified form[1].

2.2.4 Clustering

Clustering is an unsupervised technique to find groups of unlabeled data with similar properties (Duda and Hart, 1973). In this work, clustering is applied to creating independent data splits for the first paper (Betancourt et al., 2021a).

[1]They describe their CART implementation here: `https://scikit-learn.org/stable/modules/tree.html`, last access 19 September 2022

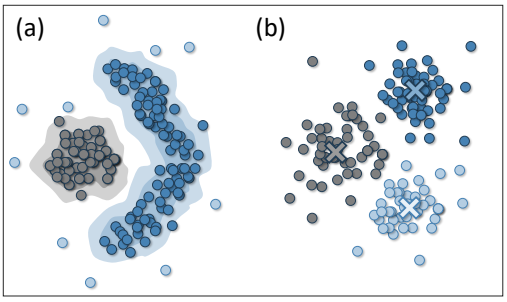


Figure 2.5: Two principles of clustering. (a) DBSCAN, where clusters are solely based on the spatial proximity of the data points to each other (grey and blue areas), and data points with no spatial proximity to any other data points are labeled as noise (light blue data points). (b) K-Means, where the data points are assigned to a fixed number of centroids by spatial proximity to these centroids (shown as grey, blue, and light blue "x").

The most common method for clustering based on spatial proximity is Density-Based Spatial Clustering of Applications with Noise (DBSCAN, Ester et al., 1996). This algorithm has a distance parameter and assigns data points that are closer than that distance parameter into one cluster. The advantages of this algorithm are that the number of clusters does not have to be specified in advance and that no specific shape is assumed for the clusters. The algorithm starts with a random data point and searches for neighboring data points within the distance parameter. If it finds a neighbor, it assigns it to the same cluster. There is an option to control cluster growth according to the number of neighbors of a data point, but it is irrelevant for this work, so we do not describe it here. The algorithm then continues to grow the cluster within the neighborhood of the cluster members. If no additional cluster neighbors are found, the cluster is complete and a new random single point is picked to grow another cluster. This process is continued until all data points are either assigned to clusters, or to noise if they are left single. Figure 2.5 (a) shows data clustered by DBSCAN. The light blue data points are denoted as noise.

Another clustering method is K-Means (Lloyd, 1982). This algorithm divides any data into a pre-set number of clusters K. The K-Means algorithm initializes K random points as first-guess cluster centers. All data samples are assigned to the cluster center with the highest spatial proximity. The cluster centers are then updated to the mean of all samples of this cluster. These two steps are repeated until convergence. The K-Means algorithm, therefore, makes cluster centers move away from each other, and assigns all data points to one of the clusters, regardless of the density of data points and their absolute spatial distance to the cluster center. This behavior is unwanted in many applications and can lead to problems if the clusters do not have a blob shape, or if an unsuitable number of clusters is given (Bishop, 2006). Figure 2.5 (b) shows a dataset grouped into three clusters with the K-Means algorithm.

2.3 Mathematical and statistical methods

2.3.1 Nearest neighbors

Nearest neighbors are a commonly used, nonparametric statistical concept. Based on a dataset and a distance metric, nearest neighbors finds samples in the dataset that are closest to a given

sample (Figure 2.6). One example application is nearest neighbor classification, where for each unlabeled sample the nearest sample with a label is searched and the label of this sample is assigned to the unlabeled sample (Duda and Hart, 1973; Bishop, 2006). The search space can be any domain, e.g. the geographical space and geographical distances. Another example of a search space is the feature space of a machine learning model, which is the multi-dimensional space spanned by its input features.

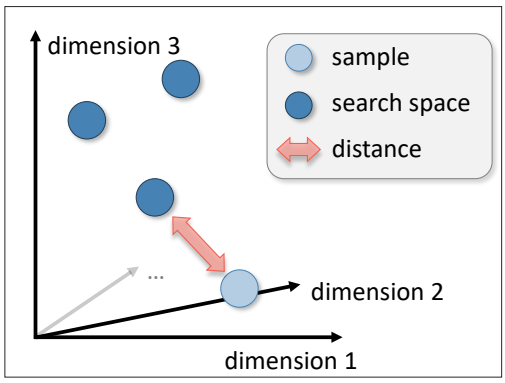


Figure 2.6: Nearest neighbor principle. For a given sample and dataset, the nearest neighbor is searched. The search space (denoted with dimension 1 - 3) can be any domain.

In this work, the location and distance of nearest neighbors are used on occasions with different search spaces. The second paper (Stadtler et al., 2022, Section 3.2) bases explanations on nearest neighbors in the latent space of the model, which is spanned by neural activations in a neural network (Figure 2.2) or a random forest (Figure 2.4) when the models make a prediction. We use nearest neighbors in the third paper (Betancourt et al., 2022, Section 3.3), to determine how safe it is to make an ozone prediction at a given location. The fourth paper (Betancourt et al., 2023, Section 3.4) uses the method of nearest neighbor regression as a baseline method. All examples and search spaces are described in detail in the papers.

We use the scikit-learn software package (Pedregosa et al., 2011) for nearest neighbor search. The documentation of this software notes that there exist different algorithms for nearest neighbor search. The brute force algorithm calculates the pairwise distance of all samples and reports the pair with the smallest distance. K-d tree and ball tree algorithms make use of the internal data structure, and can therefore be cheaper (Bentley, 1975; Omohundro, 1989). The authors of scikit-learn recommend, however, using the brute force method for high dimensional datasets because the intrinsic dimensionality of the data is generally too high for tree-based nearest neighbor algorithms. Therefore, we use the brute force method for our data.

2.3.2 Graph theory

Graphs are a “general language for describing and analyzing entities with relations or interactions” (Leskovec, 2021). Graph structures are non-Euclidean and appear in many real-world settings, such as telecommunication networks, traffic systems, social networks, or in chemical structures (Hamilton, 2020).

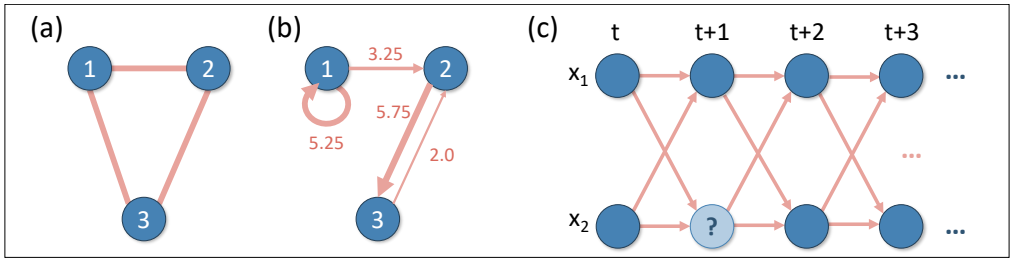


Figure 2.7: Examples of graphs. (a) A simple graph, consisting of three nodes and three edges. (b) A directed graph with a self-loop, double edges, and edge weights. (c) The graph of time series at two locations $x_{1,2}$ at time steps $t+0, 1, 2, 3$, where the question mark denotes an unlabeled node.

The entities of a graph ($\mathcal{G}$) are called nodes ($\mathcal{V}$). If a so-called edge ($\mathcal{E}$) connects them, they can interact or "pass messages" with each other. The graph is therefore defined as a set of nodes and edges: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Figure 2.7 (a) shows a simple graph with three nodes and three edges. Figure 2.7 (b) shows a directed graph with a self-loop, double edges, and edge weights. Edge weights are a measure for the intensity of interaction between two nodes. Apart from the graphical representation from Figure 2.7, an adjacency matrix ($\mathbf{A}$) is a convenient way to represent a graph. The adjacency matrix contains the weight w of the edge between two nodes u and v if they are connected and zero otherwise: $\mathbf{A}_{u,v} = w_{u \to v}$. The adjacency matrix of the graph in Figure 2.7 b) therefore reads:

$$\mathbf{A} = \begin{pmatrix} 5.25 & 3.25 & 0 \\ 0 & 0 & 5.75 \\ 0 & 2.0 & 0 \end{pmatrix} \tag{2.3}$$

Real world graphs often have thousands or millions of nodes, and sparse adjacency matrices (Hamilton, 2020; Leskovec, 2021). This makes sparse matrix operations most feasible for many graph-based applications.

The degree d of a node in a simple graph is the number of its edges. In directed, weighted graphs, it is the sum of the weight of the incoming edges.

$$d_v = \sum_{u \in \mathcal{V}} \mathbf{A}_{u,v} \tag{2.4}$$

The degree matrix $\mathbf{D}$ contains all node degrees of the graph on the diagonal.

The motivation to use graph theory in this work stems from the fact that real-world ozone measurements are of irregular placement (Figure 1.2). The spatial structure of the measurements can therefore be described as a graph. As a complementary motivation, Figure 2.7 (c) shows how a graph can be defined on two ozone time series of measurements at different locations. Here, every node is labeled with a measurement at a specific place x and time t, or unlabeled if no measurement is available. Edges connect the two time series with a time shift of one time step. In this work we use graph machine learning to interpolate gaps in ozone measurement time series (Betancourt et al., 2023, Section 3.4), combining the context of a spatial graph (Figure 1.2) and a time series graph (Figure 2.7 (c)). The graph machine learning method used in this paper uses the graph

theory described in this Section and is described in detail in the paper itself. The books by Hamilton (2020) and Barabási (2016) contain a more thorough general introduction to graphs and graph machine learning.

3. Summary of papers

The papers of this work comprise four published journal articles. Figure 3.1 illustrates how the papers build on each other. The first paper introduces AQ-Bench, a benchmark dataset for machine learning on global ozone metrics and geospatial data. The second paper devises an explainable machine learning method that allows detailed insight into how models trained on AQ-Bench represent the training data and derive their predictions. These first two papers substantiate the use of machine learning methods to predict ozone with geospatial data as input. Their findings are therefore the basis for the mapping and interpolation we conduct in the third and fourth papers. The third paper uses a model trained on AQ-Bench for global mapping of long-term ozone average values, including explanations and uncertainty assessments of the predictions. The fourth paper complements the geospatial data of AQ-Bench with time-resolved meteorological data to interpolate hourly missing ozone measurements with a graph machine learning method.

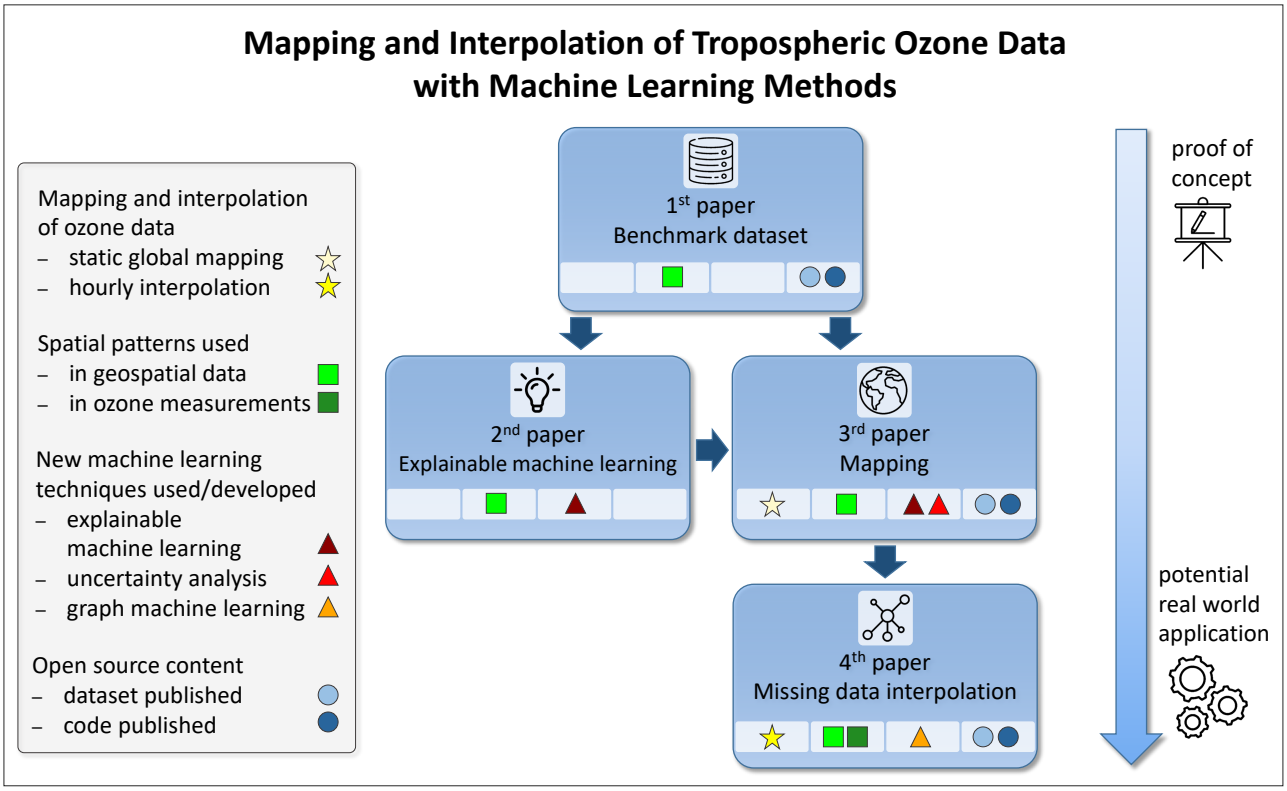


Figure 3.1: Graphical summary of the four papers of this work. The colored shapes mark contributions to the research objectives from Section 1.2, and open source content. The arrow on the right indicates that the machine learning applications range from a proof of concept to a potential real-world application. Icons by `flaticon.com`.

Based on the research objectives from Section 1.2, Figure 3.1 indicates whether papers include mapping or interpolation of ozone data, what new machine learning methods they develop, and how they use spatial patterns. We provide open data and code with the papers to ensure repro-

ducibility, as additionally indicated[1]. Fig. 3.1 also shows the evolution of the papers, starting with a proof-of-concept benchmark dataset in the first paper and ending with an interpolation method for missing ozone measurements that is suitable for operational use in the fourth paper. The following sections give an overview of the content of the papers and briefly present the most important results. We also note the distribution of tasks among the co-authors. A detailed synthesis of the scientific findings follows in chapter 4. The published versions of the papers can be found in Appendix D.

3.1 Benchmark dataset

The first paper of this work introduces AQ-Bench, short for "Air Quality Benchmark". It is a benchmark dataset for machine learning on global air quality metrics. Machine learning benchmark datasets are published so they can be reused and accelerate machine learning progress in a specific domain. They make machine learning approaches comparable to each other, and ease access to machine learning data without the need for time-consuming data preparation.

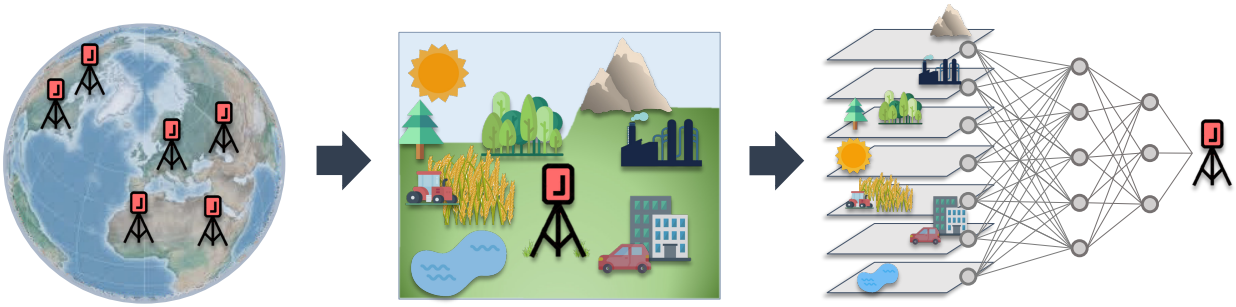
Bibliography entry:
Betancourt, C., Stomberg, T. T., Roscher, R., Schultz, M. G., and Stadtler, S. (2021a). "AQ-Bench: a benchmark dataset for machine learning on global air quality metrics". In: *Earth System Science Data* 13.6, pp. 3013–3033. DOI: 10.5194/essd-13-3013-2021.

Citation:
The paper is cited as "Betancourt et al., 2021a" throughout this work.

Published version:
The published version can be found in Appendix D.1.

Paper content. Figure 3.2 shows the graphical abstract of AQ-Bench. The focus of the dataset is on basic statistics and long-term metrics of ozone. It consists of temporally aggregated ozone metrics and static metadata at 5577 stations all over the globe. The aggregation period is from 2010 to 2014. The metrics summarize hourly ozone measurements to assess the long-term ozone burden and impacts at the measurement stations. Basic statistics include, for example, average values and percentiles, while other metrics are suitable to assess ozone impacts on health and vegetation. The metadata in the AQ-Bench dataset are easy-access geospatial data at the measurement stations. They are continuous and categorical features such as the *altitude*, the *station type*, and the *land cover* of the area around the stations. These metadata are part of the dataset because they are proxies for the determining factors of ozone given in Section 1.1.1. The data source of both ozone and geospatial data in AQ-Bench is the database of the Tropospheric Ozone Assessment Report (TOAR, Schultz et al., 2017).

[1]The 2nd paper (Stadtler et al., 2022) does not provide open data as this paper only uses the AQ-Bench dataset already published with the 1st paper (Betancourt et al., 2021a). It does not provide open code either, as we originally planned to release the machine learning method described in the paper as a separate software package.



The AQ-Bench dataset contains long-term air quality metrics and metadata at sites around the globe.

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

The proposed machine learning task is to train a machine learning algorithm which predicts long-term air quality metrics at measurement sites using the metadata as input.

Figure 3.2: Graphical abstract of the AQ-Bench dataset. Figure adapted and modified from Betancourt et al. (2021).

As advised for machine learning benchmark datasets (Ebert-Uphoff et al., 2017), we predefine a machine learning task for AQ-Bench. Users are encouraged to use machine learning for predicting the ozone metrics using the metadata as input features (Figure 3.2, right). We also propose the coefficient of determination R^2 as an evaluation score for this task. R^2 is especially suitable for this multi-target task, because it is comparable between all metrics. Baseline experiments with linear regression, a shallow neural network, and a random forest architecture serve as user references and validate the dataset. One challenge for this dataset is to find a suitable partitioning of the data into a training, validation and test set for independent evaluation of the machine learning models (Section 2.2.1), even though the ozone measurement stations are irregularly located (Figure 1.2). We obtain this data split by a clustering (Section 2.2.4) approach.

The machine learning task we pose with the AQ-Bench dataset is a real-world task that is relevant to humans and the environment. Yet, it is suitable for beginners, since the dataset is comparably small and contains tabular-style data. It is therefore aimed at atmospheric scientists who pursue machine learning and at machine learners who engage in the environmental sciences. Since we completed the time-consuming process of data acquisition and preparation, users can directly address topics such as feature engineering (Duboue, 2020), the development of new machine learning methods, or explainable machine learning. Through its extensive documentation and open source availability, AQ-Bench contributes to FAIR data (Wilkinson et al., 2016) in the environmental sciences.

Main results. The main deliverables of this paper are the AQ-Bench dataset itself, and the associated machine learning code. The AQ-Bench dataset is licensed under Creative Commons Attribution (CC-BY) and can be downloaded from
`https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f`.
The code is licensed under MIT License and hosted under
`https://gitlab.version.fz-juelich.de/esde/machine-learning/aq-bench`.
The data are FAIR (Wilkinson et al., 2016) and the code is open source. They are documented, easily accessible, and reusable.

The key scientific finding of this paper is, that static geospatial data as inputs for machine learning models have predictive power for long-term ozone metrics. This finding is a prerequisite for the following papers of this work that map and interpolate ozone using these data. It is nontrivial because the geospatial data have no direct known connection to ozone, but are merely proxies. We chose these geospatial features because they are related to drivers or factors of the chemical and physical processes leading to ozone formation and destruction (Section 1.1.1). One of the benefits of machine learning is that if a connection between the inputs and the outputs exists, it can be learned, even if it is not directly known (Section 2.2). The baseline models show good evaluation scores for most of the target metrics with R^2 scores of 0.5 or more. Only two metrics that count the days in a year with ozone concentration threshold exceedances (*nvgt* metrics) have lower evaluation scores. We hypothesize that this is because of an imbalanced data problem. There exist many stations which rarely exceed given ozone thresholds, so their *nvgt* values are zero. It is difficult to learn from these imbalanced data. The machine learning methods random forest and neural network outperform the linear regression, which shows that the nonlinearity of machine learning is beneficial when predicting ozone. The random forest has the best evaluation scores for all target metrics except the *nvgt* metrics, surpassing the neural network and the linear regression.

The third research objective of this work is to perform machine learning on spatial patterns. Regarding this research objective, we use precomputed spatial patterns as input features. They are feature engineered from available geospatial gridded data. For example, the *relative altitude* is the difference between the station altitude and the lowest altitude found in a radius of 5 km around the station and was derived from a digital elevation model and the reported station altitude. This spatial pattern is important because flow patterns are different on a high plateau and on a mountaintop, for example. Another pre-computed spatial pattern is *population density*. It is reported at the station, together with maxima in radii of 5 km and 25 km around the station. The different radii allow distinguishing between remote areas and sparsely populated areas in the vicinity of a city, which is crucial when characterizing a location with respect to ozone patterns. The *nightlight* and land cover features also entail pre-computed spatial patterns.

Own contribution. This paper is the result of a collaboration with scientists from the University of Bonn and the Forschungszentrum Jülich. Based on Schultz et al. (2017), who compile ozone data and geospatial data together in a database, I conceived the idea of linking these data with a machine learning task and baseline experiments to create a machine learning benchmark dataset. I played the key role in selecting appropriate methods, designing experiments, preparing and visualising data, writing software and drafting the first manuscript. Additionally I coordinated communication with the University of Bonn.

3.2 Explainable machine learning

The second paper of this work develops explainable machine learning approaches for two machine learning models trained on the AQ-Bench dataset. Explainable machine learning allows making scientific discoveries by overcoming the black box behavior of machine learning models. This paper is also an example of how explainable machine learning can support decision making.

Bibliography entry:
Stadtler, S., Betancourt, C., and Roscher, R. (2022). "Explainable machine learning reveals capabilities, redundancy, and limitations of a geospatial air quality benchmark dataset". In: *Machine Learning and Knowledge Extraction* 4.1, pp. 150–171. DOI: 10.3390/make4010008.

Citation:
The paper is cited as "Stadtler et al., 2022" throughout this work.

Published version:
The published version can be found in Appendix D.2.

Paper content. This paper describes a new method to obtain and use post-hoc explanations of machine learning models and applies it to two models trained on the AQ-Bench dataset. The models are a random forest and a shallow neural network, as Betancourt et al. (2021) proved that they are suitable for AQ-Bench (Betancourt et al., 2021a).

We first compare the accuracy of the models and calculate the global SHAP importances for them. Our own explainable machine learning method is based on the latent space, and uses model activations (Section 2.2), and nearest neighbors (Section 2.3.1). For every prediction of a test sample, the method searches training samples with a similar latent model activation, i.e. nearest neighbors in the latent space. For the random forest, those are training samples with the same decision path. For the neural network, the training samples have similar neural activations. We evaluate the nearest neighbors for their suitability to make the predictions, assuming that the nearest neighbors are the influential training samples for the predictions. In other words, the explanations analyze model predictions by relating them to the underlying training samples.

We use the explanations to point out the limitations of AQ-Bench and to suggest improvements. A special focus is on explaining why inaccurate predictions occur. The method can flag i) failed predictions due to inputs that are not represented in the training data, ii) training data which are not helpful to make predictions and therefore not contribute to the model accuracy, and iii) unexpected inaccurate predictions, which fail due to unknown reasons and are therefore untrustworthy. The method can also point out "Clever Hans" predictions, which are correct for the wrong reasons (Lapuschkin et al., 2019). Based on the underrepresented samples, we propose new locations for air quality monitoring stations. Likewise, we train alternative models where unnecessary training data are left out.

The black box behavior of machine learning models is unwanted in environmental sciences, and this paper overcomes it. By taking a closer look at how models are trained on the AQ-Bench dataset, and how they derive their predictions, we increase the trust in the data and models. Going beyond post-hoc analyses, this paper also proposes examples of how to bridge decision making and explainable machine learning.

Main results. Going beyond standard data analysis, the paper provides a look at the AQ-Bench dataset from a machine learning perspective. The main results are the preliminary analysis of the models, and, most importantly, the model explanations provided by our own method.

The accuracy of the random forest and the neural network are relatively similar with test R^2 values of 0.53 and 0.49, respectively. Their residuals are highly correlated, so when the random forest fails to make an accurate prediction for a specific test sample, the neural network tends to fail as well. Yet, the neural network has more "clever Hans" predictions than the random forest. This points to the fact that some samples in the dataset are generally harder to predict correctly than others. The feature *absolute latitude* has the highest global SHAP importance in both models, explaining more than 20 % of the predicted variance. Features related to the *altitude*, *nightlight*, *population density*, or land cover with water and forests are of medium importance to both models. There are some features with low or zero importance, for example, the feature *permanent wetlands in 25 km area*. We hypothesize that these features are of low importance because they have low variance in the AQ-bench dataset. For example, there are almost no stations in permanent wetlands in AQ-Bench, so a machine learning model has little to no chance of learning from this feature.

Figure 3.3 shows the main results of our explainable machine learning method on a map projection. The figure shows non-influential training stations in white. They occur in remote areas with few air quality monitoring stations, presumably because these training stations have very different features from any test station. Non-influential training stations also occur in areas with a high density of air quality monitoring stations, which points to redundant information, i.e., their features are already well represented by other stations. Stations with untrustworthy predictions are light blue in Figure 3.3. They occur in all areas, and almost all untrustworthy predictions are untrustworthy for the two models. One possible explanation for that is noise in the AQ-Bench dataset, which neither model could learn from the training data. Underrepresented test stations are shown in plum in Figure 3.3. They occur mainly in remote areas with different characteristics from the training stations. To improve these underrepresented feature combinations, we propose areas for building new air quality monitoring stations in blue, red, and purple in Figure 3.3. The areas proposed for the random forest and the neural network do not overlap entirely, because the models rely on different features, and, therefore, choose different feature combinations to improve their predictions. Notably, the new proposed locations are shaped in bands along a certain latitude because this is the most important feature in both models. Training the model on a dataset without the irrelevant samples reduces the coefficient of determination by only 1 % for the random forest and 2 % for the neural network.

In view of the goals of this work stated in section 1.2, this paper develops a new technique of explainable machine learning. It increases the trust in the machine learning models and training data. This method is also reusable for models trained on other datasets.

Own contribution. This work is a collaboration between scientists from the University of Bonn and the Forschungszentrum Jülich. The first author of this paper is Scarlet Stadtler. She and I jointly developed the explainable machine learning method this paper is based on and carried out the analyses. I assumed the primary responsibility for software development and visualization of results. Scarlet Stadtler created the first paper draft and all authors improved upon that draft.

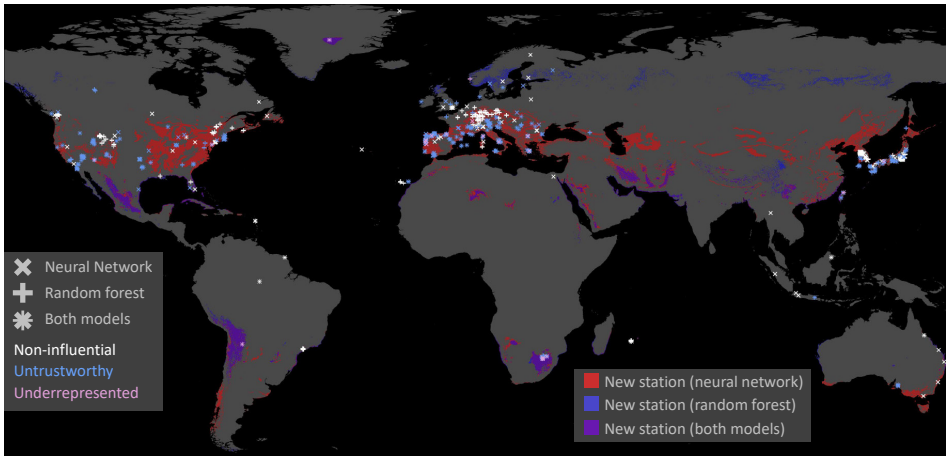


Figure 3.3: Flagged stations of the second paper on a map projection, showing stations that are non-influential for the model accuracy, untrustworthy, and underrepresented in the training data. This figure also marks the proposed regions for new building locations of air quality monitoring stations. Figure from Stadtler et al., 2022.

3.3 Mapping

The third paper of this work performs high-resolution mapping of ozone data across the global domain with machine learning. To our knowledge, this is the first completely data-driven global mapping approach for ozone. Besides proving a point that it is possible to map ozone with this method, this paper also entails explainable machine learning and uncertainty estimates, which increases the trust in the produced machine learning-based data products.

Bibliography entry:
Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., and Stadtler, S. (2022). "Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties". In: *Geoscientific Model Development* 15.11, pp. 4331–4354. DOI: 10.5194/gmd-15-4331-2022.

Citation:
The paper is cited as "Betancourt et al., 2022" throughout this work.

Published version:
The published version can be found in Appendix D.3.

Paper content. This paper uses a random forest trained on the AQ-Bench dataset to create a static, global, fine-resolution map of ozone. We perform mapping, as described in Section 2.1.1, to create a map of the average ozone concentration of the years 2010 - 2014 with a resolution of 0.1°×0.1°. We apply the random forest pixel-wise to gridded fields of geospatial data from various sources, such as satellite products. This paper also entails various techniques of explainable machine learning and uncertainty assessment as described below.

Explainable machine learning in this paper increases trust in the produced maps and checks if the used machine learning model is consistent with commonly accepted knowledge about ozone.

Before training, we apply feature engineering, merging features in AQ-Bench with similar properties together. For example, we merge all land cover features associated with forests. Feature engineering does not increase the model accuracy, yet fewer features make the model easier to interpret. We also apply a forward feature selection method proposed by Meyer et al. (2018). This method removes misleading features that favor overfitting. To ensure that the model is spatially robust and applicable in regions with sparse or no training data, we develop a two-stage spatial cross-validation method. First, we apply cross-validation on a small scale. Training and testing data are approximately 50 km apart here and therefore considered independent of each other (European Union, 2008). In the second step, we apply cross-validation on the global scale, distributing the training and testing data by their world region. For example, we train a model on data from Europe and East Asia and test this model on data from North America. The global cross-validation evaluates the global generalizability of the model. We also calculate the SHAP values of the model. We use SHAP as a tool to explain the model predictions and to check their consistency with previous ozone research.

To assess uncertainties and to ensure trustworthy predictions of our model, we define its area of applicability (Meyer and Pebesma, 2021), which flags feature combinations underrepresented in the training data as non-predictable. The decision, which input feature combination is predictable and which is not, is based on the Euclidean distance in the normalized multi-dimensional feature space. The method checks the distance of the input features in the mapping domain to the AQ-Bench training features. If the distance is greater than a threshold value, the model is not applicable, and we flag the model output as not trustable. We also assess the robustness of the model with respect to common ozone fluctuations by retraining the model several times and monitoring the variance of the resulting map. Then we perturb the inputs with a variance that could realistically occur and propagate the perturbed inputs through the model. If the resulting map has deviations from the standard produced map that are greater than the 5 ppb (Schultz et al., 2017), it means that the model is not robust.

Main results. The main deliverable of this paper is the produced ozone map and associated uncertainties (Figure 3.4). With this map, this paper adds to the first research objective of this work (Section 1.2). Both the map and pixel-wise uncertainty estimates are licensed under CC-BY and available under
`https://doi.org/10.23728/b2share.a05f33b5527f408a99faeaeea033fcdc`.
The map has ozone values between 0.4 and 56.5 ppb, and shows commonly known global ozone patterns such as higher values in mountain ranges, a north-south gradient in Europe, and low values in urban areas. Meyer and Pebesma (2022) warns against using untrustworthy global mapping models and therefore suggests giving uncertainty estimates for each pixel in a mapping domain, which we do (Figure 3.4). The associated uncertainty estimates are an RMSE of 4 ppb in regions with good spatial coverage of training data and suitable feature combinations for the model. We expect a higher RMSE of approximately 5 ppb in regions with suitable feature combinations but no spatial proximity to the training data. The associated machine learning code with MIT license is available under

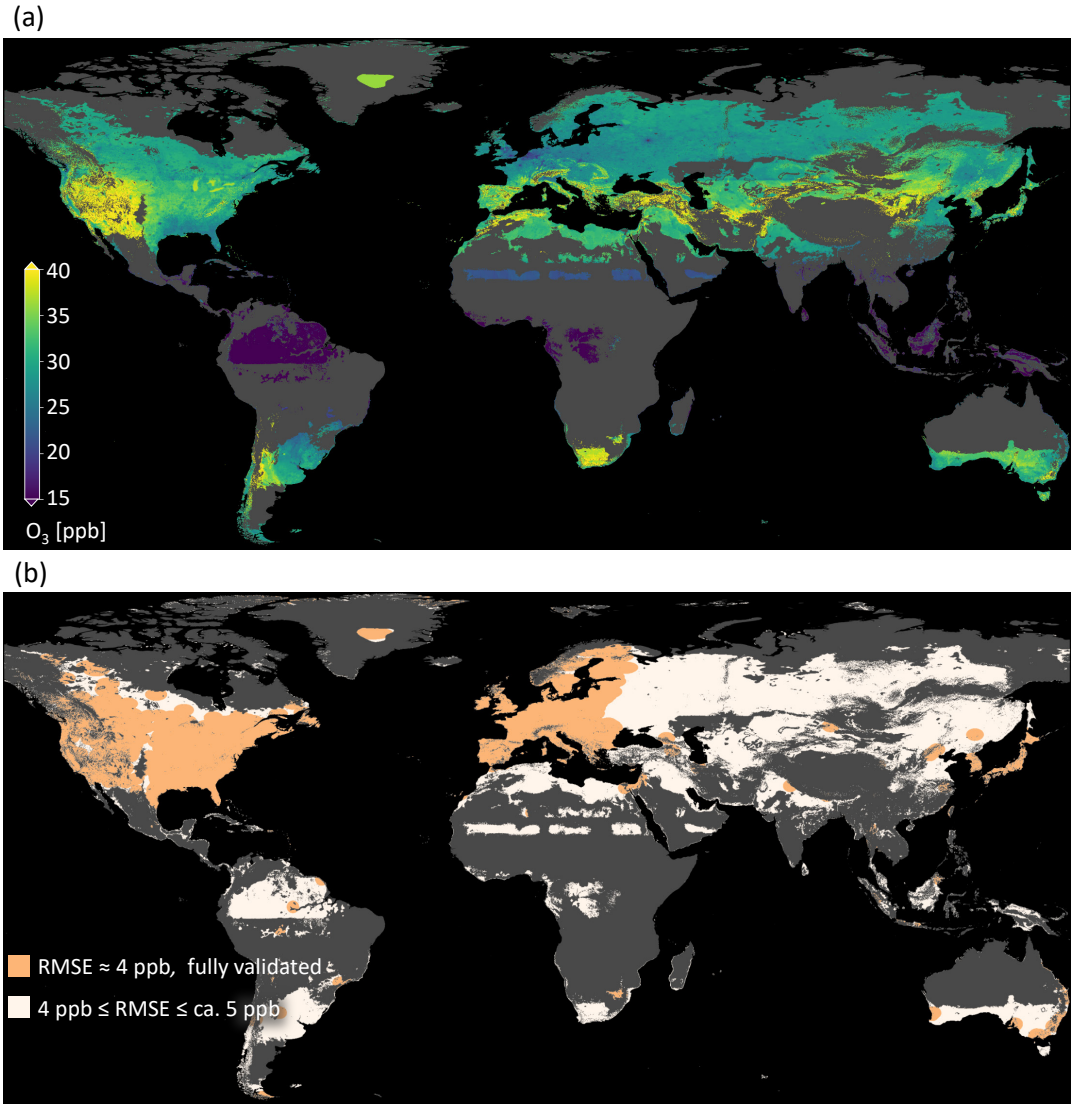


Figure 3.4: Results produced within the scope of the third paper of this work. (a) Map of average ozone values from 2010 to 2014. (b) Uncertainty estimates for every pixel. Figure adapted and modified from Betancourt et al., 2022

https://doi.org/10.34730/af084443e1c444feb12d83a93a65fa33,
to ensure reproducibility of the results.

The second research objective of this work is to generate trustworthy machine learning data products. We obtain trustworthiness through the different explainable machine learning and uncertainty experiments of this paper, and summarize their results in the following. The feature selection method discarded several features which are either not well represented in the training data, or have an unclear or contradicting effect on ozone values. For example, *snow and ice in 25 km area* is not well represented in the training data. *Cropland/natural vegetation mosaic in 25 km are* is difficult for the model to use because croplands and natural vegetation have counteracting effects on the ozone burden (Section 1.1.1), and this input feature is a mixture of the two land-cover types. We also performed a two-stage spatial cross-validation on the local and global scale. The local scale cross-validation showed RMSEs of approximately 4 ppb, while a model trained and tested on data from different world regions showed test errors of approximately 5 ppb. This shows that the model generalizes well across different world regions and is therefore globally applicable. Yet, in regions with no or sparse training data, errors are expected to be larger than in regions well covered with measurement stations. The SHAP values show a general plausibility of the model, for example, a high *altitude* has a positive effect on the predicted ozone values, which is consistent with common ozone knowledge.

The final map in Figure 3.4 only shows ozone values for regions where the model is applicable, which means that large parts of South America and Africa are not part of the map. The model is robust against fluctuations in both ozone and input features. Yet, the regions well covered with training stations are more robust against those fluctuations than regions with sparse training data. In summary, the paper entails diverse experiments related to explainable machine learning and uncertainty assessment, which draw a consistent picture of a robust and trustworthy machine learning data product.

Own contribution. This paper is a joint effort with scientists from the University of Bonn and The RWTH Aachen University. The Jülich scientists contributed their expertise on ozone, the Bonn scientists took the lead in designing explainable machine learning experiments, and the Aachen scientists took the lead in estimating uncertainty impacts. I lead the conceptualization, distributed the associated tasks, and coordinated the exchange of research results. I also took primary responsibility for curating the code and writing the first draft of the manuscript, and coordinated the feedback iterations to improve upon that first draft.

3.4 Missing data interpolation

The fourth paper of this work applies the new graph machine learning algorithm “correct and smooth” to interpolate missing hourly ozone measurements in Germany. The use of graph machine learning on an air quality monitoring network is beneficial because it can use the temporally and spatially irregular available measurements to improve the interpolation up to high precision.

Bibliography entry:
Betancourt, C., Li, C. W. Y., Kleinert, F., and Schultz, M. G. (2023). "Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data". In: *Environmental Science & Technology* 57.46, pp. 18246–18258. DOI: 10.1021/acs.est.3c05104.

Citation:
The paper is cited as "Betancourt et al., 2023" throughout this work.

Published version:
The published version can be found in Appendix D.4.

Paper content. The main idea of this paper is to interpolate missing ozone measurements with the new graph machine learning method correct and smooth (Huang et al., 2020). The dataset we interpolate comprises hourly measurements of the year 2011 at 278 stations of the German Environment Agency (Umweltbundesamt – UBA), in which 15 % of the measurements are missing. It is a preliminary dataset with frequent and larger gaps than the final validated dataset the UBA provides. We chose to develop our method on the preliminary dataset to better demonstrate its potential. The missing data show three distinct patterns: short gaps at single stations of up to 5 h length, longer gaps at single stations of up to several months length, and gaps that occurred at all stations of the UBA network simultaneously. As auxiliary data for the interpolation, we complement the geospatial data from AQ-Bench with meteorological data and model data. In more detail, we use hourly COSMO reanalysis (Bollmeyer et al., 2015), which we extract together with the ozone measurements from the TOAR database (Schultz et al., 2017). Additionally, we extract CTM data from the ECMWF Atmospheric Composition Reanalysis (EAC4, Inness et al., 2019) and emission data with monthly resolution from the CAMS Global anthropogenic emissions (version 5.3, Granier et al., 2019).

We fuse the auxiliary data and neighboring available ozone measurements in the vicinity of a missing value to interpolate the gaps. For that, we use graph machine learning. We define the graph structure so that each hourly data point at a station is a node, so there are about 2.4 million nodes. An edge exists between two nodes if their stations are 50 km in spatial distance or closer, and if their measurement times are 6 h or less apart. This results in about 240 million edges. We provide a static version of this graph in Figure 3.5, where we show one node for each station and omit the temporal component for clarity. All nodes have the meteorological and geospatial data as features, and if they have a measurement available, they are labeled with that measurement. The graph machine learning method is correct and smooth (Huang et al., 2020) which can fuse the available measurements and auxiliary data. It works in three steps. First, a simple model predicts first-guess ozone values, using the features of the node, but without using neighboring data. We tested different simple models and decided for a random forest as it performs best in our case. Second, the correct step improves the prediction of the simple model depending on the bias of the simple model at neighboring nodes with known labels. In the third step, the model smoothes over all neighboring predictions.

Missing data in ozone time series can decrease the robustness and reliability of analyses for impact assessments. For example, long-term metrics can be corrupted or have to be considered

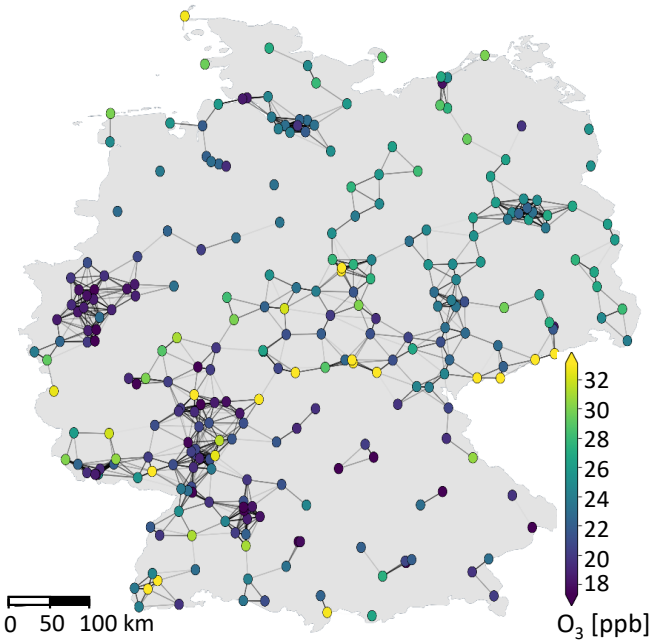


Figure 3.5: Illustration of the graph defined on the station network of the UBA. The circles shown are the stations, and their color corresponds to the mean ozone value measured there. Stations closer than 50 km are connected by an edge. The graph in this paper is dynamic, but this figure omits the temporal component for better visualization. Figure adapted and modified from Betancourt et al., 2023

missing if there are too many missing data. Moreover, machine learning applications for forecasting ozone need gap-free data for training and predictions (Kleinert et al., 2021; Sayeed et al., 2020). Therefore, the interpolation of missing data is vital for many applications. Graph machine learning offers an opportunity to solve this task by making ideal use of the available measurements and auxiliary data.

Main results. The main result of this paper is the ozone interpolation method which is part of the first research objective of this work. We successfully adapt a new graph machine learning method for the interpolation. We obtain better results with the graph-based method than with point-based machine learning with a random forest alone. The method also surpasses statistical baseline models which are frequently used for missing data interpolation (Junninen et al., 2004). The predictions with that model are so fast that the interpolation could even be done in (near-) real time. Depending on the gap type, the RMSEs of the interpolation are 2.3 to 6.18 ppb, and the index of agreement d ranges between 0.96 and 0.99. Short gaps of up to 5 h length are interpolated best with linear interpolation, and longer gaps are most accurately interpolated with correct and smooth. Also, many neighboring available measurements result in more accurate interpolations, and longer gaps are generally interpolated with less accuracy than shorther gaps. To our knowledge, this is the first application of correct and smooth on continuous data, and in the field of air quality research.

We published the code under
`https://gitlab.jsc.fz-juelich.de/esde/machine-learning/ozone-imputation`,

the training data under
http://doi.org/10.23728/b2share.59281340dd37485eb2c6a08de3587c13,
and the interpolated dataset under
http://doi.org/10.23728/b2share.04821864a81f40af89c7633889f147cb.

Another aspect of this paper is the use of spatial patterns, which is the third research objective of this work. In the previous papers, we used point data, and spatial patterns only were used by precomputed feature-engineered inputs. This approach is not the most optimal, as it leaves any neighboring available measurement data unused. In contrast, this paper makes ideal use of neighboring measurements in space and time. It also uses a larger variety of input features, namely meteorological and CTM data, which make it possible to perform the hourly interpolation. The interpolation is limited to UBA station locations in Germany and the year 2011, but we expect the method to work well in any area with good spatial coverage of air quality monitoring stations.

Own contribution. The work is a joint effort with colleagues from the Jülich Supercomputing Centre and a visiting scientist from the Max Planck Institute for Meteorology in Hamburg, Germany. I took the lead in conceptualizing the paper together with the co-authors and Franca Hoffmann from the University of Bonn, who is mentioned in the acknowledgments of this paper. I assumed the primary responsibility in programming, data analysis, visualization and writing, and was supported by the co-authors.

4. Synthesis

This chapter contains a synthesis of the scientific findings of this work. Sections 4.1 to 4.3 consolidate the results of the papers of this work into answers (**A**) to the initial research questions posed in Section 1.2. Then, Section 4.4 adds some aspects on the importance of open data and code in machine learning for the environmental sciences. We restate the original research questions in cursive type to introduce each section. We then state the research findings in bold type, followed by a detailed explanation and rationale. Each section concludes with a concise answer to the research question.

4.1 Ozone mapping and interpolation

Research question 1: *"How can we use machine learning to map and interpolate ozone from existing measurements to gapless data of any required spatio-temporal resolution?"*

A 1.1: Mapping and interpolation of ozone data with machine learning methods can yield highly accurate data products.
The third and fourth papers of this work contain two machine learning applications that produce ozone data products of high accuracy. The third paper (Betancourt et al., 2022) presents a machine learning approach to produce a global map of average ozone values of the years 2010 - 2014. The RMSE of the map in regions with good training data coverage is approximately 4 ppb, and regions with sparse or no training data have an RMSE of 4 to 5 ppb. To put these values into perspective, 5 ppb is comparable to the ozone bias in CTMs (DeLang et al., 2021), and therefore acceptable, and the accuracy of the map is high. The R^2 score of 0.55 shows that the model explains more than 50 % of the variance in the true ozone values. This value is also satisfactory given the long aggregation period of the ozone data and the simplified approach of predicting ozone using only geospatial data as input. Limiting the input features to static geospatial features underlines the proof-of-concept character of Betancourt et al. (2022). Even if the evaluation scores are acceptable with this relatively simple approach, it is clear that more accurate and time-resolved predictions will require the use of meteorological data, for example from numerical weather prediction models or reanalyses. While training localized models could also improve model accuracy, our experiments show that the limiting factor for mapping accuracy in a particular region is the available training data in that region, not the machine learning models. This finding will be the subject of further discussion later in this section.

As opposed to the proof-of-concept mapping paper, the fourth paper of this work (Betancourt et al., 2023) is a potential real-world application for missing data interpolation. Here we make time-resolved predictions of gaps in measured ozone time series, using all possible types of input features: geospatial data, meteorological data, CTM data, and emission fields. The R^2 values of the interpolation are very high with values between 0.87 and 0.97, depending on the gap type. This is a significant accuracy gain compared to the traditional statistical method multivariate nearest

neighbors and basic machine learning models (Betancourt et al., 2023). RMSEs are between 2.43 and 6.18 ppb, depending on the gap type. The higher RMSEs compared to the mapping paper is due to the higher variance in the time-resolved data, as the true values cover ozone values of up to a maximum of 113.18 ppb, while the multi-annual ozone averages of the AQ-Bench dataset used in Betancourt et al. (2022) have a maximum value of only 65.59 ppb. Both the third and fourth papers predicted extremes with less accuracy. This is a known issue in both traditional numerical modeling and machine learning (Young et al., 2018) and needs further improvement.

A 1.2: Random forest and graph machine learning are suitable machine learning methods for mapping and interpolation of ozone.

We use the machine learning algorithm random forest (Section 2.2.3) in all four papers of this work to make ozone predictions based on point data features with no spatio-temporal structure. The benefits of random forest are numerous. First of all, it has high evaluation scores, surpassing other algorithms such as neural network and multivariate nearest neighbors (Junninen et al., 2004) in our studies. Likewise, random forest was shown to excel deep learning models on tabular style data (Lundberg et al., 2020; Grinsztajn et al., 2022). Random forest has short training and inference times, and we found it to be insensitive to hyperparameters (Betancourt et al., 2022), compared to the neural network. We hypothesize that the insensitivity to hyperparameters is due to the fact that the complexity of the random forest adjusts to the complexity of the data when unlimited depth of the single trees is allowed. The properties of random forest allows for flexible testing of different scenarios and rapid execution of explainable machine learning experiments that require re-training the model. Betancourt et al. (2022) also show better generalizability of the random forest across world regions, than for the neural network. The random forest improves its generalizability through bootstrapping of the available training data (Breiman, 2001). Therefore, it is also insensitive to noise in the training data, and can also make robust predictions even on smaller datasets like AQ-Bench. We recommend exploring whether the benefits of increased generalizability of transformer architectures over random forest (as observed by Hickman et al. (2022), for example) also apply to mapping.

We found the graph machine learning algorithm correct and smooth to be useful to improve random forest predictions by making use of the spatio-temporal patterns inherent in the ozone measurements (Betancourt et al., 2023). The benefit of graph machine learning is that it can make use of data with any spatio-temporal structure such as irregularly placed ozone measurements with gaps. This is opposed to machine learning architectures like LSTM or CNN (Section 2.2), which require gap-free input on a regular grid. The suitability of graph machine learning for environmental problems has also recently become apparent as Lam et al. (2022) surpassed the accuracy of traditional numerical weather forecasts with their graph-based machine learning model GraphCast.

A 1.3: Ozone proxies chosen according to expert knowledge are suitable input features for machine learning models for ozone prediction.

We use an unprecedented abundance of static geospatial data as features for our machine learning models in all papers of this work. We chose the geospatial data according to a priori expert knowledge because they correspond to governing factors for ozone as summarized in Section 1.1. Using these data for ozone prediction is not straightforward because they are proxies to ozone

processes with an unknown connection to ozone. For example, features derived from *nightlight* point to a certain amount of traffic or industrial activity and therefore ozone precursor emissions, but not to a quantifiable amount of ozone precursors. Still, this feature proved to be important to make predictions. Our machine learning approach is therefore complementary to the traditional process-oriented CTM models for ozone (Rao et al., 2011; Schultz et al., 2018; Wagner et al., 2021).

The first paper of this work justifies the data selection and proves that ozone metrics can be predicted with only geospatial data as inputs with machine learning (Betancourt et al., 2021a). Following that, the third paper (Betancourt et al., 2022) applies a feature selection method (Meyer et al., 2018) to identify which features of the geospatial data are helpful to predict ozone. It sorts out a few features because they are not helpful, and there are plausible explanations for that. For example, the feature *snow and ice in 25 km area* is not represented well in the training data. The third paper also determines the importance and influence of the individual features on the ozone predictions using SHAP (Lundberg and Lee, 2017), which further demonstrates the consistency of our models with expert knowledge. For example, the positive influence of an increased *altitude* on the predicted ozone values is in line with previous ozone research (Chevalier et al., 2007). The fourth paper adds more input features to complement the geospatial data. Here, meteorological data and ozone precursor emission data, as well as CTM data are straightforward choices for inputs, which have been proven useful for ozone prediction in previous studies (Kleinert et al., 2021; Sayeed et al., 2020). The datetime features are more interesting in that regard, namely *day of the year*, *day of the week*, and *hour of the day*. These features are easy to obtain as they are derived from the time and date of each ozone measurement. They can represent diurnal, weekly, and yearly cycles in emission, weather, and available radiation (Section 1.1.1). These features were among the first picked by the feature selection method from Meyer et al. (2018), showing that they are comparably useful to meteorological data for predicting ozone. But, just as the geospatial data, they do not have a quantifiable connection to ozone values.

In summary, we showed that ozone proxies are useful as inputs for machine learning. Both expert knowledge and explainable machine learning agree on that finding.

A 1.4: Machine learning models offer flexible input and output resolution for ozone data products.

One benefit of machine learning is the flexible spatio-temporal resolution of the generated data products. The AQ-Bench dataset contains point data which are data at the exact measurement location. Therefore, the spatial resolution of a map generated with a model trained on AQ-Bench (like in Betancourt et al., 2022) depends only on the spatial resolution of the inputs used for the prediction, i.e., the gridded fields containing geospatial data. In the third paper, these are fields with a resolution of $0.1° \times 0.1°$. If higher-resolution input data would be available, the resolution could also be increased without retraining the model. Machine learning also offers flexibility in the temporal domain. In the third paper, we noted that a time-resolved map could be produced by adding time-resolved input features. In general, the finest resolution of inputs in both temporal and spatial domains is the finest resolution of a map that can be produced with that input data.

In principle, machine learning models can output any data product they are trained to output, and we make extensive use of that ability in this work. The AQ-Bench dataset contains ozone metrics and long-term statistics as targets, which are aggregated data products. If instead, the ap-

proach was to predict hourly ozone concentrations and then calculate metrics from those hourly concentrations, large amounts of data would need to be processed and appropriate software would need to be used. The prediction of aggregated metrics is also convenient when producing the ozone map of the third paper. The spatial resolution of 0.1°×0.1° and a global coverage results in about 6 million data points. Adding hourly resolution for the years 2010 - 2014 would increase the number of data points to nearly 300 billion, which in turn would have to be aggregated to obtain the desired metric. Therefore, our approach to create a map with long-term statistics saves time, data volume and compute capacity, and is less prone to error. This effect is diminished if the goal is to map seasonal, monthly, or daily statistics, which is a more realistic application.

A 1.5: The most important factor in the accuracy of ozone predictions with machine learning in a region is the measurements available in that region.
In both the mapping and interpolation papers (Betancourt et al., 2022; Betancourt et al., 2023) we identified good measurement coverage as a limiting factor for the accuracy of ozone predictions with machine learning. The third paper demonstrates global generalisability of our model, and therefore the ability to make predictions in a region with little or no training data, with a cross-validation experiment on different world regions. Although the model generalizes satisfactorily across world regions, we also note that using a model in the world region on which it was trained gave better predictions than using the model in a different world region. One implication of this is that it is impossible to make predictions in regions without measurement coverage as accurately as in regions with good coverage, and generalisability is limited. In addition, we flag gaps in the generated map in places with feature combinations that do not occur in the training data, because the model is not applicable there. These unknown feature combinations occur mainly in regions with sparse training data.

The fourth paper performs graph learning on the air quality monitoring network, where all nodes in the graph represent hourly measurements, and also supports the importance of data coverage. This paper proves that a well-connected node with many neighboring measurements available is more likely to be interpolated with high accuracy than a node with few or no neighbors. This is due to the fact that a well-connected node has more examples of true values (i.e. ozone measurements) from which it can learn. In summary, the findings of Betancourt et al. (2022) and Betancourt et al. (2023) come to the unanimous conclusion that good measurement coverage is crucial for an accurate ozone prediction and hints to emphasize the importance of measurement availability in all regions.

Apart from the data coverage, another problem hindering the accuracy of predictions is noisy data. There are influences on the ozone values that are not represented in the training data. This means even a well-trained model cannot make perfect ozone predictions. While the first paper just noted that the AQ-Bench dataset is noisy, the second paper conducts a thorough analysis of reasons why predictions went wrong. It assigns 238 of the 243 inaccurately predicted test samples to the group of “untrustworthy samples”. These are samples where faulty predictions cannot be pinned down to reasons like model fit or training data feature combinations. We hypothesize that these inaccurate predictions occur simply due to the fact that geospatial data cannot fully explain long-term ozone metrics. This problem is partly solved in the fourth paper where the correct step from correct and smooth accounts for these unresolved influences to ozone and therefore in-

creases the overall model accuracy. Nevertheless, it will never be possible to achieve one hundred per cent accuracy in ozone predictions with machine learning.

► **This work shows that mapping and interpolation of ozone using machine learning provides equal or higher accuracy compared to CTMs and traditional statistical methods, and offers flexibility in input and output resolution. We use random forest for point input data and graph-based machine learning on spatio-temporal ozone patterns, as these proved to be the most appropriate architectures due to their flexibility and accuracy. Their ability to use ozone proxies and measurements is particularly beneficial. We also state that machine learning and expert knowledge ideally go hand in hand in selecting appropriate input features. Finally, we note that a good spatio-temporal coverage with measurements is most crucial for gapless, accurate machine learning for ozone.**

4.2 Trustworthy machine learning

Research question 2: *"What machine learning methods can we develop or adapt to create ozone data products, and how can we make these data products trustworthy?"*

A 2.1: Out-of-the-box machine learning approaches are often not suitable for environmental science applications such as ozone mapping and interpolation.
Machine learning offers the ability to learn complex patterns, making it attractive for environmental applications. However, many out-of-the-box machine learning approaches suitable for tasks such as image recognition or natural language processing are not readily suited to environmental applications. We justify this statement below.

One aspect is the need for a proper evaluation strategy. It is necessary to test the predictive capabilities of a machine learning model on independent data (Section 2.2.1). Standard machine learning often relies on a random data split (Krizhevsky et al., 2012; Goodfellow et al., 2006). When data is sampled from a continuous domain, such as in machine learning on Earth system data, this results in non-independent data splits, and potential overestimation of the model accuracy (Section 2.2.1). Therefore, the first paper (Betancourt et al., 2021a, Section 3.1) conducts a two-step clustering (Section 2.2.4) approach for an independent data split. It relies on grouping nearby ozone observation stations and assigning the groups randomly to the training, test, and validation sets. The spatial distance of the stations is the measure of (in)dependence, and we consider stations with a distance of at least 50 km independent. The third paper (Betancourt et al., 2022, Section 3.3) goes one step further to prove the global applicability of the model. It conducts a two-step cross-validation approach based on different regional station groups, and, additionally, on stations grouped according to their continent. These two evaluation methods together allow us to estimate the error of regions with and without measurements available separately (4 and 5 ppb, respectively). We specifically tailored this evaluation technique to the application at hand, because there exists no standard machine learning procedure to evaluate the global ozone model and demonstrate its generalizability.

Like the third paper, the fourth paper of this work (Betancourt et al., 2023) needs a proper evaluation strategy. The problem here is spurious correlations in the data. Random data splitting of

single hourly time steps would be unacceptable to demonstrate the ability to interpolate longer gaps in the data. We address the evaluation issue by masking validation and test gaps of the same lengths as the missing measurement gaps we encounter in the preliminary UBA dataset. We, therefore, mask gaps ranging from 1 h to several months in length, as well as gaps at all stations simultaneously. This is another example of an evaluation strategy for a machine learning approach in environmental science needs that is carefully designed according to the statistical properties of the data and the task at hand.

Another caveat to machine learning on environmental data is that models trained on a specific dataset can only be applied to data sampled from the same statistical distribution as that training dataset. While this is not a problem for solving a game (Silver et al., 2016) or recognizing image data that comes from the same source as the training data (Krizhevsky et al., 2012), it is often a problem in environmental science. The Earth is a highly complex system, and due to the variety of conditions and environmental influences that occur on a global scale, it is obvious that a limited number of measurements can hardly cover all feature combinations. For example, there is no ozone measurement station at the summit of Mount Everest, so any machine learning model trained on ground-based ozone observations is not suitable for making predictions at such high altitudes. The third paper uses the area of applicability approach of Meyer and Pebesma (2021) to account for this generalization limitation. It is evident that any machine learning application in environmental science needs careful consideration of the input data on which the model will be used in production.

A 2.2: Transparent, interpretable machine learning architectures facilitate the uptake of machine learning for ozone research.

There are numerous deep and complex machine learning architectures (Szegedy et al., 2015; Goodfellow et al., 2006, e.g.). They have thousands of adaptive parameters and take a long time to train. Deep learning models are difficult to interpret due to their layered "deep" architecture. The machine learning counterpart to these overly complex models are simpler architectures such as a shallow neural network and random forest, or the simple message passing graph machine learning algorithm correct and smooth (Huang et al., 2020). We use these simpler architectures in all papers of this work, because they are more interpretable and flexible, and because they have a comparable accuracy to complex architectures such as deep neural networks or graph neural networks on our datasets. This is true for our datasets because they are comparatively small, ranging from thousands to millions of samples. Of course, large language model (Kaplan et al., 2020; Dis et al., 2023) would not be possible with a random forest. We argue that simple, transparent, and easily interpretable architectures are especially beneficial for starting machine learning on a new scientific topic, which in this work is ozone research. The following are some of the arguments based on our experience.

Choosing an algorithm for a machine learning problem is not trivial. One intransparent way is to simply try a number of architectures and chose the most accurate one. On the contrary, in Section 4.1 we explained in detail why we use a random forest. Similarly, we have explained why we use a graph architecture and why we defined the graph structure the way we did. In addition, we have tested our models against basic statistical methods such as linear regression and multivariate nearest neighbors to further justify the use of our machine learning models. We

did not explicitly state this in any of the papers, but deepening a network did not improve model accuracy on AQ-Bench. Similarly, using a graph neural network on the interpolated dataset of Betancourt et al. (2023) did not surpass the results of correct and smooth. We hypothesize that this is because we are in a data limited "small data" regime. Transparency of the model choice can lead to model interpretation which is beneficial for scientific machine learning. For example, knowing how the graph works on the observations, we can interpret and understand incorrect and correct predictions by attributing them to single and well-connected nodes, or to short and long gaps (Betancourt et al., 2023). This is a valuable insight especially because graph machine learning is not yet established to be used for missing data interpolation.

Another way of interpreting model predictions post-hoc is through SHAP (Lundberg and Lee, 2017, Section 2.2.2). SHAP is an out-of-the-box explainable machine learning method that works fast with tree-based architectures like random forest. It would be technically possible to use SHAP with deeper architectures, but with tree-based models, the SHAP values are much cheaper to compute and therefore faster to obtain. Also, SHAP values can be determined analytically for tree-based models, but not for (deep) neural networks. While SHAP values are interpretations, the SHAP package of python (Lundberg, 2021) suggests some ways to aggregate the interpretations into explanations. Following this, we use SHAP values in the second and third papers to understand how our models arrive at single predictions but also aggregate them further. We describe this progress further below and note that the availability of such an out-of-the-box interpretation technique is highly beneficial.

As mentioned above, transparency goes hand in hand with interpretability. The second paper of this work (Stadtler et al., 2022) visualizes the prediction modes of a shallow neural network and the random forest. This allows us to interpret individual predictions and, therefore, get an intuition about the prediction process of the model. For example, we can attribute individual nodes in the neural network to high or low ozone predictions. In addition, we can assess the training process by showing the underlying training samples of the predictions. This increases our confidence as environmental scientists in using machine learning as a tool.

A 2.3: Explaining machine learning on ozone data can lead to scientific insights.
In Section 2.2.2, we mentioned that part of explainable machine learning is the interpretation of individual predictions and that a consolidation of many interpretations is an explanation. Explanations are a prerequisite for scientific results, the ultimate goal of machine learning in environmental science. The process of progressing from interpretations of a transparent model to explanations and the derivation of scientific results is an important part of this work, and we outline some examples of that in the following.

The starting point for the second paper (Stadtler et al., 2022) is the AQ-Bench dataset and two basic machine learning models trained on that dataset. We trace the prediction process of both models in detail, namely through the activation patterns of individual predictions. Our first step is to visualize and interpret these patterns. By relating the activation patterns, and therefore the internal state of a model, to the underlying training data, we can analyze the failure and success of individual predictions in detail. For example, did the model predict a true value by chance, or could an incorrect prediction be attributed to missing patterns in the training data? While these analyses are interesting for gaining intuition and interpretation of individual predictions, only the further

aggregation of these interpretations proved valuable for answering questions about the AQ-Bench dataset as a whole. We point out limitations through underrepresented feature combinations of the whole training dataset and then suggest where new air quality monitoring stations should be built to overcome these limitations. This is a way of using explainable machine learning to make a recommendation and, therefore, use it for decision-making. On the other hand, this method can also point out samples in the training data that are redundant and therefore do little to improve the accuracy of the model. This insight can guide data acquisition. While the paper is based on a benchmark dataset and we designate it as relatively close to "prove of concept" in Figure 3.1, we believe that recommendations based on and supported by explainable machine learning methods like ours can make a contribution to guiding ongoing research and decision making in the future.

Another simple way to gain new insights into the model and the machine learning process is to look at the SHAP values. We determined these individual post-hoc interpretations for our random forest model predictions in the third paper (Betancourt et al., 2022). As in the previous example, we gain new insights by consolidating these individual interpretations into explanations. The simplest form of aggregation of SHAP values is to determine global feature importances by summing the absolute local contributions of the features to predictions. We use these global importances together with overview plots of the SHAP values to verify the scientific consistency of our trained model. The aggregated SHAP values also reveal the surprising fact that the model relies more on spatial features than on process-describing features. As mentioned in Section 3.3, the most important feature is the *absolute latitude*. The second and third most important are *relative altitude* and *altitude*. In contrast to these spatial features, we have included the chemical features NO_2 *column* and NO_x *emissions* as model inputs because they associated with the chemical processes from Figure 1.1. They have global importances below 5 %. This hints to the fact that our random forest cannot learn ozone processes but rather reflects the as-is state of ozone, correcting false expectations that one might have in light of previous knowledge of ozone processes. Again, it should be noted that this is mainly true for the small data regime, and may change when temporally resolved data are combined with a more complex or physics-guided machine learning architecture that can grasp such processes.

A 2.4: Explainable machine learning and uncertainty assessment of ozone predictions agree.
The second paper of this work (Stadtler et al., 2022) develops an explainable machine learning method for models trained on AQ-Bench. Likewise, the third paper (Betancourt et al., 2022) carries out multiple experiments of explainable machine learning and uncertainty analysis, all related to ozone mapping using models trained on the same dataset. The different experiments of the two papers allow looking at the dataset and models from different angles. Comparing their results, the most interesting finding is that the methods of explainable machine learning and uncertainty assessment agree, as we will detail in the following.

As the second and third papers both train their models on the AQ-Bench dataset, we can link the findings of the two papers. In the second paper, we point out that a "healthy" prediction is a prediction that is based on many samples (which increases robustness) and the right samples (to make an unbiased prediction). In other words, one finding of this paper is that it is crucial that there are many training samples available, that the model can learn of, i.e., that there is a high

density of training data in the feature space. We already described how that is mostly the case if the prediction is made in a region with good spatial coverage of measurements. We can link this method with the area of applicability method of the third paper, which sorts out locations with feature combinations that are not well represented in the training data. It is reasonable to hypothesize that regions that are applicable without gaps are well covered in measurements and that in turn, the borders of the area of applicability are regions that are not well covered. Proof of this hypothesis can be found in the uncertainty assessment experiments of the third paper, where we train the model under ozone fluctuations, and produce maps with these perturbed models. Through the sparsity of training data in the feature space at the borders of the area of applicability, we expect the model to be less robust, and thus more sensitive against the perturbations. This is the case, as indeed the influence of perturbations is much higher at the edges of the area of applicability than in its center.

The point we want to make is that conducting different explainable machine learning and uncertainty assessment experiments will yield a consistent picture of a model and its abilities. In a healthy, well-defined model, the different techniques will all arrive at the same findings. New research suggests that the two concepts of explainable machine learning and uncertainty assessment should be combined, as they both delve into the way a machine learning model works and behaves (Seuß, 2021).

► **The trustworthiness of machine learning products is a key focus of this work. We find that interpretable architectures like random forest and correct and smooth are beneficial for getting started with machine learning for ozone, as they make it easier to explain the models and their predictions. We also observe that machine learning evaluation for ozone needs to be carefully tailored to the problem at hand, rather than using out-of-the-box approaches. We could gain valuable insights into the AQ-Bench dataset by combining explainable machine learning with uncertainty assessment. An example of this is the consistent picture of map trustworthiness in the third paper (Betancourt et al., 2022).**

4.3 Use of spatio-temporal patterns

Research question 3: *"How can we use spatio-temporal patterns represented in geospatial data and ozone measurements effectively within machine learning models?"*

A 3.1: Hand-crafted spatial patterns are valuable inputs for machine learning for ozone.
A key problem with ozone prediction using geospatial data is their spatio-temporal patterns that need to be represented in the machine learning model. One option would be to present the raw geospatial data to the model in a gridded format and train the machine learning model to find the spatial patterns by itself. We hypothesize that the approximately 3400 training samples of AQ-Bench are not enough to train a complex model to recognize the various spatial patterns occurring globally. We, therefore, resorted to using hand-crafted features as inputs, which can be seen as a form of feature engineering (Duboue, 2020). As the features were designed using prior knowledge about ozone, we note that expert knowledge of ozone is a way to compensate for small training

datasets. We added the hand-crafted features to the AQ-Bench dataset and use them in all papers in this work. Below, we give some examples.

One feature is the *relative altitude* of a station. It is the difference between the minimum altitude within a 5 km radius of a station and the station altitude. From an ozone researcher's point of view, this feature is useful because local flow patterns are particularly influenced by the relative altitude of a site. Thus, they will be different on a plateau and a mountain top, even if the *altitude* is the same. In the third paper of this work (Betancourt et al., 2022), SHAP ranked it as the second most important feature. Other examples of spatial patterns in AQ-Bench are *population density* and *night light*. We use the *population density* and *night light* values at the stations and the maximum values within radii of 5 km and 25 km around the stations. These features can be proxies for human activity and for remote conditions. If, for example, the *population density* at a station is low, but the feature *max population density 5 km* is high, then suburban conditions are present. Similarly, if all three feature values are low, then remote conditions are present. Therefore, these hand-crafted features allow a more detailed perception of spatial patterns in the geospatial data than the raw data features.

As noted in Section 4.1, we chose all geospatial inputs according to expert knowledge of ozone and ozone impacts. The hand-crafted geospatial features were initially part of the TOAR database because the authors intended them to aid further statistical analysis of their ozone data. The use of these hand-crafted features to represent spatial patterns in geospatial data is a shortcut that greatly simplifies costly pattern recognition within machine learning. Their explanatory power, especially of spatial features, is high in our models, as discussed in Section 4.2.

A 3.2: Grasping spatio-temporal patterns in ozone measurements with machine learning requires a non-euclidean machine learning architecture.

The fourth paper of this work (Betancourt et al., 2023) aims at interpolating gaps in ozone measurements in Germany by using available neighboring ozone measurements. The ozone measurement data include spatio-temporal ozone patterns, such as unusually high concentrations in a region, and the diurnal cycle. One problem in machine learning on ozone measurements is that they are not available on a regular grid. Instead, ozone monitoring stations are irregularly placed (Figure 3.5). Some regions, such as the greater Berlin area, there is a station every few kilometers. In rural areas of Germany, the distances are much greater. Furthermore, there are gaps in the time series and not all stations have the gaps at the same time. Using these measurements as-is is impossible with many common machine learning architectures because they require inputs of fixed size. For example, random forest and a shallow neural network have a fixed number of input nodes, and input is required for all nodes. Similarly, LSTM and CNN need their inputs on a regular grid. The solution we identified is graph machine learning.

We use patterns in the measurements effectively with the graph machine learning algorithm correct and smooth to improve the predictions of a random forest. Although correct and smooth is a very simple way of exploiting the spatial patterns in irregular measurements, it proved to be highly effective. We show, for example, that ozone predictions at locations with many neighboring stations are less prone to error than at locations with few neighbors. However, a thorough analysis of the spatial patterns present in the ozone data, and how each pattern improves the predictions, is beyond the scope of this paper.

► **Spatial and spatio-temporal patterns in environmental data are complex. In this work, we successfully use two complementary ways of representing these patterns in machine learning models. First, we represent spatial patterns in geospatial data as hand-crafted input features. These require expert knowledge to design, but are compatible with standard machine learning models such as random forest. Second, we use the graph machine learning algorithm correct and smooth which is capable of representing patterns in ozone measurements, even if they are irregularly placed, i.e. non-Euclidean.**

4.4 Open data and code

4.1: FAIR data enables large-scale data-driven ozone research.
This work uses large amounts of ozone data available in the TOAR database (Schultz et al., 2017). For the AQ-Bench dataset alone, we aggregated approximately 200 million hourly values between the years 2010 and 2014 from 5577 stations all over the globe by using the TOAR data portal. We also retrieved 2 million hourly ozone values from Germany for the fourth paper of this work (Betancourt et al., 2023). The collection and preparation of such big data would be infeasible for a single researcher to complete. Therefore, this work would not have been possible without the data preparation done by the TOAR community.

TOAR data are, therefore, an example of how FAIR data (findable, accessible, interoperable, reusable, Wilkinson et al., 2016) can accelerate research. TOAR data are also reused in other studies, as shown by the database description paper by Schultz et al. (2017) which has been cited 191 times already[1]. This means, they are highly "reusable". Not only does the TOAR database contain all the measurements, and make them available ("findable" and "accessible" in FAIR) they also provide tools to access and further aggregate the data, and we benefit from these tools. This is an example of "interoperable" in FAIR.

4.2: Benchmark datasets accelerate machine learning research for ozone.
Machine learning developments have always been driven by benchmark datasets. Prominent examples are the MNIST (LeCun et al., 2010) and Imagenet (Deng et al., 2009) image recognition datasets. Users can download them from open repositories, practice machine learning basics or develop new state of the art methods. As machine learning in the environmental sciences is evolving, Ebert-Uphoff et al. (2017) noted that benchmark datasets are also needed in this field. One difficulty in this area is the high level of domain knowledge required to make sense of environmental data. They, therefore, noted that a benchmark dataset for environmental science requires a description of the underlying problem. With AQ-Bench (Betancourt et al., 2021a) we provided one of these datasets. We have based two further studies on this dataset (Stadtler et al., 2022; Betancourt et al., 2022), which were faster to accomplish because they did not require additional training data preparation. AQ-Bench has already been reused in other studies and recognised as a good example of environmental machine learning benchmark datasets (Balamurugan et al., 2022; Dueben et al., 2022).

[1] According to `https://scholar.google.com/`, the number of citations was determined on 9 February 2023

4.3: Open data and code in environmental machine learning fight the machine learning reproducibility crisis.

Machine learning suffers a reproducibility crisis (Hutson, 2018; Gibney, 2022). Researchers frequently develop new machine learning methods, that improve the current state-of-the-art, but the results cannot be reproduced by others. One reason is that the training data is often not openly available, and another is that the method is not described in every detail, and the code is not provided. This hampers further development of the methods, and also increases the error proneness, as researchers cannot control each other, or build on top of each other.

Reproducibility can be obtained only when the data and experiments are properly described. It is best if the data and code are accessible and documented (Pineau et al., 2021; Nature editorial, 2021; Gundersen et al., 2018). Therefore, we made code and data of this work accessible wherever possible[2]. This applies for the code and data in Betancourt et al. (2021), Betancourt et al. (2022) and Betancourt et al. (2023). The code of these papers is either available in an open repository or has a Digital Object Identifier (DOI). We also include a readme file and proper code licensing as best practice. The prepared training datasets are also openly available. Through python environments, readmes, and stable data repositories, we hope that our research stays reproducible.

► **We accelerate the advances in machine learning for ozone research through open data and code. For example, we use FAIR data from TOAR to conduct studies on global ozone measurements that would be impossible to gather alone. We have also made the AQ-Bench benchmark dataset openly available, including extensive code and documentation, to better meet the needs of machine learners. Lastly, we followed best practices of open data and code to advance machine learning in the field of ozone research and make it reproducible.**

[2]Stadtler et al. (2022) does not provide open data as this paper only uses the AQ-Bench dataset already published with the Betancourt et al. (2021). It does not provide open code either, as we originally planned to release the machine learning method described in the paper as a separate software package.

5. Conclusion and new research directions

Tropospheric ozone is a toxic greenhouse gas with a highly variable spatio-temporal distribution. To better assess the distribution and impacts of ozone in places where measurements are not available, we have developed machine learning methods for spatio-temporal mapping and interpolation of ozone.

To perform the mapping and interpolation, we fused available ozone measurements with geospatial and meteorological features in machine learning models. We used the large amount of ozone measurements available from the Tropospheric Ozone Assessment Report database (TOAR, Schultz et al., 2017) as the main FAIR data source. We first compiled the machine learning benchmark dataset AQ-Bench with long-term ozone statistics (years 2010 - 2014) and geospatial data from the TOAR database to show that it is possible to predict ozone using geospatial features as inputs. In a second step, we explained how machine learning models can learn from AQ-Bench. We then performed static global high-resolution (0.1°$\times$ 0.1°) mapping using a machine learning model trained on AQ-Bench. Although the method of relying only on geospatial features for ozone mapping is a proof of concept rather than a mature ozone modeling approach, the resulting map captures known ozone patterns and has a low bias of 4 - 5 ppb. In contrast to the static mapping, we then performed hourly interpolations of missing ozone measurements for the year 2011 in Germany by including time-resolved meteorological input data. The index of agreement of our interpolation method ranges from 0.96 to 0.99 depending on the gap characteristics. This is a higher interpolation accuracy than the traditional statistical methods to which we compared.

The random forest machine learning architecture proved valuable for both mapping and interpolation. We note that the AQ-Bench training dataset we provided is in the small data regime. We have found that random forest is a very suitable architecture for a training dataset of this size because it offers interpretability, robustness, and flexibility. We used spatial patterns in the geospatial data, such as land cover or population density, at different radii around an air quality monitoring station through feature engineering. This simplified approach proved to be valuable. For interpolation, we used a more sophisticated approach to incorporate spatio-temporal patterns of ozone measurements, namely graph machine learning on the irregularly placed stations of the air quality monitoring network.

Rather than using machine learning as a black box, we made our ozone data products trustworthy through proper evaluation, explainable machine learning, and uncertainty estimation. For example, we used SHAP values to explain model predictions, evaluated our models using a spatial cross-validation approach, and developed a new explainable machine learning method. As a result, we were able to provide pixel-wise uncertainty estimates for the global map. We also pointed out regions in the global domain where our machine learning model is not applicable because they have characteristics too different from the training data. A result of the different methods is also that the accuracy of the mapping and interpolation is mainly ozone data limited. Regional characteristics and their relationship to ozone patterns must be learned from observational data to

provide a good estimate of ozone distribution. Generalization to regions with lower measurement density is possible, but with the trade-off of lower accuracy of ozone predictions.

Based on our experience, we find that it is beneficial to include explainable machine learning in environmental machine learning projects. We recommend including similar techniques in the big data regime and when using deeper machine learning architectures, even though this may be more difficult to realize. We also note that domain knowledge of the problem at hand is essential for successful machine learning. Domain knowledge enters the machine learning pipeline at several points, from selecting appropriate input data, to finding appropriate machine learning algorithms, to finally being able to evaluate and make sense of the results. Although our results are promising in terms of accuracy and trustworthiness of the generated data products, we stay relatively close to proof-of-concept applications and do not contribute to the generation of operational or reusable ozone data products. The focus of the method development and evaluation in this work is on ozone, but we expect that the methods we have developed will be transferable to other environmental variables because they have similar statistical properties to ozone. With this work, we hope to contribute to the further establishment of machine learning for ozone, air quality, and environmental applications on local to global scales, and from hourly to long-term time scales.

Below we propose three new research ideas that further develop the results of this work. Each of the following three paragraphs builds on one of the three research objectives from Section 1.2.

As one new research direction, we propose building an operational machine learning-based service that provides state-of-the-art ozone data products to end users. In this work, we have developed machine learning methods for ozone mapping and interpolation as a first research objective. It gradually progresses from full proof-of-concept studies to real-world applications (Figure 3.1). The next step is implementing an operational machine learning model for the generation of ozone data products. We propose to produce maps of 96 h ozone forecasts within the service because these data products are routinely provided by CTM (Marécal et al., 2015). The service would be an economical alternative to the current CTMs, which provide similar data products but are computationally costly to operate. Two machine learning steps are necessary for this: 1) derive the is-state of the ozone from real-time measurements by mapping, 2) generate gridded fields of ozone forecasts by performing forward calculations of this is-state in real-time. Machine learning methods that can accomplish these tasks have recently been developed in this work (Betancourt et al., 2022), and by Kleinert et al. (2021) and Leufen et al. (2022), who develop machine learning-based ozone forecast models. To create a service that combines these works, the machine learning models need to be fused into a single framework and brought to a common domain. We recommend aiming for global coverage, but as we detailed in Section 4.1, the spatial domain where reliable data products are available will be limited to regions with good measurement coverage. Modifying the method of Betancourt et al., 2022 is, therefore, necessary as it provides static instead of hourly resolved mapping. Likewise, Leufen et al. (2022) needs modification as they provide forecasts for station locations in central Europe, not over a global gridded domain. More training data and thorough evaluation are needed to generalize their machine learning method to other world regions. We have mentioned how the accuracy of our models is comparable to or better than numerical models (Section 4.1). Similarly, Leufen et al. (2022) surpassed the forecast accuracy of a CTM with their machine learning model. Training this fused machine learning model will require substan-

tial computing resources. But once trained, the computational cost of the operational machine learning ozone service is expected to be much lower than that of a complex operational CTM since predictions of the machine learning models are computationally cheap. We expect this approach to work for other pollutants as well.

Our second proposal for new research is to explore causal relations between ozone and geospatial features, as an addition to explainable machine learning. We have used explainable machine learning in the scope of the second research objective from Section 1.2 to increase trust in our machine learning models. Yet, the explained models are still of pure statistical nature, neglecting the difference between causality and correlation. As Schölkopf (2022) notes, "A causal model [thus] contains genuinely more information than a statistical one". Therefore, exploring causality is a complementary approach to our way of using ozone proxies with no known connection to ozone. I could also improve the poor process understanding of our models that we note in Section 4.2. Causality exploration in environmental science is seen as a new and promising idea for using the large amounts of available data (Runge et al., 2019a). A first starting point to explore causality in big ozone datasets could be the works of Gerhardus and Runge, which enables to explore causality in time series data with their PCMCI algorithm (Runge et al., 2019b; Gerhardus and Runge, 2020). This algorithm could already resolve teleconnections for the reconstruction of the Walker circulation (Runge et al., 2019a) and midlatitude winter circulation (Kretschmer et al., 2016) from meteorological data. They provide the Python package tigramite (Runge et al., 2022) and the platform causeme[1] for an easy start with the PCMCI algorithm. Furthermore, they aim to generalize their algorithm so that it can be applied to time series of the same variables at different locations (Gerhardus, 2022, personal communication), which would be useful for the training data we have published with Betancourt et al. (2023). Applying PCMCI to these data is an easy way to combine expert knowledge, machine learning, and causality for ozone research.

Thirdly, we propose to use transformer architectures in ozone interpolation with machine learning. This work has presented two approaches to use spatiotemporal patterns in machine learning for ozone as an answer to the third research question (Section 1.2). In Section 4.3 we have detailed the two complementary approaches. One approach was to use completely hand-crafted geospatial features within a basic machine learning model. The other approach allows more flexibility in the machine learning model by using a graph machine learning algorithm. The ladder interpolates missing ozone data by considering all available measurements within the spatio-temporal range of 50 km and 6 h around a missing value. This graph-based approach allows flexibility in the number of neighboring measurements considered, but by ignoring all measurements outside this radius, it makes rather rigid assumptions about which measurements are important for the interpolation and which are not. Yet, from previous research (Section 1.1), we know that some ozone patterns cover larger spatio-temporal ranges, such as repeating diurnal patterns under the same meteorological conditions, or long-range transport. These larger patterns are impossible to capture with the current setup. One machine learning architecture that allows more flexibility is the transformer (Vaswani et al., 2017; Phuong and Hutter, 2022). Transformers were originally developed for natural language processing, where they replaced LSTM as the new state-of-the-art (Lin et al., 2021). The novelty is that instead of the fixed sequential order in which an LSTM con-

[1] `https://causeme.uv.es/`, last access 15 February 2022

siders the words in a sentence, a transformer makes it possible to relate all words to each other with the concept of attention. Applying this idea to ozone measurements and patterns, all available measurements on a domain could be used within a transformer, and their relevance to make a prediction for a specific place and time could be determined by its attention mechanism. Yet, this would require a large model and substantial computing resources, for the amount of available ozone measurements is in the order of millions (Section 4.1). Transformer-based models are currently adapted for meteorology (Nguyen et al., 2023), and for ozone forecasting (Hickman et al., 2022), and we expect these models to be well suited for ozone interpolation as well.

The benefits and challenges of machine learning in environmental science, and especially in air pollution, is a widely discussed topic. Researchers are beginning to shape its new developments following an increased use of these new techniques (Tuia et al., 2021; Hsieh, 2022; Schultz et al., 2021; Liu et al., 2022). In summary, we have presented state-of-the-art mapping and interpolation of tropospheric ozone with trustworthy machine learning. It demonstrates the value of machine learning for Earth system data, and ozone in particular. Machine learning for ozone may eventually complement or replace CTMs because it is highly accurate and computationally inexpensive.

A. Lists of figures and tables

Figures

Tables

B. Abbreviations

A answer (to a research question)

CART classification and regression trees

CC-BY Creative Commons Attribution

CNN convolutional neural network

CTM chemical transport model

d index of agreement

DBSCAN density-based spatial clustering of applications with noise

DOI Digital Object Identifier

EU European Union

FAIR findable, accessible, interoperable, reusable

h hour

IID independent with an identical statistical distribution

km kilometer

LSTM long short term memory neural network

ODE ordinary differential equation

ppb parts per billion

ppm parts per million

R^2 coefficient of determination

RMSE root mean square error

SHAP Shapley additive explanations

TOAR Tropospheric Ozone Assessment Report

UBA Umweltbundesamt

WHO World Health Organization

C. Bibliography

Adadi, A. and Berrada, M. (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

American Meteorological Society (2021). *Interpolation. Glossary of Meteorology*. URL: https://glossary.ametsoc.org/wiki/Interpolation.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin". In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 173–182. DOI: 10.48550/arXiv.1512.02595.

Archibald, A. T., Neu, J. L., Elshorbany, Y. F., Cooper, O. R., Young, P. J., Akiyoshi, H., Cox, R. A., Coyle, M., Derwent, R. G., Deushi, M., et al. (Dec. 2020). "Tropospheric Ozone Assessment Report: A critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100". In: *Elementa: Science of the Anthropocene* 8.1. 034. DOI: 10.1525/elementa.2020.034.

Ashmore, M., Emberson, L., Büker, P., Briolat, A., and Gillies, D. (2017). *Deposition of Ozone for Stomatal Exchange (DO_3SE) model [software]*. v3.1.0. URL: https://www.sei.org/projects-and-tools/tools/do3se-deposition-ozone-stomatal-exchange/.

Balamurugan, V., Balamurugan, V., and Chen, J. (2022). "Importance of ozone precursors information in modelling urban surface ozone variability using machine learning algorithm". In: *Scientific Reports* 12.1, pp. 1–8. DOI: 10.1038/s41598-022-09619-6.

Barabási, A. (2016). *Network Science. Cambridge University Press, Cambridge (United Kingdom)*. URL: http://networksciencebook.com.

Bentley, J. L. (1975). "Multidimensional Binary Search Trees Used for Associative Searching". In: *Communications of the ACM* 18.9, pp. 509–517. DOI: 10.1145/361002.361007.

Betancourt, C. (2021). *webDO_3SE [web service]*. v1. URL: https://toar-data.fz-juelich.de/do3se/api/v1/.

Betancourt, C., Stomberg, T. T., Roscher, R., Schultz, M. G., and Stadtler, S. (2021a). "AQ-Bench: a benchmark dataset for machine learning on global air quality metrics". In: *Earth System Science Data* 13.6, pp. 3013–3033. DOI: 10.5194/essd-13-3013-2021.

Betancourt, C., Küppers, C., Piansawan, T., Sager, U., Hoyer, A. B., Kaminski, H., Rapp, G., John, A. C., Küpper, M., Quass, U., et al. (2021b). "Firewood residential heating – local versus remote influence on the aerosol burden". In: *Atmospheric Chemistry and Physics* 21.8, pp. 5953–5964. DOI: 10.5194/acp-21-5953-2021.

Betancourt, C., Hagemeier, B., Schröder, S., and Schultz, M. G. (2021c). "Context aware benchmarking and tuning of a TByte-scale air quality database and web service". In: *Earth Science Informatics* 14.3, pp. 1597–1607. DOI: 10.1007/s12145-021-00631-4.

Betancourt, C., Li, C. W. Y., Kleinert, F., and Schultz, M. G. (2023). "Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data". In: *Environmental Science & Technology* 57.46, pp. 18246–18258. DOI: 10.1021/acs.est.3c05104.

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., and Stadtler, S. (2022). "Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties". In: *Geoscientific Model Development* 15.11, pp. 4331–4354. DOI: 10.5194/gmd-15-4331-2022.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. 1st ed. Springer. ISBN: 978-0387-31073-2.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., et al. (2015). "Towards a high-resolution regional reanalysis for the European CORDEX domain". In: *Quarterly Journal of the Royal Meteorological Society* 141.686, pp. 1–15. DOI: 10.1002/qj.2486.

Brasseur, G., Orlando, J. J., and Tyndall, G. S., eds. (1999). *Atmospheric chemistry and global change*. 1st ed. New York, US: Oxford University Press. ISBN: 0-19-510521-4.

Brasseur, G. P., Prinn, R. G., and Pszenny, A. A. P., eds. (2003). *Atmospheric chemistry in a changing world: an integration and synthesis of a decade of tropospheric chemistry research*. Springer. ISBN: 978-3540430506.

Breiman, L. (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. 1st ed. Routledge. ISBN: 9781315139470.

Cabaneros, S. M., Calautit, J. K., and Hughes, B. R. (2019). "A review of artificial neural network models for ambient air pollution prediction". In: *Environmental Modelling & Software* 119, pp. 285–304. DOI: 10.1016/j.envsoft.2019.06.014.

Chapelle, O., Schölkopf, B., and Zien, A., eds. (2006). *Semi-Supervised Learning*. MIT Press. ISBN: 978-0-262-03358-9.

Chevalier, A., Gheusi, F., Delmas, R., Ordóñez, C., Sarrat, C., Zbinden, R., Thouret, V., Athier, G., and Cousin, J.-M. (2007). "Influence of altitude on ozone levels and variability in the lower troposphere: a ground-based study for western Europe over the period 2001–2004". In: *Atmospheric Chemistry and Physics* 7.16, pp. 4311–4326. DOI: 10.5194/acp-7-4311-2007.

Comes, F. J. (1994). "Recycling in the Earth's Atmosphere: The OH Radical—Its Importance for the Chemistry of the Atmosphere and the Determination of Its Concentration". In: *Angewandte Chemie International Edition in English* 33.18, pp. 1816–1826. DOI: https://doi.org/10.1002/anie.199418161.

Comrie, A. C. (1997). "Comparing Neural Networks and Regression Models for Ozone Forecasting". en. In: *Journal of the Air & Waste Management Association* 47.6, pp. 653–663. DOI: 10.1080/10473289.1997.10463925.

Cressie, N. (1988). "Spatial prediction and ordinary kriging". In: *Mathematical Geology* 20.4, pp. 405–421. DOI: 10.1007/BF00892986.

DeLang, M. N., Becker, J. S., Chang, K.-L., Serre, M. L., Cooper, O. R., Schultz, M. G., Schröder, S., Lu, X., Zhang, L., Deushi, M., et al. (2021). "Mapping Yearly Fine Resolution Global Surface Ozone through the Bayesian Maximum Entropy Data Fusion of Observations and Model Output for 1990–2017". In: *Environmental Science & Technology* 55.8, pp. 4389–4398. DOI: 10.1021/acs.est.0c07742.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (June 2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

Dis, E. A. van, Bollen, J., Zuidema, W., Rooij, R. van, and Bockting, C. L. (2023). "ChatGPT: five priorities for research". In: *Nature* 614.7947, pp. 224–226. DOI: 10.1038/d41586-023-00288-7.

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. 1st ed. Cambridge University Press. ISBN: 9781108671682. DOI: 10.1017/9781108671682.

Duda, R. O. and Hart, P. E. (1973). *Pattern recognition and scene analysis*. 1st ed. Wiley, New York. ISBN: 978-0471223610.

Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., and McGovern, A. (2022). "Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook". In: *Artificial Intelligence for the Earth Systems* 1.3, e210002. DOI: 10.1175/AIES-D-21-0002.1.

Ebert-Uphoff, I., Thompson, D. R., Demir, I., Gel, Y. R., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E., and Smyth, P. (2017). "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. URL: https://par.nsf.gov/biblio/10143795.

Emberson, L. (2020). "Effects of ozone on agriculture, forests and grasslands". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378.2183, p. 20190327. DOI: 10.1098/rsta.2019.0327.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Vol. 96. 34, pp. 226–231.

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". In: *Official Journal of the European Union* OJ L.152, pp. 1–44. URL: http://data.europa.eu/eli/dir/2008/50/oj.

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., et al. (2018). "Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health". In: *Elementa: Science of the Anthropocene* 6.1, p. 12. DOI: 10.1525/elementa.273.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P. .-., Cuesta, J., Cuevas, E., et al. (2018). "Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation". In: *Elementa: Science of the Anthropocene* 6.1, p. 39. DOI: 10.1525/elementa.291.

Gerhardus, A. (2022). *Reliable causal discovery in time series*. Workshop on Artificial Intelligence, Causality and Personalised Medicine (AICPM2022) – September 8-9, Hannover, Germany [presentation].

Gerhardus, A. and Runge, J. (2020). “High-recall causal discovery for autocorrelated time series with latent confounders”. In: *Advances in Neural Information Processing Systems* 33, pp. 12615–12625. DOI: 10.48550/arXiv.2007.01884.

Gibney, E. (2022). “Could machine learning fuel a reproducibility crisis in science?” en. In: *Nature* 608.7922, pp. 250–251. DOI: 10.1038/d41586-022-02035-w.

Goetz, S. J., Baccini, A., Laporte, N. T., Johns, T., Walker, W., Kellndorfer, J., Houghton, R. A., and Sun, M. (2009). “Mapping and monitoring carbon stocks with satellite observations: a comparison of methods”. In: *Carbon Balance and Management* 4.2, pp. 1–7. DOI: 10.1186/1750-0680-4-2.

Goodfellow, I., Bengio, Y., and Courville, A. (Nov. 2006). *Deep Learning*. 1st ed. Adaptive Computation and Machine Learning series. Cambridge, UK: The MIT Press. ISBN: 0262035618.

Google Maps (2022). *Google Maps excerpt of the city of Jülich*. retrieved 21 September 2022. URL: https://www.google.de/maps.

Granier, C., Darras, S., Gon, H. D. van der, Doubalova, J., Elguindi, N., Galle, B., Gauss, M., Guevara, M., Jalkanen, J., Kuenen, J., et al. (2019). “The Copernicus Atmosphere Monitoring Service global and regional emissions (April 2019 version)”. In: *Copernicus Atmosphere Monitoring Service (CAMS) [report]* 4. DOI: 10.24380/d0bn-kx16.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). “Why do tree-based models still outperform deep learning on tabular data?” In: *arXiv [preprint]*. DOI: 10.48550/arXiv.2207.08815.

Gundersen, O. E., Gil, Y., and Aha, D. W. (2018). “On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications”. In: *AI Magazine* 39.3, pp. 56–68. DOI: 10.1609/aimag.v39i3.2816.

Hamilton, W. L. (2020). “Graph representation learning”. In: *Synthesis Lectures on Artifical Intelligence and Machine Learning* 14.3, pp. 1–159.

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., Moninger, W., Lakshmanan, V., Tissot, P., and Williams, J. K. (2022). “The History and Practice of AI in the Environmental Sciences”. In: *Bulletin of the American Meteorological Society* 103.5, E1351–E1370. DOI: 10.1175/BAMS-D-20-0234.1.

Henschel, S., Chan, G., Organization, W. H., et al. (2013). “Health risks of air pollution in Europe-HRAPIE project: new emerging risks to health from air pollution-results from the survey of experts”. In: *Copenhagen: WHO Regional Office for Europe*. URL: https://apps.who.int/iris/handle/10665/108632.

Hickman, S., Griffiths, P., Archibald, A., Nowack, P., and Alhajjar, E. (2022). “Forecasting European Ozone Air Pollution With Transformers”. In: URL: https://www.climatechange.ai/papers/neurips2022/33.

Hochreiter, S. and Schmidhuber, J. (1997). “Long short-term memory”. In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Houghton, J. T., Ding, Y. D. J. G., Griggs, D. J., Noguer, M., Linden, P. J. van der, Dai, X., Maskell, K., and Johnson, C. (2001). *Climate change 2001: the scientific basis: contribution of Working Group I to the third assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. ISBN: 0521 80767 0.

Hsieh, W. W. (2009). *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. Cambridge University Press. DOI: 10.1017/CBO9780511627217.

Hsieh, W. W. (2022). "Evolution of machine learning in environmental science—A perspective". In: *Environmental Data Science* 1, e3. DOI: 10.1017/eds.2022.2.

Huang, Q., He, H., Singh, A., Lim, S.-N., and Benson, A. R. (2020). "Combining label propagation and simple models out-performs graph neural networks". In: *arXiv preprint arXiv:2010.13993*. DOI: 10.48550/ARXIV.2010.13993.

Hutson, M. (2018). "Artificial intelligence faces reproducibility crisis". In: *Science* 359.6377, pp. 725–726. DOI: 10.1126/science.359.6377.725.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., et al. (2019). "The CAMS reanalysis of atmospheric composition". In: *Atmospheric Chemistry and Physics* 19.6, pp. 3515–3556. DOI: 10.5194/acp-19-3515-2019.

IPCC (2022). *Global Warming of 1.5°C: IPCC Special Report on impacts of global warming of 1.5°C above pre-industrial levels in context of strengthening response to climate change, sustainable development, and efforts to eradicate poverty*. en. 1st ed. Cambridge University Press. DOI: 10.1017/9781009157940.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). "Methods for imputation of missing values in air quality data sets". In: *Atmospheric Environment* 38.18, pp. 2895–2907. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2004.02.026.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361*. DOI: 10.48550/arXiv.2001.08361.

Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D., and Tessum, C. W. (2020). "Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System". In: *Journal of Geophysical Research: Atmospheres* 125.23, e2020JD032759. DOI: 10.1029/2020JD032759.

Kington, J. A. (1990). "Daily weather mapping from 1781". In: *Climatic Change* 3.1, pp. 7–36. DOI: 10.1007/BF00144983.

Kleinert, F., Leufen, L. H., and Schultz, M. G. (2021). "IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany". In: *Geoscientific Model Development* 14.1, pp. 1–25. DOI: 10.5194/gmd-14-1-2021.

Kretschmer, M., Coumou, D., Donges, J. F., and Runge, J. (2016). "Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation". In: *Journal of Climate* 29.11, pp. 4069–4081. DOI: 10.1175/JCLI-D-15-0654.1.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., et al. (2022). "GraphCast: Learning skillful medium-range

global weather forecasting". In: *arXiv preprint arXiv:2212.12794*. DOI: 10.48550/arXiv.2212.12794.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). "Unmasking Clever Hans predictors and assessing what machines really learn". In: *Nature Communications* 10.1, pp. 1–8. DOI: 10.1038/s41467-019-08987-4.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

LeCun, Y. (1988). "A theoretical framework for back-propagation". In: *Proceedings of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA*. Ed. by D. Touretzky, G. Hinton, and T. Sejnowski. Morgan Kaufmann, pp. 21–28.

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

Lefohn, A. S., Knudsen, H. P., Logan, J. A., Simpson, J., and Bhumralkar, C. (1987). "An evaluation of the kriging method to predict 7-h seasonal mean ozone concentrations for estimating crop losses". In: *JAPCA* 37.5, pp. 595–602. DOI: 10.1080/08940630.1987.10466247.

Leskovec, J. (2021). *Stanford University CS224W: Machine Learning with Graphs [lecture]*. URL: http://web.stanford.edu/class/cs224w/.

Leufen, L. H., Kleinert, F., and Schultz, M. G. (2022). "O3ResNet: A deep learning based forecast system to predict local ground-level daily maximum 8-hour average ozone". In: *Artificial Intelligence for the Earth Systems* 1. Publisher: American Meteorological Society. URL: https://b2share.fz-juelich.de/api/files/40a46d8a-ba80-4e12-a01d-fdf61a64f6b9/Leufen%20et%20al%20(2022)%20O3ResNet%20Preprint.pdf.

Li, J., Siwabessy, J., Huang, Z., and Nichol, S. (2019). "Developing an Optimal Spatial Predictive Model for Seabed Sand Content Using Machine Learning, Geostatistics, and Their Hybrid Methods". In: *Geosciences* 9.4. DOI: 10.3390/geosciences9040180.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2021). "A survey of transformers". In: *arXiv [preprint]*. DOI: 10.48550/arXiv.2106.04554.

Linnainmaa, S. (1970). "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors". In: *Master's Thesis (in Finnish), University of Helsinki*. URL: https://people.idsia.ch/~juergen/linnainmaa1970thesis.pdf.

Liu, L.-J. S. and Rossini, A. (1996). "Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto—a pilot study". In: *Environment International* 22.6, pp. 677–692. DOI: 10.1016/S0160-4120(96)00059-1.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., and Bi, J. (2020). "Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach". In: *Environment International* 142, p. 105823. DOI: https://doi.org/10.1016/j.envint.2020.105823.

Liu, X., Lu, D., Zhang, A., Liu, Q., and Jiang, G. (2022). "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems". In: *Environmental Science & Technology* 56.4, pp. 2124–2133. DOI: 10.1021/acs.est.1c06157.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Lundberg, S. M. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017 proceedings)*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Long Beach, CA, USA, pp. 4765–4774. DOI: 10.48550/arXiv.1705.07874.

Lundberg, S. M. (2021). *shap [code]*. v0.38.1. URL: https://github.com/slundberg/shap/.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., et al. (2015). "A regional air quality forecasting system over Europe: the MACC-II daily ensemble production". In: *Geoscientific Model Development* 8.9, pp. 2777–2813. DOI: 10.5194/gmd-8-2777-2015.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al. (2021). "Climate change 2021: the physical science basis". In: *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, p. 2.

Mathieu, M., Couprie, C., and LeCun, Y. (2015). "Deep multi-scale video prediction beyond mean square error". In: *arXiv [preprint]*. DOI: 10.48550/arXiv.1511.05440.

McCulloch, W. S. and Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133. DOI: 10.1007/BF02478259.

McGovern, A., Lagerquist, R., Gagne, D., Jergensen, G., Elmore, K., Homeyer, C., and Smith, T. (2020). "Using machine learning and model interpretation and visualization techniques to gain physical insights in atmospheric science". In: *AI for Earth Sciences Workshop*. URL: https://ai4earthscience.github.io/iclr-2020-workshop/papers/ai4earth16.pdf.

Meyer, H. and Pebesma, E. (2021). "Predicting into unknown space? Estimating the area of applicability of spatial prediction models". In: *Methods in Ecology and Evolution* 12.9, pp. 1620–1633. DOI: 10.1111/2041-210X.13650.

Meyer, H. and Pebesma, E. (2022). "Machine learning-based global maps of ecological variables and the challenge of assessing them". en. In: *Nature Communications* 13.1, p. 2208. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29838-9.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation". In: *Environmental Modelling & Software* 101, pp. 1–9. DOI: 10.1016/j.envsoft.2017.12.001.

Mills, G., Pleijel, H., Malley, C. S., Sinha, B., Cooper, O. R., Schultz, M. G., Neufeld, H. S., Simpson, D., Sharps, K., Feng, Z., et al. (2018a). "Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation". In: *Elementa: Science of the Anthropocene* 6.1, p. 47. DOI: 10.1525/elementa.302.

Mills, G., Sharps, K., Simpson, D., Pleijel, H., Broberg, M., Uddling, J., Jaramillo, F., Davies, W. J., Dentener, F., Van den Berg, M., et al. (2018b). "Ozone pollution will compromise efforts to

increase global wheat production". In: *Global change biology* 24.8, pp. 3560–3574. DOI: 10.1111/gcb.14157.

Mills, G., Sharps, K., Simpson, D., Pleijel, H., Frei, M., Burkey, K., Emberson, L., Uddling, J., Broberg, M., Feng, Z., et al. (2018c). "Closing the global ozone yield gap: Quantification and cobenefits for multistress tolerance". In: *Global Change Biology* 24.10, pp. 4869–4893. DOI: 10.1111/gcb.14381.

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub, lulu.com. ISBN: 978-0244768522.

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., et al. (2015). "Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer". In: *Atmospheric Chemistry and Physics* 15.15, pp. 8889–8973. DOI: 10.5194/acp-15-8889-2015.

Nature editorial (2021). "Moving towards reproducible machine learning". In: *Nature Computational Science* 1.10, pp. 629–630. DOI: 10.1038/s43588-021-00152-6.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. (2023). "ClimaX: A foundation model for weather and climate". In: *arXiv preprint arXiv:2301.10343*. DOI: 10.48550/arXiv.2301.10343.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). "A Review of Relational Machine Learning for Knowledge Graphs". In: *Proceedings of the IEEE* 104.1, pp. 11–33. DOI: 10.1109/JPROC.2015.2483592.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A. (2018). "Evaluation of digital soil mapping approaches with large sets of environmental covariates". In: *SOIL* 4.1, pp. 1–22. DOI: 10.5194/soil-4-1-2018.

Omohundro, S. M. (1989). *Five balltree construction algorithms [technical report]*. International Computer Science Institute Berkeley.

Oxford Learner's Dictionary (2022). Oxford University Press. URL: oxfordlearnersdictionaries.com/.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., et al. (2014). *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Phuong, M. and Hutter, M. (2022). "Formal algorithms for transformers". In: *arXiv preprint*. DOI: 10.48550/arXiv.2207.09238.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., and Beygelzimer, A. (2021). "Improving Reproducibility in Machine Learning Research". In: *Journal of Machine Learning Research* 22. DOI: 10.48550/arXiv.2003.12206.

Rao, S. T., Galmarini, S., and Puckett, K. (2011). "Air Quality Model Evaluation International Initiative (AQMEII) advancing the state of the science in regional photochemical modeling and its appli-

cations". In: *Bulletin of the American Meteorological Society* 92.1, pp. 23–30. DOI: 10.1175/2010BAMS3069.1.

Ren, X., Mi, Z., and Georgopoulos, P. G. (2020). "Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States". In: *Environment International* 142, p. 105827. DOI: 10.1016/j.envint.2020.105827.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). "Explainable machine learning for scientific insights and discoveries". In: *IEEE Access* 8, pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019a). "Inferring causation from time series in Earth system sciences". In: *Nature communications* 10.1, p. 2553. DOI: 10.1038/s41467-019-10105-3.

Runge, J., Gillies, E., Strobl, E. V., and Palachy-Affek, S. (2022). *Tigramite [software]*. v5.1. URL: https://github.com/jakobrunge/tigramite.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019b). "Detecting and quantifying causal associations in large nonlinear time series datasets". In: *Science advances* 5.11, eaau4996. DOI: 10.1126/sciadv.aau4996.

Samuel, A. L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Sander, S. P., Golden, D. M., Kurylo, M. J., Moortgat, G. K., Wine, P. H., Ravishankara, A. R., Kolb, C. E., Molina, M. J., Finlayson-Pitts, B. J., Huie, R. E., et al. (2006). *Chemical kinetics and photochemical data for use in Atmospheric Studies Evaluation Number 15 [dataset]*. Version V1. DOI: 2014/41648.

Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., and Jung, J. (Jan. 2020). "Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance". In: *Neural Networks* 121, pp. 396–408. DOI: 10.1016/j.neunet.2019.09.033.

Schmitz, S., Towers, S., Villena, G., Caseiro, A., Wegener, R., Klemp, D., Langer, I., Meier, F., and Schneidemesser, E. von (2021). "Unravelling a black box: an open-source methodology for the field calibration of small air quality sensors". In: *Atmospheric Measurement Techniques* 14.11, pp. 7221–7241. DOI: 10.5194/amt-14-7221-2021.

Schölkopf, B. (2022). "Causality for machine learning". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. Ed. by H. Geffner, R. Dechter, and J. Y. Halpern, pp. 765–804. DOI: 10.1145/3501714.

Schultz, M. G., Stadtler, S., Schröder, S., Taraborrelli, D., Franco, B., Krefting, J., Henrot, A., Ferrachat, S., Lohmann, U., Neubauer, D., et al. (2018). "The chemistry–climate model ECHAM6.3-HAM2.3-MOZ1.0". In: *Geoscientific Model Development* 11.5, pp. 1695–1723. DOI: 10.5194/gmd-11-1695-2018.

Schultz, M. G., Akimoto, H., Bottenheim, J., Buchmann, B., Galbally, I. E., Gilge, S., Helmig, D., Koide, H., Lewis, A. C., Novelli, P. C., et al. (2015). "The Global Atmosphere Watch reactive gases measurement network". In: *Elementa: Science of the Anthropocene* 3.000067. DOI: 10.12952/journal.elementa.000067.

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. (1999). "On the origin of tropospheric ozone and NOx over the tropical South Pacific". In: *Journal of Geophysical Research: Atmospheres* 104.D5, pp. 5829–5843. DOI: 10.1029/98JD02309.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). "Can deep learning beat numerical weather prediction?" In: *Philosophical Transactions of the Royal Society A* 379.2194, p. 20200097. DOI: 10.1098/rsta.2020.0097.

Schultz, M. G., Kleinert, F., Leufen, L. H., Betancourt, C., Schröder, S., Gong, B., Stadtler, S., Langguth, M., and Mozaffari, A. (2022). "Artificial intelligence for air quality". In: *The Project Repository Journal* 12.1, pp. 70–73. DOI: 10.54050/PRJ1218384.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O., 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". In: *Elementa: Science of the Anthropocene* 5, p. 58. DOI: 10.1525/elementa.244.

Seinfeld, J. H. and Pandis, S. N. (2006). *Atmospheric Chemistry and Physics, From Air Pollution to Climate Change*. 1st ed. John Wiley & Sons Inc. ISBN: 0-471-17815-2.

Seuß, D. (2021). "Bridging the gap between explainable AI and uncertainty quantification to enhance trustability". In: *arXiv preprint arXiv:2105.11828*. DOI: 10.48550/arXiv.2105.11828.

Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T., et al. (2020). "SpaceNet 6: Multi-sensor all weather mapping dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 196–197. DOI: 10.48550/arXiv.2004.06500.

Silva, S. J., Heald, C. L., Ravela, S., Mammarella, I., and Munger, J. W. (2019). "A Deep Learning Parameterization for Ozone Dry Deposition Velocities". In: *Geophysical Research Letters* 46.2, pp. 983–989. DOI: 10.1029/2018GL081049.

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. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, p. 484. DOI: 10.1038/nature16961.

Singh, A. and Agrawal, M. (2007). "Acid rain and its ecological consequences". In: *Journal of Environmental Biology* 29.1, p. 15.

Skeie, R. B., Myhre, G., Hodnebrog, Ø., Cameron-Smith, P. J., Deushi, M., Hegglin, M. I., Horowitz, L. W., Kramer, R. J., Michou, M., Mills, M. J., et al. (2020). "Historical total ozone radiative forcing derived from CMIP6 simulations". In: *NPJ Climate and Atmospheric Science* 3.1, p. 32. DOI: 10.1038/s41612-020-00131-0.

Sonnewald, M. and Lguensat, R. (2021). "Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning". In: *Journal of Advances in Modeling Earth Systems* 13.8, e2021MS002496. DOI: 10.1029/2021MS002496.

Stadtler, S., Betancourt, C., and Roscher, R. (2022). "Explainable machine learning reveals capabilities, redundancy, and limitations of a geospatial air quality benchmark dataset". In: *Machine Learning and Knowledge Extraction* 4.1, pp. 150–171. DOI: 10.3390/make4010008.

Steffensen, J. F. (2013). *Interpolation*. 2nd ed. Dover Publications. ISBN: 9780486154831.

Sun, F., DAI, Y., and Yu, X. (2017). "Air pollution, food production and food security: A review from the perspective of food system". In: *Journal of Integrative Agriculture* 16.12, pp. 2945–2962. DOI: 10.1016/S2095-3119(17)61814-8.

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*, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

Tarasick, D., Galbally, I. E., Cooper, O. R., Schultz, M. G., Ancellet, G., Leblanc, T., Wallington, T. J., Ziemke, J., Liu, X., Steinbacher, M., et al. (2019). "Tropospheric Ozone Assessment Report: Tropospheric ozone from 1877 to 2016, observed levels, trends and uncertainties". In: *Elementa: Science of the Antropocene* 7.1, p. 39. DOI: 10.1525/elementa.376.

Tiao, G. C., Box, G. E. P., and Hamming, W. J. (1975). "Analysis of Los Angeles Photochemical Smog Data: A Statistical Overview". In: *Journal of the Air Pollution Control Association* 25.3, pp. 260–268. DOI: 10.1080/00022470.1975.10470082.

Tuia, D., Roscher, R., Wegner, J. D., Jacobs, N., Zhu, X., and Camps-Valls, G. (2021). "Toward a Collective Agenda on AI for Earth Science Data Analysis". In: *IEEE Geoscience and Remote Sensing Magazine* 9.2, pp. 88–104. DOI: 10.1109/MGRS.2020.3043504.

United Nations (2022). *The Sustainable Development Goals Report 2022*. United Nations Publications. ISBN: 978-92-1-101448-8.

Van Dingenen, R., Dentener, F. J., Raes, F., Krol, M. C., Emberson, L., and Cofala, J. (2009). "The global impact of ozone on agricultural crop yields under current and future air quality legislation". In: *Atmospheric Environment* 43.3, pp. 604–618. DOI: 10.1016/j.atmosenv.2008.10.033.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30. DOI: 10.48550/arXiv.1706.03762.

Voigt, S., Giulio-Tonolo, F., Lyons, J., Kučera, J., Jones, B., Schneiderhan, T., Platzeck, G., Kaku, K., Hazarika, M. K., Czaran, L., et al. (2016). "Global trends in satellite-based emergency mapping". In: *Science* 353.6296, pp. 247–252. DOI: 10.1126/science.aad8728.

Von Storch, H. and Zwiers, F. W. (2002). *Statistical analysis in climate research*. Cambridge university press. ISBN: 978-0521012300.

Wagner, A., Bennouna, Y., Blechschmidt, A.-M., Brasseur, G., Chabrillat, S., Christophe, Y., Errera, Q., Eskes, H., Flemming, J., Hansen, K. M., et al. (May 2021). "Comprehensive evaluation of the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis against independent observations: Reactive gases". In: *Elementa: Science of the Anthropocene* 9.1. DOI: 10.1525/elementa.2020.00171.

Wallace, J. and Hobbs, P. (Feb. 2006). *Atmospheric Science: An Introductory Survey*. 2nd ed. Vol. 92. International Geophysics Series. Burlington, MA, USA: Elsevier Academic Press, pp. 1–488. ISBN: 978-0127329512.

Wang, S. and Jiang, J. (2015). "Learning natural language inference with LSTM". In: *arXiv preprint arXiv:1512.08849*. DOI: 10.48550/arXiv.1512.08849.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2022). "Transformers in time series: A survey". In: *arXiv preprint arXiv:2202.07125*. DOI: 10.48550/arXiv.2202.07125.

Weng, X., Forster, G. L., and Nowack, P. (2022). "A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019". In: *Atmospheric Chemistry and Physics* 22.12, pp. 8385–8402. DOI: 10.5194/acp-22-8385-2022.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1, p. 160018. DOI: 10.1038/sdata.2016.18.

World Health Organization (2006). *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide : global update 2005 : summary of risk assessment*. URL: https://apps.who.int/iris/handle/10665/69477.

World Health Organization (2021). *Review of evidence on health aspects of air pollution: REVIHAAP project: technical report*. Technical documents, 302 p. URL: https://apps.who.int/iris/handle/10665/341712.

World Health Organization (2022). *4.2 million deaths every year occur as a result of exposure to ambient (outdoor) air pollution*. Accessed: 2021-12-12. URL: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1, pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.

Xu, J., Ma, J. Z., Zhang, X. L., Xu, X. B., Xu, X. F., Lin, W. L., Wang, Y., Meng, W., and Ma, Z. Q. (2011). "Measurements of ozone and its precursors in Beijing during summertime: impact of urban plumes on ozone pollution in downwind rural areas". In: *Atmospheric Chemistry and Physics* 11.23, pp. 12241–12252. DOI: 10.5194/acp-11-12241-2011.

Yi, J. and Prybutok, V. R. (1996). "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area". In: *Environmental pollution* 92.3, pp. 349–357. DOI: 10.1016/0269-7491(95)00078-X.

Young, P. J., Naik, V., Fiore, A. M., Gaudel, A., Guo, J., Lin, M., Neu, J., Parrish, D., Rieder, H., Schnell, J., et al. (2018). "Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends". In: *Elementa: Science of the Anthropocene* 6.10. DOI: 10.1525/elementa.265.

Zhao, Z., Chen, W., Wu, X., Chen, P. C., and Liu, J. (2017). "LSTM network: a deep learning approach for short-term traffic forecast". In: *IET Intelligent Transport Systems* 11.2, pp. 68–75. DOI: 10.1049/iet-its.2016.0208.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al. (2023). "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt". In: *arXiv preprint arXiv:2302.09419*. DOI: 10.48550/ARXIV.2302.09419.

Ziemke, J., Chandra, S., and Bhartia, P. (2005). "A 25-year data record of atmospheric ozone in the Pacific from Total Ozone Mapping Spectrometer (TOMS) cloud slicing: Implications for ozone trends in the stratosphere and troposphere". In: *Journal of Geophysical Research: Atmospheres* 110.D15. DOI: 10.1029/2004JD005687.

D. Papers

D.1 First paper (Betancourt et al., 2021a)

AQ-Bench: a benchmark dataset for machine learning on global air quality metrics

Clara Betancourt, Timo T. Stomberg, Ribana Roscher, Martin G. Schultz, and Scarlet Stadtler

Journal: Earth System Science Data (2021), Vol. 13, No. 6
Status: published (June 2021)
DOI: 10.5194/essd-13-3013-2021

Earth Syst. Sci. Data, 13, 3013–3033, 2021
https://doi.org/10.5194/essd-13-3013-2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Open Access Earth System Science Data

AQ-Bench: a benchmark dataset for machine learning on global air quality metrics

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

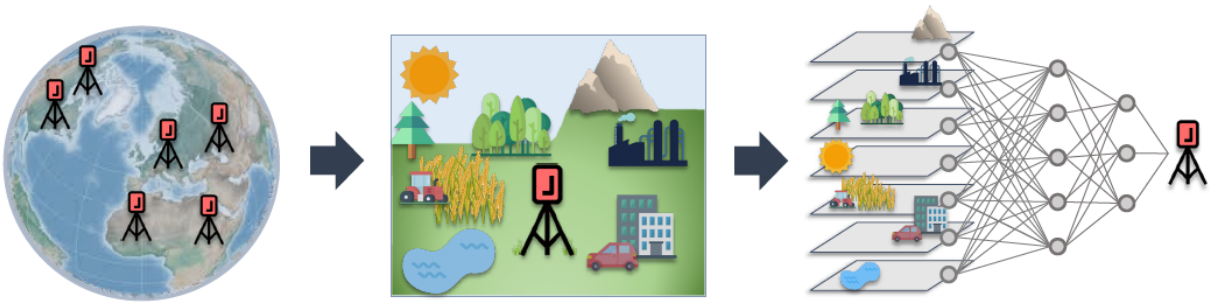
Received: 8 December 2020 – Discussion started: 14 January 2021
Revised: 9 April 2021 – Accepted: 20 May 2021 – Published: 24 June 2021

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. The purpose of this dataset is to produce estimates of various long-term ozone metrics based on time-independent local site conditions. We combine this task 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 and validation. 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/esde/machine-learning/aq-bench (Betancourt et al., 2021). AQ-Bench thus provides a blueprint for environmental benchmark datasets as well as an example for data re-use according to the FAIR principles.

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 has aroused the interest of environmental scientists in exploring 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, for example, in Wise and Comrie, 2005; Porter et al., 2015).

In air quality research, there is one pollutant which is especially challenging to track: tropospheric ozone, a toxic trace gas which harms human health and 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 airborne 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 inter-



The AQ-Bench dataset contains long-term air quality metrics and metadata at sites around the globe. Map by Wessel et al., 2019

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 long-term air quality metrics at measurement sites.

connected 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 (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 following days at individual locations. The principle behind recent machine learning applications in ozone research is often a similar principle to the one Schultz et al. (2021) described for weather data: the input data are directly mapped to a specific data product, e.g., from meteorological and past ozone measurements to the next day's maximum ozone value. In recent studies, Sayeed et al. (2020) and Kleinert et al. (2021) 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. Moreover, within computationally complex components of atmospheric chemistry models, machine learning techniques are used as emulators or surrogate models. They replace for example costly atmospheric chemistry and micro-physical calculations to improve computational performance of the models (Kelp et al., 2020). In addition, machine learning is applied in the calibration of low-cost sensors for air quality measurements in order to account for the diverse sources of interference with these measurements (Schmitz et al., 2021; Wang et al., 2020). Nevertheless, to our knowledge there are currently no machine learning projects that attempt to analyze and predict 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 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 these 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 long-term air quality metrics and metadata compiled from the TOAR database (Tropospheric Ozone Assessment Report; Schultz et al., 2017). To make these data usable for machine learning developments, this paper also describes the specific task of mapping between the metadata and the air quality metrics (see graphical abstract). 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/esde/machine-learning/aq-bench (Betancourt et al., 2021), 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 dataset. 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.

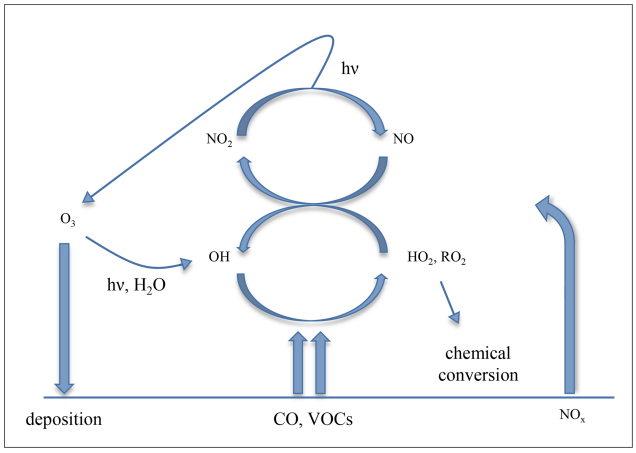


Figure 1. Simplified scheme describing the ozone chemical cycle. Figure adapted and modified from Jacob (2000). See text for elaboration.

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 they are 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).

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 Fig. 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 economic activities. Agriculture machinery emits similar trace gases to those emitted by 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. Agricultural fields, forests and grasslands therefore 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 order of 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 nonlinear system, where O_3 concentrations depend on the ratio of VOCs + CO and NO_x ($= NO + NO_2$) concentrations. During the 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 into this rather complex system. For more details on ozone chemistry we refer to Brasseur et al. (1999).

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 downward-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 (Schraudner et al., 1997), leading to reduced crop yields (Mills et al., 2011). 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 how the influences of ozone precursor emission, chemistry, transport and loss (described in Sect. 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 is 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 valleys 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" within which ozone is determined by nearby precursor emissions and deposition surfaces is typically about 25 km in mid-latitude areas (European Union, 2008). All in all, ozone concentrations measured at a station are determined by many interconnected influences from precursor emissions, land use and land cover, and the local weather conditions. Many of these factors are poorly quantified, and often the interconnections have not yet been understood well (Schultz et al., 2017). With AQ-Bench and the machine learning task described below, we want to explore a novel way of using a multitude of geographical features to predict ground-level ozone 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 the 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 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 C). 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. The data providers conduct quality control on these data by calibrating the measurement devices and setting suitable instrument parameters. In a second step of data curation, the TOAR database administrators conduct a statistical analysis of the data to identify and remove low-quality data (Schultz et al., 2017). 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) into 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

[1] https://join.fz-juelich.de/ (last access: 21 June 2021).

[2] https://join.fz-juelich.de/services/rest/surfacedata/ (last access: 21 June 2021).

[3] https://doi.org/10.1594/PANGAEA.876108 (last access: 21 June 2021).

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, basic statistics such as averages, medians and percentiles are also 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 Sect. 4. Documentation and further information on all available metrics including data capture criteria are available in Schultz et al. (2017) and Lefohn et al. (2018).

3.2 Station metadata

The TOAR database also contains geographical information on air quality measurement station locations, i.e., station metadata. Metadata give background information on the measurement site where the data were retrieved from and thus enable the characterization of the location. These metadata are collected from different sources. Some data, for instance station coordinates and altitude, are given by the data providers and quality controlled by TOAR. Others were derived from data sources with individual quality control, such as satellite Earth observations. For a complete list of the available metadata attributes see Schultz et al. (2017) and the REST interface (see footnote 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 Sect. 4.

4 AQ-Bench dataset description

The AQ-Bench dataset consists of metadata and aggregated 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 them to processes, and aggregate air quality data to metrics in a way that it is representative of long time periods and meaningful in a machine learning context.

Three key points in the conception of this benchmark dataset are as follows: (1) as targets, we use aggregated air quality metrics over 5 years. These are not influenced by short-term weather and emission forcings but by site conditions on the climatological timescale. (2) Many known environmental influences on ozone are on short timescales (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 contain 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 leveraged 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 account for long-term air quality conditions at a site, and daily or hourly influence on ozone variations is 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 originate from the TOAR database (Sect. 3); see Appendices A and C for details on the data sources and harmonization for machine learning purposes. The metadata contain proxies for environmental influences on ozone on the climatological scale. In the following, we give two examples.

As mentioned in Sect. 2, ozone is influenced by weather. Likewise, ozone on longer timescales is influenced by climate. One variable in the AQ-Bench dataset is the *climatic zone* in which the site is located. The climatic 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 is 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. This is the average intensity of light during the 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 give not only population density and stable nightlights 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 timescale. 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.

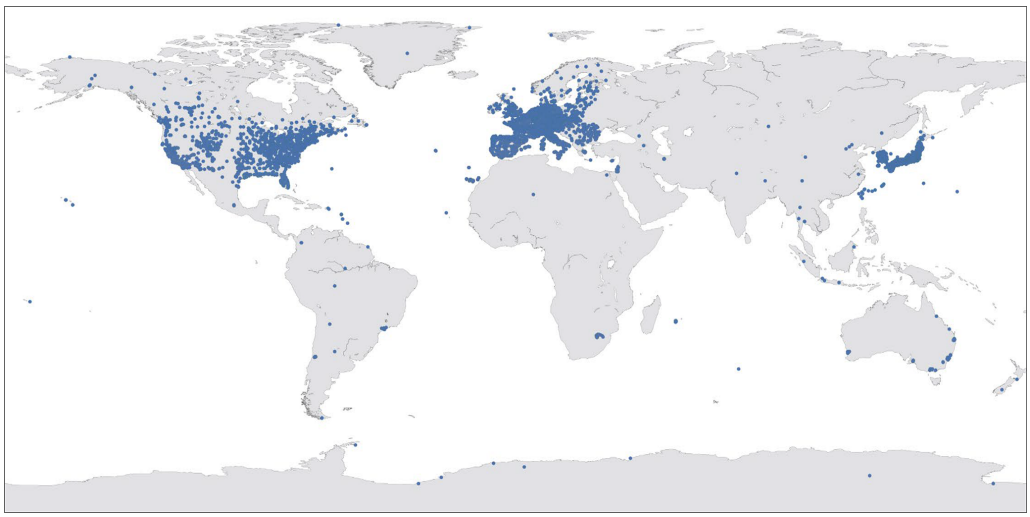


Figure 2. Worldwide measurement stations which are part of AQ-Bench, selected from the TOAR database. Map by Wessel et al. (2019).

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) obtaining up to five yearly metrics between 2010–2014 from hourly measurements, including data cover criteria to validate the metrics, and (2) averaging over these 5 years. If fewer 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 eliminated, as documented in Appendix C. A summary of all metrics and their data capture criteria is given in Table 2. More details on the process of ensuring robustness through data capture are given in Appendix B.

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 for our baselines is 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 denotes the total number of samples, $\hat{y}_m$ denotes a predicted output value and y_m denotes 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. For further evaluation of machine learning results, cross validation can be applied. 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 is commonly used for datasets of this size. It

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 (last access: 21 June 2021)
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 25 km 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 in ozone formation
Barren or sparsely vegetated in 25 km area	%	continuous	Emissions, deposition
Wheat production	1000 t	continuous	Agricultural emissions
Rice production	1000 t	continuous	Agricultural emissions
NO_x emissions	$g\,m^{-2}\,yr^{-1}$	continuous	NO_x emissions
NO_2 full column	10^5 molec. cm^{-2}	continuous	NO_2
Population density	persons km^{-2}	continuous	Human emissions
Max population density 5 km	persons km^{-2}	continuous	Human emissions nearby
Max population density 25 km	persons km^{-2}	continuous	Human emissions in area of influence

Table 1. Continued.

Variable	Unit	Type	Proxy for
Nightlight 1 km	brightness index	continuous	Industrial activity
Nightlight 5 km	brightness index	continuous	Industrial activity nearby
Max nightlight 25 km	brightness index	continuous	Industrial activity in area of influence

is indicated in the dataset whether an example belongs to the 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 to each other (European Union, 2008). Large station clusters were split again into smaller ones to ensure similar statistical distributions of the training, validation and test datasets. The final clusters were randomly assigned to the three datasets. This way, all stations within a spatially dependent cluster are allocated to the same dataset.

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 are scaled, each by a quantile range from 25 % to 75 % to avoid influence from outliers. Categorical metadata are one-hot encoded, resulting in 135 input features in total. We drop the *longitude* from our baseline experiments, since this is a circular variable and cannot be used without additional feature engineering. The preprocessed metadata are called input data in the following. Ozone metrics, which are the targets, are not scaled.

Methods are as follows:

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

- *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 an MSE (mean squared error) loss function, L2 regularization and ReLU (rectified linear unit) as the activation function (Goodfellow et al., 2016). Training is performed independently for each ozone metric. We optimized the learning rate and regularization parameter by empirical studies and random search. Through further empirical analyses, we decided on the hyperparameters 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, based on empirical studies. As in the case of the neural network, we use the MSE as an optimization criterion. 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 most targets except, e.g., *75th percentile*, where it is the 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-lowest scores. These two targets count exceedances of a certain threshold, so many values equal zero, which might be problematic for standard machine learning algorithms to capture. Except for 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 prove 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.

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

Table 2. The ozone metrics of AQ-Bench. The unit is ppb (parts per billion) for all metrics except the nvgt metrics, where it is the number of days.

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 1 year. At least 10 hourly values must be present to accept a daily median value as valid.	Basic statistics
25th percentile	25th percentile of daily values in 1 year. At least 10 hourly values must be present to accept a daily percentile value as valid.	Basic statistics
75th percentile	As "25th percentile" but for the 75th percentile.	Basic statistics
90th percentile	As "25th percentile" but for the 90th percentile.	Basic statistics
98th percentile	As "25th percentile" but for the 98th percentile.	Basic statistics
dma8eu	Daily maximum 8 h average statistics according to the EU definition. For 24 bins, 8 h averages are calculated 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 h. If fewer than 75 % of the data are present (i.e., less than 6 h), the average is considered missing. For annual aggregation, the 26th-highest daily 8 h 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 h 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-month running mean of daily maximum 1 h mixing ratios during the aggregation period of 1 year.	Human health
W90	Daily maximum W90 5 h experimental exposure index: EI = SUM($w_i C_i$) with weight $w_i = 1/[1 + M \exp(-AC_i/1000)]$, where M is 1400 and A is 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 h sums, requiring that at least 4 of the 5 h is present (75 %). If a sample consists of only four data points, a fifth value shall be constructed from averaging the four present mixing ratios. For annual aggregation, the fourth-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 the 08:00 until 19:59 solar time interval. AOT40 is defined as cumulative ozone above 40 ppb. If fewer than 75 % of hourly values (i.e., less than 9 out of 12 h) are present, the cumulative AOT40 is considered missing. When there exists 75 % or greater data capture in the daily 12 h window, the scaling by fractional data capture ($n_\mathrm{total}/n_\mathrm{present}$) is utilized. For annual statistics, the daily AOT40 values are accumulated over the aggregation period and scaled by ($n_\mathrm{total}/n_\mathrm{valid}$) 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
nvgt100	Number of days with exceedance of the daily max 1 h values above 100 ppb. The value is marked as missing if less than 75 % of days contain data.	Human health

Table 3. R^2 scores of the test set in percent. 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**
25th percentile	52.77	56.12	**62.75**
75th percentile	51.75	45.92	**55.65**
90th percentile	49.48	50.41	**58.54**
98th 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**

approaches. By doing this, we enable a quick entry into machine learning in air quality research on a global scale with reduced machine learning overhead. Our approach enables the use of 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 of using 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 (see Sect. 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. (2021), our approach is to output the demanded metrics directly and thus to obtain the required data products directly from machine learning. Further applications of AQ-Bench could be developed, such as a classification of ozone sites into "healthy" or "unhealthy". Our dataset fits with the vision for benchmark datasets described by Ebert-Uphoff et al. (2017).

[4]https://join.fz-juelich.de/ (last access: 21 June 2021).

6.2 Limitations of AQ-Bench

AQ-Bench includes ozone metrics and metadata from 5577 stations and spans a time period of 5 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 the 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 (cf. 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 are fixed and have no temporal evolution, an assumption which we can make because we average over 5 years of ozone metrics. It cannot be ruled out 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 5 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. We note that some uncertainty is introduced by the relatively lax requirement of two annual ozone metric values to form a

valid 5-year average value (see Appendix B): if both yearly averages correspond to the beginning or to the end of the time period in question, a bias may be introduced if the ozone concentrations exhibit a strong trend or if the region experienced rapid changes, such as urbanization.

Another topic is the complexity of the problem, compared to the dataset size. It is doubtful whether simple machine learning models are intricate enough to grasp all complex relationships between ozone and environmental factors. On the other hand, very deep neural networks, which may be capable of learning such patterns, cannot be trained on a dataset with only 5577 samples. In Sect. 5.3 we gave some basic machine learning approaches to find a mapping between the metadata and the target ozone metrics. We assume that the inaccuracies in our baselines partly arise from the complex relationships of ozone with the environment compared to the input dataset size and complexity of these basic machine learning approaches. Furthermore, through a longer aggregation period, we emphasize robust, static features. This aggregation reduces the size of the dataset and makes global coverage possible. Due to our focus on spatial relationships we consciously ignore time-resolved patterns. 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. For a more complete description of ozone processes, more input data, additional input variables and 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 machine 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 is not captured well. This is explained by the fact that there are very few ($< 25\,\%$) stations which experience occasional 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 http://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/esde/machine-learning/aq-bench (Betancourt et al., 2021). 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 the exploration of 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 and to optimize the results with hyperparameter tuning and data engineering. The usability of the dataset is documented through the results from our three baseline machine learning solutions. These methods show robust relations between the input data (geospatial features) and the targets (ozone metrics), and these relations are understandable from an atmospheric chemistry point of view. As data-driven techniques for air quality research are emerging, we present a first benchmark dataset on the global scale. The purpose and significance of AQ-Bench is twofold: first, it has never been tried before to exploit a rich collection of geospatial datasets to find out which fraction of ozone pollution can be attributed to such more or less static geographical features. Second, this problem definition makes some low-level air quality analysis easily accessible to data scientists with little or no background in atmospheric chemistry. 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 as follows:

- *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 can quickly iterate through different algorithms and configurations. Yet noise, the small size of the dataset 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/esde/machine-learning/aq-bench (last access: 21 June 2021). Running the introductory notebook on the 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 a basis for future experiments and research.

Appendix A: Technical details on the station metadata of the AQ-Bench dataset

Table A1. Technical details on the station metadata of the AQ-Bench dataset, updated from Schultz et al. (2017). Please note that in order to keep this table uncluttered, we have summarized all types of land cover in a 25 km area, all population density and all nightlight variables in one row each.

Variable	Data source	Reference
Country	Information given by data providers	
HTAP region	Derived from gridded data: Tier-1 regions from the Task Force on Hemispheric Transport of Air Pollution with an original resolution of 0.1°	Koffi et al. (2016)
Climatic zone	Derived from gridded data: IPCC 2006 classification scheme for default climate regions with a resolution of 5′	https://esdac.jrc.ec.europa.eu/projects/RenewableEnergy/ (last access: 23 Mar 2021)
Longitude	Information given by data providers. Quality controlled by TOAR database administrators	
Latitude	Information given by data providers. Quality controlled by TOAR database administrators	
Altitude	Information given by data providers. Quality controlled by TOAR database administrators	
Relative altitude	Derived from the ETOPO1 digital elevation model and the station altitude	Amante and Eakins (2009)
Type	Information given by data providers	
Type of area	Information given by data providers	
Land cover in 25 km area	Derived from gridded data: yearly land cover type L3 from the MODIS MD12C1 collection with an original resolution of 0.05°. The year 2012 and the IGBP classification scheme were used	https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD12C1/ (last access: 23 Mar 2021)
Wheat production	Derived from gridded data: annual wheat production of the year 2000 according to the Global Agro-Ecological Zones data, version 3, with an original resolution of 5′	https://www.fao.org/ (last access: 23 Mar 2021)
Rice production	Derived from gridded data: annual rice production of the year 2000 according to the Global Agro-Ecological Zones data, version 3, with an original resolution of 5′	https://www.fao.org/ (last access: 23 Mar 2021)
NO_x emissions	Derived from gridded data: annual NO_x emissions of the year 2010 from the EDGAR HTAP inventory V2 with an original resolution of 0.1°	Janssens-Maenhout et al. (2015)
NO_2 full column	Derived from gridded data: 5-year average (2011–2015) tropospheric NO_2 column value from the Ozone Monitoring Instrument (OMI) instrument on NASA's Aura with an original resolution of 0.1°	Krotkov et al. (2016)
Population density	Derived from gridded data: GPWv3 population density of the year 2010 with an original resolution of 2.5′	CIESIN (2005)
Nightlight	Derived from gridded data: stable nighttime lights of the year 2013 extracted from the NOAA DMSP product with an original resolution of 0.925 km	https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html (last access: 23 Mar 2021)

Appendix B: Data capture criteria

The data capture criteria applied in this work ensure robustness of the ozone metrics. Data capture criteria of hourly to annual metrics are applied through the JOIN web service (https://join.fz-juelich.de/, last access: 21 June 2021), as described in Schultz et al. (2017). The 5-year mean and its data capture criterion were applied in this work. One exception is the average value metric which does not have a data capture criterion in JOIN. Here we have verified that more than 2200 hourly values are processed to calculate the metric and that the average hourly data capture of all stations is above 50 %. The flowchart in Fig. B1 shows an example data capture criterion as applied in the AQ-Bench dataset. All data capture criteria are summarized in Table 2 of this work.

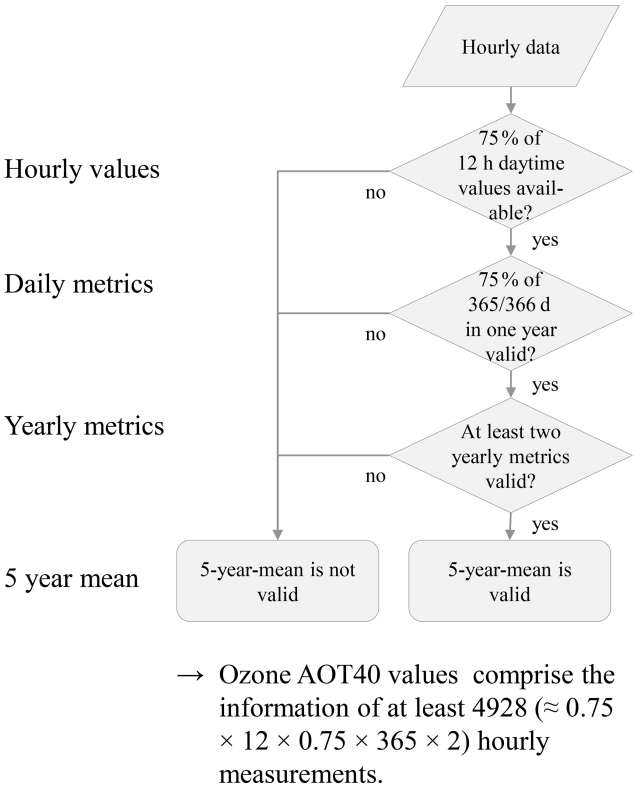


Figure B1. Data capture criteria for the AOT40 metric.

Appendix C: Data editing

Some data from TOAR–JOIN were modified in order to make them more understandable and user-friendly.

- *HTAP region* was updated according to the number code (see Table C1).
- Climatic zone was updated according to the number code (see Table C2).
- 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:
  - The types agricultural, commercial, other-agricultural and other-marine were replaced with other.
  - The type rural was replaced with background.
  - The type 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:
  - The types alpine grasslands, background, forest and marine were replaced with unknown.
  - The types rural-nearcity and rural-regional were replaced with rural.
  - The type rural-remote was replaced with remote.
  - The type Urban was replaced with urban.

  Five types of area remain: rural, urban, suburban, remote and unknown.
- The station with ID 4587 was eliminated because it was a remote background station in Romania which reported an *o3_average value* that was one of the highest of all stations (65.5899 ppb), and it had low data coverage. We suspect its values are faulty.
- The station with ID 4589 was eliminated because it reported a *max_population_density_5km* of ca. $1 \times 10^6\ \mathrm{km}^{-2}$ which we suspect is faulty.

Table C1. HTAP region number code.

No.	Replaced with	Description
2	OCN	Non-Arctic and Antarctic Ocean
3	NAM	USA and Canada (up to 66° N, polar circle)
4	EUR	Western and eastern Europe and 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	Southeast Asia
8	PAN	Pacific, Australia and New Zealand
9	NAF	Northern Africa, Sahara and Sahel
10	SAF	Sub-Saharan and sub-Sahel Africa
11	MDE	Middle East: Saudi Arabia, Oman, Iran, Iraq, etc.
12	MCA	Mexico, Central America, the 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) and Greenland
17	SPO	Antarctic

Table C2. Climatic zone number code.

No.	Replaced with
1	warm_moist
2	warm_dry
3	cool_moist
4	cool_dry
5	polar_moist
6	polar_dry
7	boreal_moist
8	boreal_dry
9	tropical_montane
10	tropical_wet
11	tropical_moist
12	tropical_dry

Appendix D: Hyperparameters for baselines

Table D1. Hyperparameters for the neural network training in Sect. 5.3. They are determined from empirical studies and random search.

Target	Learning rate	L2 lambda	Batch size	Epochs
Average values	1.0×10^{-4}	1.0×10^{-2}	32	250
Daytime average	1.0×10^{-4}	1.0×10^{-2}	32	250
Nighttime average	1.0×10^{-4}	1.0×10^{-2}	32	250
Median	1.0×10^{-4}	1.0×10^{-2}	32	250
25th percentile	1.0×10^{-3}	1.0×10^{-2}	64	100
75th percentile	1.0×10^{-3}	1.0×10^{-2}	256	250
90th percentile	1.0×10^{-3}	1.0×10^{-2}	256	250
98th percentile	1.0×10^{-3}	1.0×10^{-2}	256	250
dma8eu	1.0×10^{-3}	1.0×10^{-2}	128	250
avgdma8epax	1.0×10^{-4}	1.0×10^{-2}	32	250
drmdmax1h	2.0×10^{-4}	1.0×10^{-2}	32	150
W90	1.0×10^{-4}	1.0×10^{-2}	32	250
AOT40	1.0×10^{-2}	1.0×10^{-2}	128	250
nvgt070	1.0×10^{-4}	1.0×10^{-2}	32	150
nvgt100	1.0×10^{-5}	1.0×10^{-2}	32	200

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. Martin G. Schultz is a topic editor of the *ESSD* journal.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Special issue statement. This article is part of the special issue "Benchmark datasets and machine learning algorithms for Earth system science data (ESSD/GMD inter-journal SI)". It is not associated with a conference.

Acknowledgements. The authors gratefully acknowledge the computing resources granted by the Jülich Supercomputing Centre (JSC). The graphical abstract of this paper was designed with icons from Flaticon (https://www.flaticon.com/, last access: 21 June 2021). We gratefully acknowledge the comments and suggestions of the two anonymous reviewers and the topical editor.

Financial support. This research has been supported by the European Research Council, H2020 Research Infrastructures (IntelliAQ (grant no. ERC-2017-ADG#787576)) and the Helmholtz-Gemeinschaft (Supercomputing and Big Data of the Helmholtz Association's research field Key Technologies).

Review statement. This paper was edited by David Carlson and reviewed by two anonymous referees.

References

Amante, C. and Eakins, B. W.: ETOPO1 arc-minute global relief model: procedures, data sources and analysis, Tech. rep., NOAA National Geophysical Data Center, available at: https://repository.library.noaa.gov/view/noaa/1163/noaa_1163_DS1.pdf (last access: 21 June 2021), 2009.

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, arXiv [preprint], arXiv:1512.02595, pp. 173–182, 8 December 2015.

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, J. Geophys. Res.-Atmos., 101, 29239–29253, https://doi.org/10.1029/96JD00126, 1996.

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

Betancourt, C., Stadtler, S., and Stomberg, T.: AQ-Bench Git repository, GitLab – JSC [data set], available at: https://gitlab.version.fz-juelich.de/esde/machine-learning/aq-bench, last access: 21 June 2021.

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

Breiman, L.: Random forests, Mach. Learn., 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 PM_{10} Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model, Water Air Soil Poll., 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, available at: https://keras.io (last access: 21 June 2021), 2015.

CIESIN: Gridded Population of the World, Version 3 (GPWv3): Population Count Grid, Center for International Earth Science Information Network – CIESIN – Columbia University, United Nations Food and Agriculture Programme – FAO, and Centro Internacional de Agricultura Tropical – CIAT, Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), https://doi.org/10.7927/H4639MPP, 2005.

Comrie, A. C.: Comparing Neural Networks and Regression Models for Ozone Forecasting, J. Air Waste Manage., 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, Miami, FL, USA, 20–25 June 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848, 2009.

Duboue, P.: The Art of Feature Engineering: Essentials for Machine Learning, 1 edn., Cambridge University Press, Cambridge, UK, 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, 20–22 September 2017, 2017.

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, Adv. Environ. Res., 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, Environ. Pollut., 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, available at: http://data.europa.eu/eli/dir/2008/50/oj (last access: 21 June 2021), 2008.

Field, R., Goldstone, M., Lester, J., and Perry, R.: The sources and behaviour of tropospheric anthropogenic volatile hydrocarbons, Atmos. Environ. A-Gen., 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, 1 edn., MIT press Cambridge, Cambridge, UK, 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, Atmos. Environ., 34, 2131–2159, https://doi.org/10.1016/S1352-2310(99)00462-8, 2000.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M.: HTAP_v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmos. Chem. Phys., 15, 11411–11432, https://doi.org/10.5194/acp-15-11411-2015, 2015.

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), Atmos. Environ., 41, 9273–9287, https://doi.org/10.1016/j.atmosenv.2007.09.027, 2007.

Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D., and Tessum, C. W.: Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System, J. Geophys. Res.-Atmos., 125, e2020JD032759, https://doi.org/10.1029/2020JD032759, 2020.

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, Geosci. Model Dev., 14, 1–25, https://doi.org/10.5194/gmd-14-1-2021, 2021.

Koffi, B., Dentener, F., Janssens-Maenhout, G., Guizzardi, D., Crippa, M., Diehl, T., Galmarini, S., and Solazzo, E.: Hemispheric Transport Air Pollution (HTAP): Specification of the HTAP2 experiments – Ensuring harmonized modelling, Tech. rep., EUR 28255 EN, Luxembourg: Publications Office of the European Union, 2016.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States, https://doi.org/10.1145/3065386, pp. 1097–1105, 2012.

Krotkov, N. A., McLinden, C. A., Li, C., Lamsal, L. N., Celarier, E. A., Marchenko, S. V., Swartz, W. H., Bucsela, E. J., Joiner, J., Duncan, B. N., Boersma, K. F., Veefkind, J. P., Levelt, P. F., Fioletov, V. E., Dickerson, R. R., He, H., Lu, Z., and Streets, D. G.: Aura OMI observations of regional SO_2 and NO_2 pollution changes from 2005 to 2015, Atmos. Chem. Phys., 16, 4605–4629, https://doi.org/10.5194/acp-16-4605-2016, 2016.

LeCun, Y., Cortes, C., and Burges, C. J.: MNIST handwritten digit database, available at: http://yann.lecun.com/exdb/mnist/ (last access: 21 June 2021), 2010.

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, 27, 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, Atmos. Chem. Phys., 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, Glob. Change Biol., 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, Atmos. Chem. Phys., 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, J. Mach. Learn. Res., 12, 2825–2830, available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (last access: 21 June 2021), 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, Atmos. Chem. Phys., 15, 10349–10366, 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, J. Adv. Model. Earth Sy., 12, e2020MS002203, 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, 42200–42216, 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.

Schmitz, S., Towers, S., Villena, G., Caseiro, A., Wegener, R., Klemp, D., Langer, I., Meier, F., and von Schneidemesser, E.: Unraveling a black box: An open-source methodology for the field calibration of small air quality sensors, Atmos. Meas. Tech. Discuss. [preprint], https://doi.org/10.5194/amt-2020-489, in review, 2021.

Schraudner, M., Langebartels, C., and Sandermann, H.: Changes in the biochemical status of plant cells induced by the environmental pollutant ozone, Physiol. 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., Flocke F., Gregory, G. L., Heikes, B. G., Sachse, G. W., Sandholm, S. T., Shetter, R. E., Singh, H. B., and Talbot, R. W.: On the origin of tropospheric ozone and NO_x over the tropical South Pacific, J. Geophys. Res.-Atmos., 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, 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. H., Mozaffari A., and Stadtler S.: Can deep learning beat numerical weather prediction?, Philos. T. R. Soc. A., 379, 20200097, https://doi.org/10.1098/rsta.2020.0097, 2021.

Sillman, S.: The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments, Atmos. Environ., 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, Geophys. Res. Lett., 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., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484–489, 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., Pierce, T. E., Puxbaum, H., Shearer, M., Skiba, U., Steinbrecher, R., Tarrasón, L., and Öquist, M. G.: Inventorying emissions from nature in Europe, J. Geophys. Res.-Atmos., 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, Nutr. Cycl. Agroecosys., 48, 69–77, https://doi.org/10.1023/A:1009725319290, 1997.

Wagstaff, K.: Machine learning that matters, arXiv [preprint], arXiv:1206.4656, 18 June 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, 2006.

Wang, S., Ma, Y., Wang, Z., Wang, L., Chi, X., Ding, A., Yao, M., Li, Y., Li, Q., Wu, M., Zhang, L., Xiao, Y., and Zhang, Y.: Mobile monitoring of urban air quality at high spatial resolution by low-cost sensors: impacts of COVID-19 pandemic lockdown, Atmos. Chem. Phys., 21, 7199–7215, https://doi.org/10.5194/acp-21-7199-2021, 2021.

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

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., and Tian, D.: The Generic Mapping Tools Version 6, Geochem. Geophy. Geosy., 20, 5556–5564, https://doi.org/10.1029/2019GC008515, 2019.

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. C., 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, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

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

Xu, J., Ma, J. Z., Zhang, X. L., Xu, X. B., Xu, X. F., Lin, W. L., Wang, Y., Meng, W., and Ma, Z. Q.: Measurements of ozone and its precursors in Beijing during summertime: impact of urban plumes on ozone pollution in downwind rural areas, Atmos. Chem. Phys., 11, 12241–12252, https://doi.org/10.5194/acp-11-12241-2011, 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., 8, 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, Environ. Pollut., 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], arXiv:1707.08114, 25 July 2017.

D.2 Second paper (Stadtler et al., 2022)

Explainable machine learning reveals capabilities, redundancy, and limitations of a geospatial air quality benchmark dataset

Scarlet Stadtler, Clara Betancourt, and Ribana Roscher

Journal: Machine Learning and Knowledge Extraction (2022), Vol. 4, pp 150 – 171
Status: published (February 2022)
DOI: 10.3390/make4010008

machine learning & knowledge extraction

MDPI

Article

Explainable Machine Learning Reveals Capabilities, Redundancy, and Limitations of a Geospatial Air Quality Benchmark Dataset

Scarlet Stadtler [1,*], Clara Betancourt [1] and Ribana Roscher [2,3]

[1] Jülich Supercomputing Centre, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany; c.betancourt@fz-juelich.de
[2] Institute of Geodesy and Geoinformation, University of Bonn, Nußallee 17, 53115 Bonn, Germany; ribana.roscher@uni-bonn.de
[3] Data Science in Earth Observation, Technical University of Munich, Lise-Meitner Street 9, 85521 Ottobrunn, Germany
* Correspondence: s.stadtler@fz-juelich.de

Abstract: Air quality is relevant to society because it poses environmental risks to humans and nature. We use explainable machine learning in air quality research by analyzing model predictions in relation to the underlying training data. The data originate from worldwide ozone observations, paired with geospatial data. We use two different architectures: a neural network and a random forest trained on various geospatial data to predict multi-year averages of the air pollutant ozone. To understand how both models function, we explain how they represent the training data and derive their predictions. By focusing on inaccurate predictions and explaining why these predictions fail, we can (i) identify underrepresented samples, (ii) flag unexpected inaccurate predictions, and (iii) point to training samples irrelevant for predictions on the test set. Based on the underrepresented samples, we suggest where to build new measurement stations. We also show which training samples do not substantially contribute to the model performance. This study demonstrates the application of explainable machine learning beyond simply explaining the trained model.

Keywords: explainable machine learning; air quality; k-nearest neighbors; neural network; random forest

check for updates

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

Academic Editor: Andreas Holzinger

Received: 22 December 2021
Accepted: 26 January 2022
Published: 11 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Air pollution poses a significant environmental risk to human health, leading to 4.2 million premature deaths every year [1]. Therefore, air quality monitoring networks are established in many countries to warn the public, monitor compliance to regulations concerning air pollutant emissions, and analyze observations to assist with the development of new regulations [2,3]. Tropospheric ozone is a toxic air pollutant. In contrast to stratospheric ozone, which protects humans and plants from harmful ultraviolet radiation, tropospheric, near-surface ozone harms humans and plants. It is also a greenhouse gas [4]. Uncovering the spatial variability of air pollutants such as ozone is crucial for controlling air pollution and assessing human exposure.

Machine learning is a complementary approach to established physics-based chemistry-transport modeling [5–7]. Data-driven techniques and machine learning are increasingly explored for air quality modeling [8–12] because many observations are available on the one hand. On the other hand, these methods were proven to capture complex relationships while being easy to implement [12].

The downside of these easy-to-implement methods is the problem of opaque models. For atmospheric scientists, it is essential to understand the internal functioning of their models. Investigating machine learning approaches to predict ozone values based on environmental data can help pinpoint influential factors for ozone values or predict the

spatial variability of ozone. In addition, for decision-making, trustworthy and reliable models are required. Understanding the models' capabilities and limitations is a way to increase trust in a model. Explaining how a trained machine learning model arrives at its predictions gives us insights into its core functioning.

As stated in the beginning, air pollution monitoring is essential to design policies to protect the public and for research to understand air pollution chemistry. Inspired by the increasing application of data-driven techniques to air quality research, Betancourt et al. [11] combine environmental data with air quality observations for the challenge to model air pollutant tropospheric ozone. An impression of the AQ-Bench dataset is given in Figure 1. Figure 1a shows the locations of ozone observation stations distributed around the globe and their ozone values, while Figure 1b gives a histogram of the target data distribution. In AQ-Bench, the authors model the target ozone metrics derived from air quality measurement stations based on various geospatial datasets using different machine learning algorithms [11]. Betancourt et al. [11] show differing scores for the coefficient of determination for a random forest and a two-layer shallow neural network. They compare the coefficient of determination of the three data-driven approaches and found that the nonlinear methods had a higher score than linear regression. They conclude a similar performance of the shallow neural network and the random forest. What is rarely done, to our knowledge, is to explain the differences between various machine learning architectures applied to the same task.

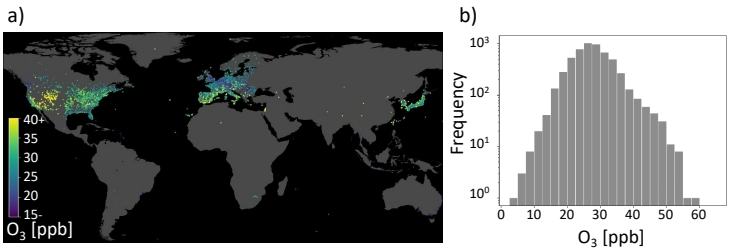


Figure 1. Geospatial air quality benchmark dataset (AQ-Bench). (**a**) Measurements of the target on a map projection. It is the average ozone from 2010 to 2014 of the AQ-Bench dataset and is given in ppb (parts per billion). (**b**) Histogram of the average ozone values in the AQ-Bench dataset [13].

In this study, we explain the similarity of a shallow neural network and a random forest, which are two different algorithms trained on the same dataset by showing similar behavior in the models' representation space. Thus, the contribution of this study is two-fold. On the one hand, we uncover the core functionality of two different machine learning approaches trained on the same benchmark dataset AQ-Bench. On the other hand, we use the models' explanations to gain a deeper understanding of the underlying dataset. The explanations reveal the representation of AQ-Bench in the machine learning models. With our analysis, we flag untrustworthy data samples, identify training data samples irrelevant for prediction, and recommend where to build new near-surface ozone measurement stations based on underrepresented test samples. The uniqueness of our approach is that we use machine learning explanations based on analysis of the models' representation space to derive understanding and make recommendations in the geographical space.

2. Related Work

Earth system science research faces challenges when applying machine learning methods to environmental data. Tuia et al. [14] point out the challenges that arise from the basic machine learning chain to derive input-output relations from Earth system data. The input data are complex and, at the same time, limited. The black box behavior of the models has to be overcome, and the output results should be turned to an explainable, reliable, and scientifically consistent outcome. A full review is beyond the scope of this study, but in the following sections, we (i) emphasize data-driven ways to model air pollution (Section 2.1); (ii) show examples of overcoming the black box behavior in atmospheric

science (Section 2.2); and (iii) highlight studies turning their results into scientific outcomes (Section 2.3). We focus on studies within the Earth system science domain to meet the goal of this study, which aims to use machine learning explanations for further use in ozone research.

2.1. Data-Driven Air Pollution Modeling

Different machine learning approaches have been used in air pollution research in recent years. Algorithms capable of learning nonlinear relationships in the input data are needed to process complex air pollution data. Tree-based algorithms, such as random forest and sophisticated neural networks, are commonly applied to model different air pollutants such as fine particulate matter (PM2.5, PM10) and gases such as near-surface ozone.

Brokamp et al. [15] use a random forest for land-use regression and assessment of several particulate pollutant species. They conclude to use random forests in land-use models for more accurate exposure assessment in the future. Similarly, Mallet [16] states that the best performing model in their study is the random forest, which can model 59% of the variance in PM10. They use a range of meteorological, environmental, and temporal variables as predictors. Althuwaynee et al. [17] judge the random forest to provide clear insights about the PM10 pollution distribution. They implement a random forest and extreme-gradient boosting to map the PM10 susceptibility index onto probability and classification index maps. Tian et al. [18] find that their random forest outperforms other models, suggesting that the relationship between air quality and spatial configurations of the urban elements such as the urban infrastructure is most likely nonlinear. They use a random forest and a neural network to combine meteorological factors with urban elements to explore intra-urban PM2.5 concentrations. Lu et al. [19] conclude that deviations of hourly ozone prediction by their numerical chemistry transport model can be significantly reduced by machine learning postprocessing. Their postprocessing involves Lasso regression, random forest, and a long short-term memory recurrent neural network. Alimissis et al. [20] compare the application of neural networks and multiple linear regression to spatial interpolation of the urban air pollutants nitrogen oxides, ozone, carbon monoxide, and sulfur dioxide. They conclude that neural networks are significantly superior in most cases. Cabaneros et al. [8] review 139 papers using neural networks for air pollution modeling between January 2001 and February 2019. Wen et al. [21] propose using convolutional long short-term memory to predict PM2.5. Their results show that their machine learning model achieves a better performance than current state-of-the-art models for monitoring stations in China. Based on meteorological and air quality data, convolutional neural networks are applied to forecast ozone at several hundred measurement locations [9,22].

2.2. Explainable Machine Learning in Earth Science

Mcgovern et al. [23] state that the ultimate goal of Earth scientists is to deepen their understanding of the Earth system. Therefore, incorporating machine learning into a cycle of knowledge discovery is a means to get closer to this goal. To integrate machine learning into the cycle of knowledge discovery, explainable AI and interpretation techniques are required to understand the core functioning of the machine learning models. Review articles list explainable AI methods [24,25]; here, we highlight Earth science studies that explain their machine learning models.

Mcgovern et al. [23] use interpretation techniques such as saliency maps [26], backward optimization [27], and neuron ranking by their discrimination ability to examine their tornado predictions. Gu et al. [12] note that different models favor different predictor variables and result in different interpretation abilities by interpreting their data-driven air quality models using SHAP value-based explanations. Yan et al. [28] develop an interpretable deep learning model to retrieve surface fine particle air pollution from satellite data. They can extract spatio-temporal features from their model, which agrees with their physics-based numerical model. Bennett et al. [29] analyze their neural network for simulating latent and sensible heat fluxes using layer-wise relevance propagation [30], and

they show that even simple neural networks can extract physically plausible relationships. They suggest that explainable AI methods offer ways to learn from trained neural networks instead of just making predictions. However, to reach the ultimate goal, as stated by Mcgovern et al. [23], explaining a model is not the end, and scientific insights need to be generated from the results.

2.3. Scientific Insights through Explainable AI

In their abstract, Roscher et al. [31] write: "An exciting and relatively recent development is the uptake of machine learning in the natural sciences, where the major goal is to obtain novel scientific insights and discoveries from observational or simulated data". According to Roscher et al. [31], an essential component is domain knowledge, which is needed to increase models' and results' explainability and enhance scientific consistency. Their article reviews various explainable machine learning approaches and highlights how they are used in combination with domain knowledge from different disciplines. Some studies go a step further than simply determining the explanation of used machine learning models; they leverage the achieved explanations to gain a deeper understanding of the Earth system. Stirnberg et al. [10], for example, use explanations based on SHAP values [32] to reveal meteorological factors driving fine particulate air pollution variability. With their SHAP value analysis, they gain process understanding at individual air pollution measurement sites. Toms et al. [33] apply an explainable neural network as a tool to identify patterns of Earth system predictability. Their neural network is trained to predict decadal oceanic variability and explained it by applying layer-wise relevance propagation [30]. They conclude that explainable neural networks are useful in determining patterns of predictability. Schramowski et al. [34] introduce a method called explanatory interactive learning for deep convolutional neural networks with the task of plant phenotyping. They use explanations by saliency maps to uncover correctly classified samples affected by the Clever-Hans effect [35] and correct these predictions to arrive at an explainable and trustworthy model. In perspective, Tuia et al. [14] argue that learning causal relationships is crucial for understanding the Earth system. The link between explainability and actual causal relationships is strong since relationships determined by machine learning can be paired with domain knowledge to formulate hypotheses that can help to uncover novel cause-and-effect relationships.

3. AQ-Bench Dataset

AQ-Bench is a machine learning benchmark dataset designed to empirically relate ozone statistics observed at air quality measurement stations to geospatial data. It contains aggregated ozone statistics from over 5500 measurement stations of the years 2010–2014. These stations are distributed globally, although not evenly (see Figure 1). The primary source of the ozone statistics is the Tropospheric Ozone Assessment report database [3]. Most of the stations are in Europe, North America, and East Asia. Although AQ-Bench contains different ozone statistics, this study only focuses on the average ozone as a target variable.

The geospatial features in AQ-Bench characterize the measurement site. Although there are no functional relationships available as prior knowledge for machine learning in the dataset, these geospatial features were selected because they serve as proxies for ozone formation, destruction, and transport processes. Features such as 'population density' in different radii around the station indicate human activity and, therefore, ozone precursor emissions. In addition, features such as 'altitude'/'relative altitude' are used as proxies for local flow patterns and ozone sinks. A complete description of the features in AQ-Bench and their relation to ozone processes can be found in [11].

4. Methods

We combine the following methods to gain novel scientific insights about the AQ-Bench dataset. First, we use the methods to understand how the trained models work.

Second, we use our knowledge about the models' functioning to explain inaccurate predictions. We train our models on a dataset that consists of input feature vectors $\mathbf{x}_i$ and target values y_i. Both machine learning models predict $\hat{y}_i$ based on the input feature vector:

$$\hat{y}_i = f_{\text{model}}(\mathbf{x}_i)\,. \tag{1}$$

To gain novel insights, we uncover the models' functioning by calculating SHAP global importance for both models; see Section 4.1 and visualizing prediction patterns. Since we use a random forest and a neural network, we implement visualization methods tailored to the specific architectures. Section 4.2 presents the neural network visualization method and Section 4.3 presents the random forest visualization method. These visualizations help us to explain individual predictions. Nevertheless, interpreting individual predictions does not yield a global understanding of the trained models. Therefore, we move from single predictions to studying prediction patterns. For this, we use k-nearest neighbors on both models for explaining inaccurate predictions; see Section 4.4.

4.1. SHAP

As Lundberg et al. [32] proposed, we use SHapley Additive exPlanations (SHAP) to explain local and global predictions [36]. SHAP values are derived by a model-agnostic post hoc explainable machine learning method and therefore are suitable for comparison of our two different machine learning algorithms. The SHAP values quantify the contribution of each feature to the model prediction. Contribution refers to the deviation from the base rate, which is the expected value of the training dataset, where features with high absolute contributions are considered more important. For example, a feature with a negative SHAP value causes the model to predict a value lower than the expected value of the training set. Since features with large SHAP absolute values are considered important for a single prediction, averaging absolute SHAP values per feature across data results in an estimate for global importance based on SHAP.

4.2. Neural Network Activation

For the neural network, Equation (1) takes the form:

$$\hat{y} = f_{\text{nn}}(\mathbf{x}, \mathbf{W}, \mathbf{b}) \tag{2}$$

where $\mathbf{W}$ and $\mathbf{b}$ represent the neural network's parameters [37]. Our trained, shallow neural network can be easily visualized by representing the node structure and expressing the values of weights and biases as colors (Figure 2, left). During inference, the trained neural networks parameters $\mathbf{W}, \mathbf{b}$ are combined with the input feature vector $\mathbf{x}$ and the activation function σ in each layer:

$$\mathbf{A}^{[1]} = \sigma(\mathbf{W}^{[1]\intercal}\mathbf{x} + \mathbf{b}^{[1]}) \tag{3}$$

$$\mathbf{A}^{[l]} = \sigma(\mathbf{W}^{[l]\intercal}\mathbf{A}^{[l-1]} + \mathbf{b}^{[l]}) \tag{4}$$

where $\mathbf{W}^{[l]}$ and $b^{[l]}$ the weights and biases of layer l [37]. Therefore, we can also visualize the trained neural network during inference by plotting the activation $\mathbf{A}$. The neural network signals are obtained by visualizing Equations (3) and (4); see Figure 2 (right).

4.3. Random Forest Activation

A random forest consists of decision trees $h(\mathbf{x}, \theta_k)$, where θ_k are independent and identically distributed random vectors. The random forest prediction is the average over all K decision tree predictions. Thus, Equation (1) takes the form:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} h(\mathbf{x}, \theta_k), \tag{5}$$

given the input **x** [38]. Typically, a random forest consists of hundreds of decision trees [39]. Therefore, visualization of the individual decision trees is possible, but hardly useful due to their sheer number and complexity.

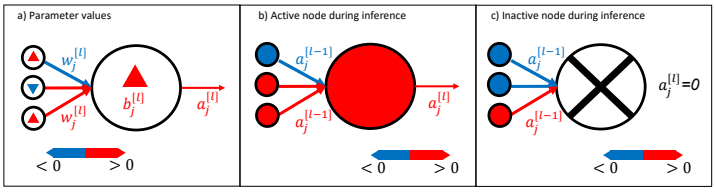


Figure 2. Visualization method for neural networks. It is possible to visualize the trained neural network weights **W** and biases **b**, as shown in (**a**). During inference, active neurons (activation **A**) transport a signal, as shown in (**b**), whereas it is also possible to have inactive neurons that do not transport any signal; see (**c**). Note that we use lowercase letters to indicate the components of the vectors and matrices.

Since we can represent our data in the geographical space, we use a more intuitive way of visualizing the basis set of influential training samples that the random forest used for its prediction. By visualizing the location of the basis set used for prediction on a global map, we display the random forests' functioning. We name this type of visualization *leaf activation* to emphasize the similarity to an activated neural network during prediction. The steps to create this kind of visualization are illustrated in Figure 3 and listed in the following:

1. Propagate all training samples through the trained random forest. Keep track of the tree IDs, leaf node IDs, and corresponding training sample IDs.
2. Propagate a single test sample through the random forest. Track the corresponding responsible tree IDs and leaf node IDs for the prediction.
3. To identify training samples that are most relevant for a given prediction, keep track of the relative frequency of the training samples contributing to the leaf node predictions responsible for a given test sample prediction.
4. Since each training sample has geographical information; influential training samples can be visualized on a map. The marker size indicates the frequency of a specific training sample contributing to the leaf nodes responsible for a particular prediction.

As decision trees split the data according to their features, these groups of training samples should have similar features as the target test sample. These training samples took the same decision path through the decision trees and ended up in the same leaf node as the test sample.

4.4. Explaining Inaccurate Predictions with k-Nearest Neighbors

Figure 4 shows how to use k-nearest neighbors to explain inaccurate model predictions as proposed by Bilgin and Gunestas [40], who explain their deep learning models through post hoc analysis of k-nearest neighbors. For an inaccurately predicted test sample, they extract the k-nearest neighbors in the training dataset and feed them into the trained model. By comparing the prediction based on the nearest neighbors in the training set and the inaccurate prediction of the test sample, they derive an interpretation of the model's response and identify different cases. Bilgin and Gunestas [40] apply their method to two standard machine learning benchmark datasets: IRIS and CIFAR10. They originally tested their method on supervised classification tasks, and we adapted and applied it to our supervised regression task.

Since our goal is to explain the functioning of our two machine learning models, we search the k-nearest neighbors in their respective representation spaces. For the random forest, we defined the nearest neighbors as samples in the same leaf nodes (Section 4.3). For the neural network, we defined the nearest neighbors as samples leading to similar

activation patterns (Section 4.2), i.e., a group of neurons activated. To search the neural network activation pattern space, we use the Euclidean distance

$$L(a_1, a_2) = \sqrt{\sum_{i=1}^{n}(a_{1,i} - a_{2,i})} \tag{6}$$

where a_1 and a_2 are a pair of neighboring activation patterns in the n-dimensional neural network activation space.

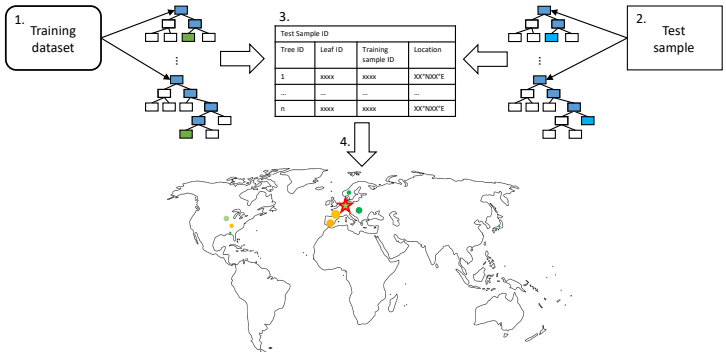


Figure 3. *Leaf activation visualization* pipeline as described in the text as enumerated bullet points. We match the training samples contributing to a specific test sample prediction and determine the influence each training sample has on the prediction. The basis set of training samples, the relative influence of a training sample, and the target test sample are visualized as a scatter plot on the map.

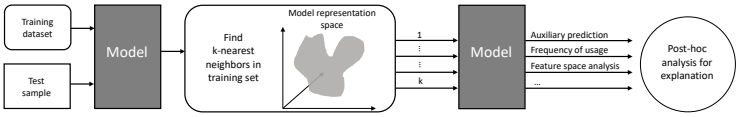


Figure 4. Explaining inaccurate predictions by k-nearest neighbors pipeline (adapted and extended from [40]). After identifying the k-nearest neighbors in the model representation space, we analyze the auxiliary predictions based on the nearest neighbors regarding the accuracy, relevance, and distance to the test sample in the feature space.

We define the following prediction scenarios for inaccurate predictions; see Figure 5:

- Case-I-A: A sample of k-nearest neighbors leads to consistent, inaccurate predictions. The k-nearest neighbors and the test station are located next to each other in the feature space. The inaccurate prediction of the test sample is not unexpected. In this case, the model might not be fitted well.
- Case-I-B: A sample of k-nearest neighbors leads to consistent, inaccurate predictions. The k-nearest neighbors and the test station are not located next to each other in the feature space. The inaccurate prediction of the test sample is not unexpected. In this case, the model might not be fitted well, and the test sample is not well represented–too many problems.
- Case-II-A: The model accurately predicts a sample of k-nearest neighbors, while it inaccurately predicts the test sample. The k-nearest neighbors and the test station are located next to each other in the feature space. Therefore, the inaccurate prediction of the test sample is unexpected. This could point to either an erroneous test sample or a model limitation. In any case, this prediction is untrustworthy.
- Case-II-B: A sample of k-nearest neighbors leads to accurate prediction, while the test sample is inaccurately predicted. The k-nearest neighbors and the test station are not located next to each other in the feature space. Thus, the inaccurate prediction of the test sample is not unexpected. This points to an underrepresented test sample.

- Case-III-A: A sample of k-nearest neighbors leads to scattered accurate predictions. The test sample is accurately predicted. In the feature space, the accurately predicted test sample has nearest neighbors. The models are predicting a correct value. This is the usual case for a healthy prediction.
- Case-III-B: A sample of k-nearest neighbors leads to scattered predictions; both accurate and inaccurate predictions are possible. The test sample is accurately predicted but due to the wrong reasons. The accurately predicted test sample has no nearest neighbors in the feature space. The models are predicting a correct value but due to the wrong reason. We can flag this case as the Clever-Hans effect [35].

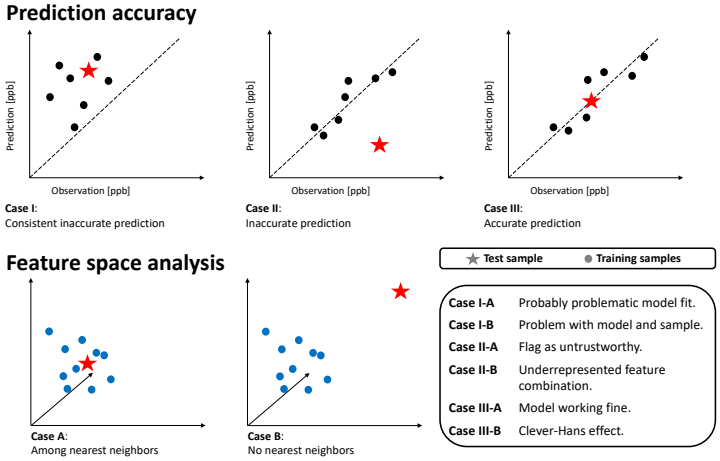


Figure 5. Schematic overview of possible cases. The upper columns show the predictions on k-nearest neighbors training samples and the target test sample with respect to the prediction error. The lower columns depict the k-nearest neighbor analysis in the feature space. Dots represent training samples, while stars depict the test sample.

For the search of the k-nearest neighbors, we prepared the feature space by (i) using scaled features such that all features have a comparable range of values and (ii) weighting the features with the respective SHAP importance value. The weighting of the feature space follows the method by Meyer and Pebesma [41], which calculates the distances' multidimensional feature space, with features being weighted by their respective importance in the model. Then, a Euclidean distance in this scaled and weighted feature represents the distance relevant to the model prediction.

5. Experimental Setup

This Section gives an overview of the experimental setups of model training and the application of explainable machine learning methods to our models. We describe the model training in Section 5.1 and the evaluation in Section 5.2. We compare the feature importance of both models with SHAP, as described in Section 5.3. To gain an insight into the representation of AQ-Bench in the trained machine learning models, we visualize single predictions, as described in Section 5.4. By investigating the predictions made on the test set in relation to the training samples that this prediction is based upon, we gain an understanding of prediction accuracy. We present in Section 5.5 how we use k-nearest neighbors for explaining inaccurate predictions.

5.1. Model Training

We train a shallow neural network and a random forest to solve the task posed by Betancourt et al. [11]: given geospatial data describing the environmental features, infer the ozone metrics. In this study, we focus on predicting one ozone metric, the average ozone. We want to solve the task of predicting average ozone values by training two machine learning models on a subset of AQ-Bench features. AQ-Bench originally contains over 100 features. Following the feature selection method by Meyer et al., here, we only

use 31 of them (features listed in Appendix A, Table A1), because fewer features decrease model complexity and enable more comprehensible explanations. Ref. [13] showed that forward feature selection applied on AQ-Bench leads to 31 features. The data split is kept as in AQ-Bench with 60% training (approximately 3300 samples) and 20% validation and test samples (roughly 1110 samples, respectively).

We trained a two-layer shallow neural network and a random forest to predict the average ozone value based on this subset of geospatial data. The hyperparameters of both machine learning models are summarized in Table A2 in Appendix B.

5.2. Evaluation Metrics

To evaluate the performance of our models, we use common evaluation metrics in the field of machine learning. We calculate the Root Mean Square Error (RMSE) and the coefficient of determination (R^2) based on the following formulas:

$$R^2 = 1 - \frac{\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}{\sum_{m=1}^{M}(y_m - \langle y \rangle)^2} \quad \text{with} \quad \langle y \rangle = \frac{1}{M}\sum_{m=1}^{M} y_m \tag{7}$$

$$\text{RMSE} = \sqrt{\sum_{m=1}^{M} \frac{(y_m - \hat{y}_m)^2}{M}}. \tag{8}$$

Moreover, we consider deviations between the prediction and the reference value as residuals. Residual Δ is calculated by subtracting the prediction $\hat{y}$ from the observed ozone value y:

$$\Delta = y - \hat{y}. \tag{9}$$

Therefore, negative residuals point to an overestimation by the prediction, while positive residuals depict underestimation.

5.3. SHAP Values

We aim to compare machine learning models based on different algorithms. Gu et al. [12] propose to treat SHAP (Section 4.1) as a unifying framework for the comparison of different machine learning models. Thus, we use SHAP feature importance to rank features of both trained random forest and neural network according to their relevance. SHAP values for the random forest are are calculated analytically, whereas the SHAP values for the neural network area are approximations. Details about the calculation of the SHAP values and the software we used can be found in [32].

We expect that both models use similar features to predict average ozone, i.e., a subset of features that are among the most important for both models.

5.4. Visualization of Individual Predictions

By visualizing the predictions patterns of an accurate prediction and an inaccurate prediction, we aim to show that the underlying patterns leading to an accurate prediction can be differentiated from the patterns leading to an inaccurate prediction. Here, we choose two example test samples for visualization where the models had to predict high ozone values. The one example shows accurate predictions by both models, while the second example displays an inaccurate prediction with a positive residual, which is also called underestimation by the models. We chose test samples to be geographically close to each other; both are located in southern Europe. An overview of the selected test sample stations, observed average ozone value, predicted values, and residuals is given in Table 1.

Table 1. Example stations with the station IDs to identify them in the Tropospheric Ozone Assessment Report database. Station 6952 is located in Spain, while Station 8756 is situated in Greece. The subscripts point to the corresponding model abbreviation, nn for the neural network and rf for the random forest.

Station ID	Observation y	$\hat{y}_{nn}$	Δ_{nn}	$\hat{y}_{rf}$	Δ_{rf}
6952	42.44	42.00	0.43	41.99	0.44
8756	40.17	29.32	10.86	27.45	12.72

5.5. Identify k-Nearest Neighbors and Classify Predictions

We aim to test our hypothesis that certain feature combinations lead to activation patterns in both models related to prediction accuracy. Moreover, we increase our understanding of how the models function and identify different reasons for inaccurate predictions. To do so, we use auxiliary predictions on the k-nearest neighbors, as described in Section 4.4. We identify the k-nearest neighbors for the auxiliary predictions in the models' representations spaces and compare if these k-nearest neighbors are also the test sample's k-nearest neighbors in the feature space. To automatically classify our test samples to the different cases (Figure 5), we determine 11 nearest neighbors in the training set of a given test sample. Then, we calculate the average residual of the training samples and compare it to the test sample's residual. In addition, we calculate the average distance between the group of k-nearest neighbor training samples in the feature space and compare it to the average distance between the test sample and its k-nearest neighbors. Based on these values, we can classify our samples into different cases.

We expect both models to lead to similar classifications of the test stations to the cases.

5.6. Train on a Reduced Dataset

We hypothesize that removing non-influential training samples will not affect the performance of machine learning models. To test the hypothesis, we re-train our models on a reduced dataset. We identify the 10% training samples that are not influential for the predictions on the test samples. To identify which samples are non-influential, we used the identified 100 nearest neighbors for each test sample and ranked the whole training dataset according to the proximity to the test samples. We eliminated the 10% of data with the lowest proximity to the test samples in the models' representation spaces from the training dataset. This leads to a training dataset of the size 3000 training samples. For evaluation, we use the evaluation metrics introduced in Section 5.2. The hyperparameters of both models are kept unchanged.

We do not expect significant performance losses of both models. The random forest is less sensitive to changes in the training dataset than the neural network, such that we expect a slightly higher performance loss of the neural network than the random forest.

6. Results

6.1. SHAP Global Importance

Table 2 gives an overview of the global feature importance for the trained random forest and neural network. In both models, the absolute value of the latitude is the most influential feature, with global feature importance of 23.96% (RF)/20.50% (NN). For the subsequential most important features, the models differ. The trained random forest heavily relies on features related to topography, i.e., altitude, and then uses environmental characteristics connected to an anthropologically influenced environment. The topography-related features are also of relevance to the neural network. Both models attribute some importance to the forest in the surrounding 25 km area, while for the neural network, this feature is two times as important as it is for the random forest. There are several features with low importance attributed by both models. These are mainly tropical, boreal, and polar climatic zones, which are not well represented in the AQ-Bench dataset. The differing

feature importance of both models leads to differently weighted features spaces when searching the k-nearest neighbors.

Table 2. Global feature importance derived by SHAP for our trained random forest (RF) and neural network (NN). The first column lists short descriptions of the AQ-Bench features; we kept the order from the most important to the least important in the test set of our random forest. Percentage values for the random forest are shown in the second column, and corresponding values of the neural network are shown in the third column. The largest importance values of both models were underlined, the second largest are shown in bold, and the third largest are shown in italic font. For a table with the AQ-Bench feature names, see Appendix A Table A1.

Feature Description	Importance RF [%]	Importance NN [%]
Absolute latitude	23.96	20.5
Relative altitude	**16.21**	*11.93*
Altitude	*10.44*	8.16
Nightlight in 5 km area	9.73	4.35
Forest in 25 km area	5.37	**13.54**
Population density	4.41	1.8
Nightlight in 1 km area	4.12	8.77
Water in 25 km area	4.11	7.5
Maximum population density in 25 km area	3.54	0.32
NO_2 emissions	3.51	6.31
Maximum population density in 5 km area	3.04	1.21
Savannas in 25 km area	2.68	0.84
Croplands in 25 km area	1.73	1.74
Grasslands in 25 km area	1.68	0.77
NO_x emissions	1.62	0.84
Warm, dry climate	1.18	5.27
Shrublands in 25 km area	0.65	1.67
Maximum nightlight in 25 km area	0.35	0.39
Warm, moist climate	0.33	2.49
Cool, moist climate	0.32	0.14
Rice production	0.3	0.58
Permanent wetlands in 25 km area	0.27	0.15
Cool, dry climate	0.25	0.13
Tropical, dry climate	0.14	0.13
Tropical, wet climate	0.03	0.11
Tropical, moist climate	0.02	0.09
Boreal, moist climate	0.0	0.11
Polar, moist climate	0.0	0.07
Boreal, dry climate	0.0	0.1
Polar, dry climate	0.0	0.0
Tropical, montane climate	0.0	0.0

6.2. Comparison of Neural Network and Random Forest Performance and Residuals

The coefficients of determination for the neural network and random forest for the training set, validation set, and test set can be found in Table 3. We calculated all performance metrics using the observed values (ground truth) and Equations (7) and (8). The coefficient of determination R^2 is over 95% for the training set for the random forest, while it is 64.21% for the neural network. The difference between the R^2 and RMSE on the test set is smaller than the difference on the training set. The random forest has a slightly higher R^2 with 53.03% than to the neural network with 49.46%. The difference between the RMSE of both models is smaller, with 4.46 ppb for the random forest and 4.59 ppb for the neural network. From these two scores, the models perform comparably well on the test set, while it is apparent that the random forest performance is slightly better.

Table 3. Coefficient of determination and RMSE for the training and validation test set.

Random Forest	R^2 [%]	**RMSE [ppb]**
Training	95.75	1.33
Validation	56.99	4.08
Test	53.03	4.46
Neural Network		
Training	64.21	3.52
Validation	58.34	3.87
Test	49.46	4.59

The focus of this study lies on the test set; therefore, we take a closer look at the residuals of both models. The residual is defined in Section 5.2, Equation (9). Figure 6 shows the residuals of the random forest together with the residuals of the neural network. Both models mainly predict the average ozone on the test set with a residual error below 5 ppb. We consider predictions with residuals below 5 ppb as accurate, considering the conservative measurement error estimation of 5 ppb [3]. From 1110 test samples, the random forest accurately predicted 867 samples, and the neural network accurately predicted 842 samples. The correlation between the residuals of the random forest and the neural network is shown in Figure 6. The correlation is high, so apparently, some test samples are difficult to predict for both models. The following Sections focus on these 268 (neural network) and 243 (random forest) inaccurately predicted samples.

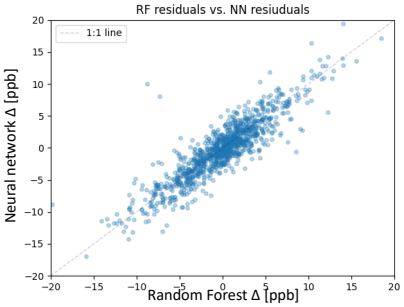


Figure 6. Scatter plot of the residuals Δ of the random forest on the x-axis and neural network on the y-axis for the test set.

6.3. Visualization of Individual Predictions

In this section, we take a closer look at the visualization of single prediction patterns by both machine learning models. As described in Section 4.3, it is possible to show in geographical space upon which samples the random forest bases its prediction. Table 2 gives the coordinates, and the stations are displayed in Figure 7. Both test samples are located in the Mediterranean area, in Spain (a,b) and Greece (c,d). Figure 3a,b show the accurate prediction, where the random forest bases its prediction on stations with similar features and similar average ozone values. The most influential training samples are located next to the target test station in the geographical space. The predicted value is 41.99 ppb compared to the observed value of 42.44 ppb. In contrast, Figure 7c,d shows an inaccurate prediction with a residual of 12.72 ppb, while the target average ozone was 40.17 ppb, which is indicated by the bright yellow star in Figure 7d. In this case, the influential training samples have much lower average ozone values. Moreover, there are hardly any influential stations with larger contribution but many with small markers, even in South Korea and Japan. The accurate prediction for station 6952 is based upon 63 training stations, while the inaccurate prediction of station 8756 is based upon 122 training stations.

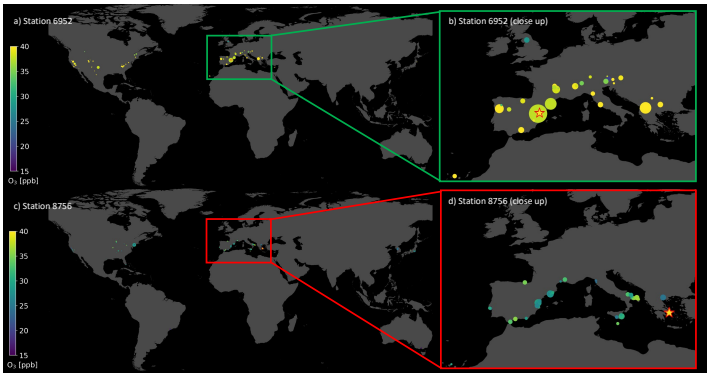


Figure 7. Map with training stations (dots) upon which the trained random forest bases its predictions to predict the test station (star). The colors of the dots indicate the observed ozone value at the respective training and test station. Plot (**a**) and (**b**) depict an accurate prediction with a small residual. Plot (**c**) and (**d**) show an inaccurate prediction with a large residual. Plots (**b**) and (**d**) are close-ups over Europe.

In Section 4.2, we introduced a way of visualizing shallow neural networks. We visualize the neural network while it infers the average ozone values for the same two example test stations presented in Table 2. Figure 8a shows the activation pattern caused by the accurate prediction for the station in Spain, while b shows the activation pattern for the inaccurately predicted station in Greece. The residual error for the accurate prediction is comparable to that of the random forest, 0.43ppb, while the inaccurate prediction misses the target value by 10.56 ppb. As a reference, we also show the weights and biases of the network in Appendix C, Figure A1. Left of the input nodes, we noted down the feature names, also found in Appendix A, Table A1. Red nodes indicate active input features important for the respective prediction. In both cases (a) and (b), most input nodes are light blue, indicating slightly negative values. While in (a), there are many connections activating and deactivating nodes in the first hidden layer, the inaccurate prediction (b) has a first hidden layer with mainly slightly activated nodes. The signals to the second hidden layer are visible in (a), while again, in (b), there is hardly any departure from the mean state. In the second hidden layer, the first node from the top can reduce the value of the output node, while the other four nodes increase the output (see also Appendix C, Figure A1). In (a), this node is correctly deactivated, leading to a high and accurate prediction. In (b), this node is activated such that the second hidden layer increases and decreases the output value, leading to a prediction near the average.

6.4. Explaining Inaccurate Predictions

To get a general impression of how our trained models work, we look at the 11 nearest neighbors of the entire test set. As described in Section 4.4, the whole test set is needed to classify all test samples into the three cases using the k-nearest neighbor algorithm to identify the nearest neighbors. After classifying all test samples, we mainly focus on inaccurate predictions with a residual larger than 5 ppb, which can only appear in case-I and case-II. Figure 9 show inaccurately predicted test stations ordered according to their predicted average ozone value. Each vertical sequence of dots represents one test sample. Each dot is one nearest neighbor of the test sample. The nearest neighbors were identified in the respective models' representation. Afterward, we also checked the Euclidean distance of these nearest neighbors in the weighted feature space. In the vertical, nearest neighbors are ordered according to the weighted Euclidean distance in the feature space, meaning the further away from zero the dot is placed, the more different the features of the neighbor and the target test sample. The colors represent negative (red) and positive residuals (blue). The dot size indicates proximity/importance in the models' representation spaces. Figure 9a shows the random forest results, and Figure 9b shows the neural network results.

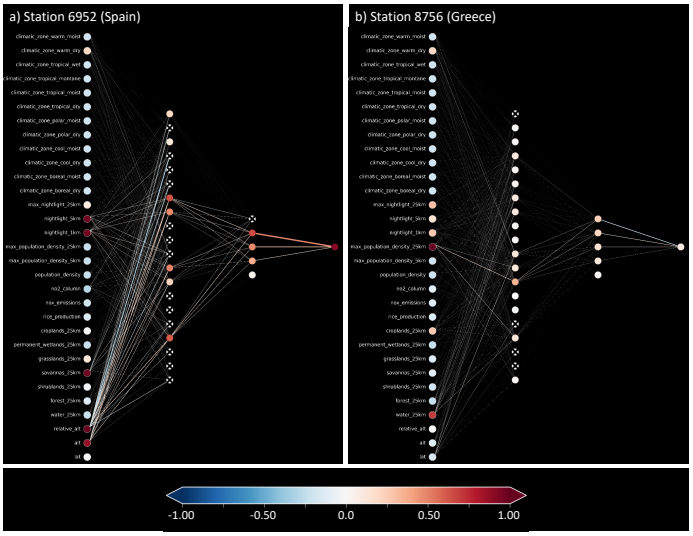


Figure 8. Trained shallow neural network while making predictions for two test samples. Plot (**a**) depicts an accurate prediction with a small residual. Plot (**b**) shows an inaccurate prediction with a large residual.

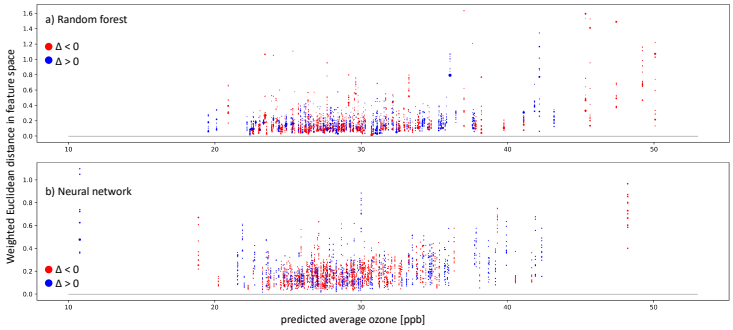


Figure 9. Plots showing inaccurate predictions (absolute values of $\Delta > 5$ ppb) on the test set by the random forest (**a**) and neural network (**b**). The weighted Euclidean distance in the respective feature space is on the y-axis. On the x-axis, the predicted average ozone is shown. The colors indicate the residual: red points mean negative residuals and model overestimation, while blue points show positive residuals and model underestimation. The size of the single dots points to the distance between the test sample and the training sample in the representation space. A larger dot means that the test sample and respective training sample lead to similar activation patterns in the model.

Inaccurate predictions can be found in both models over the average ozone distribution. We have far more test samples between 25 and 30 ppb than outside of this range. Below 21 ppb and above 35 ppb, both models have trouble finding nearest neighbors in the weighted feature space. This is visible through the vertical sequence of dots being farther away from the zero line in black. Moreover, in the well-represented range between 25 and 30 ppb, some samples have nearest neighbors in the weighted feature space, but still, they fail to produce accurate predictions. We further analyze the nature of the failed prediction using the cases presented in Section 4.4 in Figure 10.

Figure 10 shows all inaccurately predicted test samples and accurately predicted samples that do not have nearest neighbors in the weighted feature space. The colors denote the case describing the reason for the inaccurate prediction. Blue dots represent case II-A (untrustworthy sample and/or prediction), plum-colored dots represents case II-B (underrepresented feature combination), and gray dots refer to case III-B (Clever-Hans effect). Figure 10a shows the analysis for the random forest, and Figure 10b shows the analysis for the shallow neural network. The prediction error on the test samples is

mainly close to ≈0 ppb, never leading to case-I for both models. Thus, all inaccurate predictions belong to case-II for AQ-Bench. The Random Forest incorrectly predicted 243 test samples, of which 238 belong to case II-A, while five belong to case II-B. There are a total of 268 inaccurate predictions for the neural network: 246 categorized as case II-A and 22 categorized as case II-B. We also checked the nearest neighbors in the weighted feature space for all accurately predicted test samples and found six samples belonging to case-III-B for the random forest and 52 for the neural network.

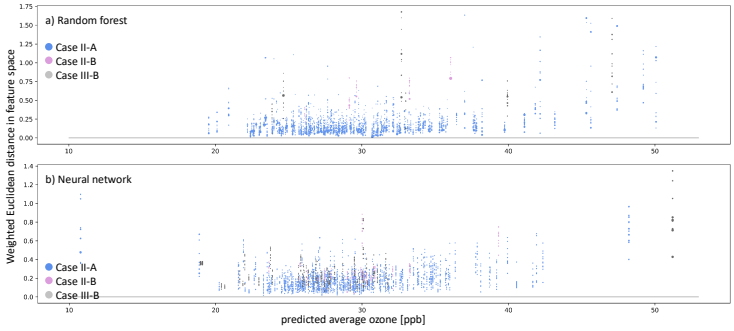


Figure 10. Similar to Figure 9, but the colors indicate the type of inaccurate prediction (navy blue and plum dots). Moreover, additionally to inaccurate predictions, we also show instances of Clever-Hans, where the residual error $\Delta < 5$ ppb, but there are no nearest neighbors in the models' representation space (gray dots).

We can pick the inaccurately predicted samples and plot their geospatial locations based on this classification. Figure 11 illustrates untrustworthy predictions, underrepresented test stations, and training stations that were non-influential (Section 6.5) for all predictions on the test set. We derive areas where new data acquisition would improve our data-driven models for the underrepresented samples. Given our current test station features, we searched the global feature space to identify locations where we recommend additional observations. Figure 11 shows the areas in blue, red, and violet (overlap blue and red) on the globe where we recommend building new ozone monitoring stations. For example, both models recommend building stations around the underrepresented stations in Greenland and Chile. Thus, these new training data samples would improve our machine learning models given the current test set.

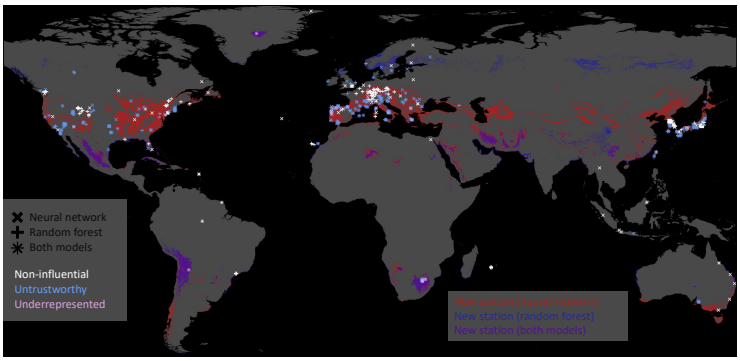


Figure 11. Map showing non-influential training stations in white, untrustworthy predictions on the test set in blue, and underrepresented test stations in plum. The marker indicates the model on which these stations were derived. The neural network's marker is $\times$, the random forest's marker is $+$, and where those two symbols overlap, we get something like an asterisk $*$. Moreover, we indicate regions where the models recommend building new stations in red, blue, and violet. The neural network recommends building new stations in red-colored areas; the random forest recommends building new stations in the blue-colored areas. Violet represents the intersection of regions recommended by both models.

6.5. Training Without Irrelevant Training Samples

We noticed that many training stations are not nearest neighbors to any of the stations within the test set during our analysis of the test set predictions. The random forest does not base any prediction upon them, and for the neural network, these training stations cause different activation patterns, leading them to not be even amongst the 100 nearest neighbors of any test sample. Given the assumption that we only want to predict the average ozone on our test set, we can argue that these might be irrelevant training samples. An overview of the location of these non-influential stations is given in Figure 11. We trained both machine learning models on a reduced training dataset from scratch to test if we could get a comparable performance by leaving out these training samples. Table 4 shows the RMSE and coefficient of determination of the models trained on the reduced training dataset.

Table 4. Scores on the test set for the reduced training set excluding 10% of the non-influential training data samples.

Random Forest	R^2 [%]	**RMSE [ppb]**
Reference	53.03	4.46
Test	52.32	4.49
Neural Network		
Reference	49.46	4.59
Test	47.45	4.72

Leaving out the non-influential training stations leads to slight decreases in the coefficient of determination on the test set. For the random forest, the R^2 value decreases by 1%, while the loss in accuracy is slightly higher for the neural network, around 2%. The RMSE values of both models increase by 0.03 ppb for the random forest and 0.13 ppb for the neural network.

7. Discussion

The following discussion is based on several assumptions. First, we assume that the SHAP values, which indicate the impact a feature has on the prediction, are related to the global importance of a feature when taking the entire set of SHAP values into account. Moreover, to use the Euclidean distance as a measure for similarity, we assume that the weighted feature space and the representation space are smooth. On top of this, we suppose that the Euclidean distance in the weighted feature space and representation space reflects similar samples and similar prediction patterns. We also assume that the weights in the neural network and the structure of the decision trees within the random forest have meaning. Finally, we assume that the k-nearest neighbors in the representations space are the influential training samples for the prediction. This assumption is weak for the random forest since we identified the training samples sharing leaf nodes with the predicted test sample. It is a somewhat stronger assumption for the neural network, where we cannot verify if the training stations we identified as k-nearest neighbors in the representation space are the stations on which the prediction on the test sample is based.

The random forest achieves a higher R^2 score and a lower RMSE than the neural network on the training data set. However, both models achieve similar R^2 scores differing by 3.5% on the test set (Section 6.2). The comparison of the residuals of the neural network and random forest shows that both models have difficulties of accurately predicting a subset of the test samples, which points to shortcomings of the AQ-Bench dataset rather than poorly fitted models.

To understand the difference between an accurate prediction and an inaccurate prediction in the models' representation space, we visualize the signal activation of the neural network and the leaf activation of the random forest (Section 6.3). In both cases, the patterns within the models' activation differ between an accurate and an inaccurate prediction

(Figures 7 and 8). These prediction patterns, which are representations of AQ-Bench samples in the model representation space, can be used to classify inaccurate predictions by the reason the prediction failed. Section 4.4 defines cases based upon the distance to the nearest neighbors in the model representation space and the weighted feature space. The numerals in the names of the cases point to the model's representation of the data. Case-I points to consistent inaccurate predictions, case-II points to inaccurate predictions, and case-III points to accurate predictions. On top of the model representation, we analyze the weighted feature space where we defined cases. In case-A, the test sample is among its nearest neighbors of the training set, while in case-B, it is far away from the training samples classified as nearest neighbors in the model representation space. In the following, we first discuss case-I and case-III because case-II gives more insights to AQ-Bench.

The first conclusion we draw from the analysis in Section 6.4 is that samples are assigned to any case except case-I-A and case-I-B, which means that both models are well fitted. Furthermore, case-III represents all accurate predictions. Over 93% of the predictions can be assigned to case-III-A for both models, which shows that most of these samples are not affected by the Clever-Hans effect (case-III-B). Although there is a difference between the neural network and the random forest, the neural network detected nearly nine times more often Clever-Hans predictions than the random forest.

In contrast to case-I and case-III, the discussion of case-II is diverse. Test predictions assigned to case-II are unexpected and inaccurate, while the k-nearest neighbor predictions are accurately predicted. Based on the examination of the weighted feature space, it is possible to identify underrepresented samples and untrustworthy predictions. The explanations lead to further insights about the AQ-Bech dataset and both models' predictions, as discussed in the following.

Overall, we found 0.5% underrepresented test samples for the random forest and 2% for the neural network. We suppose the data split causes the low rates of underrepresented test samples because Betancourt et al. [11] follow good practices of a dataset design, taking into account spatial correlations, data distribution, and representation ability.

Nevertheless, there is an overlap between the test samples identified as underrepresented in the training dataset, leading to areas where we recommend building new ozone observation stations based on both models (violet areas, see Figure 11). We chose machine learning as an alternative method to propose new station locations, which is a task that is also tackled by using an atmospheric chemistry model [42]. Although we show that the number of underrepresented test samples is not a significant issue for the prediction on the test dataset, underrepresented locations become problematic in the case of applying the models to areas outside the AQ-Bench dataset, e.g., in (global) mapping studies [13,41,43].

We also identified training samples that are non-influential when making predictions on the test set shown in Section 5.6. Those samples were either rarely or not included in the set of the 100 nearest neighbors and never used as auxiliary predictions. Neural network and random forest show slight differences regarding which subset of training samples are non-influential, but both agree on a set of roughly 5%. The non-influential stations are either located in data-dense regions or data-sparse regions. We interpret non-influential stations appearing in a data-dense region as redundancy in the training dataset. In contrast, non-influential stations in data-sparse areas are attributed to rare feature combinations not present in the test dataset and therefore are not needed to make accurate predictions on the test set. We further observe training samples in areas with sparse observations that are non-influential for one model but influential for the other one (Figure 11). One model recommends adding more stations in these areas while the other model flagged the available station as non-influential, highlighting the differences in the models' representations. This is underlined by the SHAP importance (Table 2) that shows that the models primarily base their predictions on different features. The spatial distribution of the new building locations in Figure 11 shows the strong influence of the feature absolute latitude. Areas, where we recommend building new stations based on

the model's results, are distributed across zonal bands and are characterized by relevant feature combinations.

The majority of the inaccurately predicted test samples of approximately 22% for both models belong to the case-II-A of an unexpected inaccurate prediction (Section 6.4). Here, the auxiliary predictions of the 11 nearest neighbors are accurate, and these training samples are also nearest neighbors in the weighted features space. We flag these predictions as untrustworthy because we do not trust the decision process they follow, as detailed in the following. There are two possible reasons for an inaccurate test sample prediction while the nearest neighbor predictions are accurate. The first reason is that the test sample's ozone value is erroneous, which might be due to an error in the observation. The AQ-Bench is a reliable benchmark dataset originating from a trustworthy data source. Errors in ozone values could occur in single cases, but it is doubtful that 22% of the data are erroneous. The second reason is the relationship between the features and their importance and the target average ozone deviates for these samples. The features in AQ-Bench are a variety of characteristics describing the environment around the measurement station and are proxies for precursors and atmospheric variables. There is no direct chemical relationship between environmental characteristics and average ozone. As a result, possible relevant features are missing, and the relation between features and target cannot be represented sufficiently because the system is underdetermined. Therefore, we attribute the untrustworthy samples to unique relationship between features and targets not reflected in the learned models.

8. Conclusions

In this study, we present various ways of using explainable machine learning to understand the core functionality of different machine learning models to support our understanding of the underlying dataset. Although AQ-Bench consists of proxies for chemical processes, we can gain new scientific insights and understand how different machine learning architectures use the input data to derive their predictions. By analyzing inaccurate predictions within the representation space of the machine learning models and assessing their k-nearest neighbors of the inaccurate predictions in the feature space, we draw conclusions about data representation and flag untrustworthy predictions. Moreover, our analysis also shows that given our current test dataset, irrelevant training samples exist, which we can drop without significant deterioration of model performance. Our experiments conclude that our machine learning models trained on geospatial air quality data do not represent the chemical relationships but rather found patterns in comparable training samples. Based on these learned patterns, both models construct the predictions with slightly different feature importance. Therefore, both models need enough representative and variable training samples to correctly reproduce prediction patterns required for the full range of predictions.

Author Contributions: Conceptualization, S.S. and C.B.; methodology, S.S. and C.B.; software, C.B.; validation, R.R.; formal analysis, S.S.; investigation, S.S. and C.B.; data curation, C.B.; writing—original draft preparation, S.S.; writing—review and editing, S.S., C.B. and R.R.; visualization, S.S. and C.B.; supervision, R.R.; project administration, S.S. All authors have read and agreed to the submitted version of the manuscript.

Funding: S.S. and C.B. acknowledge funding from ERC-2017-ADG#787576 (IntelliAQ). S.S also acknowledges funding by the German Federal Ministry of Education and Research 67KI2043 (KI:STE). R.R. acknowledges funding by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

Data Availability Statement: The data presented in this study are openly available. The AQ-Bench dataset [11] is available under the DOI http://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f. The gridded data of the AQ-Bench variables are available under the DOI http://doi.org/10.23728/b2share.9e88bc269c4f4dbc95b3c3b7f3e8512c.

Acknowledgments: First of all, we acknowledge the unlimited trust by Martin Schultz and the freedom he gave us to experiment, explore, and apply explainable machine learning for air quality research. We also acknowledge the fruitful discussions with Hanna Meyer and Marvin Ludwig. We want to thank our colleagues Vincent Gramlich, Lukas H. Leufen, Felix Kleinert, and Ankit Patnala for their feedback on our challenges. We are grateful for the technical support by Timo Stomberg and Ann-Kathrin Edrich. We also want to thank our supportive family and friends..

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. SHAP Importance with Feature Variable Names

Table A1. Feature importance derived by SHAP. The first column lists the names of the AQ-Bench features; we kept the order from the most important to the least important in the test set of our random forest. Percentage values for the random forest are shown in the second column; corresponding values of the neural network are shown in the third column.

Feature	Importance RF [%]	Importance NN [%]
lat	23.96	20.5
relative_alt	16.21	11.93
alt	10.44	8.16
nightlight_5km	9.73	4.35
forest_25km	5.37	13.54
population_density	4.41	1.8
nightlight_1km	4.12	8.77
water_25km	4.11	7.5
max_population_density_25km	3.54	0.32
no2_column	3.51	6.31
max_population_density_5km	3.04	1.21
savannas_25km	2.68	0.84
croplands_25km	1.73	1.74
grasslands_25km	1.68	0.77
nox_emissions	1.62	0.84
climatic_zone_warm_dry	1.18	5.27
shrublands_25km	0.65	1.67
max_nightlight_25km	0.35	0.39
climatic_zone_warm_moist	0.33	2.49
climatic_zone_cool_moist	0.32	0.14
rice_production	0.3	0.58
permanent_wetlands_25km	0.27	0.15
climatic_zone_cool_dry	0.25	0.13
climatic_zone_tropical_dry	0.14	0.13
climatic_zone_tropical_wet	0.03	0.11
climatic_zone_tropical_moist	0.02	0.09
climatic_zone_boreal_moist	0.0	0.11
climatic_zone_polar_moist	0.0	0.07
climatic_zone_boreal_dry	0.0	0.1
climatic_zone_polar_dry	0.0	0.0
climatic_zone_tropical_montane	0.0	0.0

Appendix B. Hyperparameters

Table A2. Hyperparameters of the random forest and the neural network.

Random Forest	
Number of trees	100
Criterion	RMSE
Depth	Unlimited
Bootstrapping	Training samples
Neural Network	
Learning rate	1.0×10^{-4}
L2 lambda	5.0×10^{-2}
Batch size	No mini batches
Number of epochs	15,000

Appendix C. Trained Neural Network Visualization

Figure A1 shows the trained neural network's weights and biases. The strength between the connections of input features and the first layer is consistent with the global importance given by SHAP. Moreover, we can note nodes with certain roles in the second hidden layer: A node activated to reduce the final predicted value (blue connection to output node) and three nodes (red/orange connections) that can be activated to increase the final prediction.

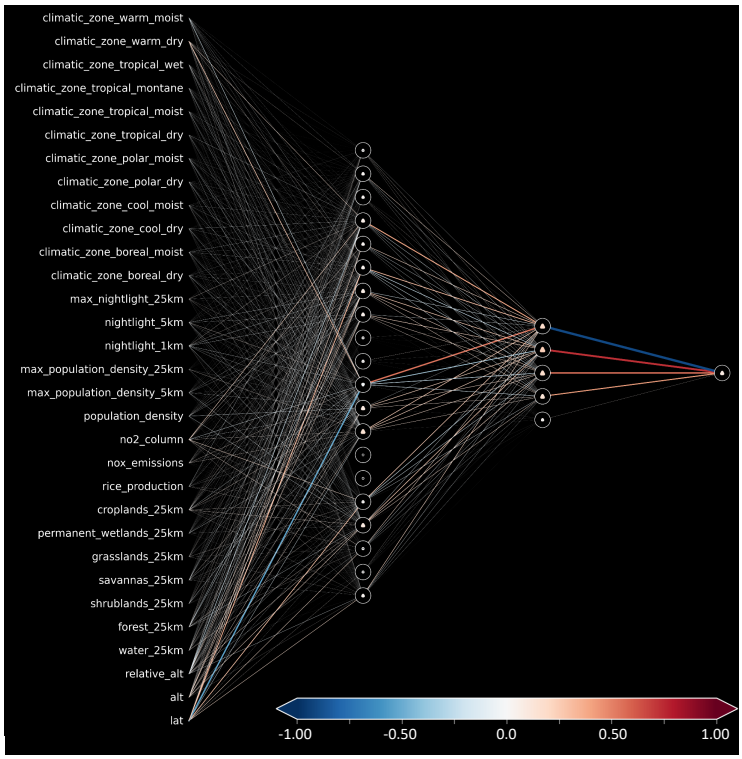


Figure A1. Weight and biases for our trained shallow neural network.

References

1. 4.2 Million Deaths Every Year Occur as a Result of Exposure to Ambient (Outdoor) Air Pollution. Available online: https://www.who.int/health-topics/air-pollution#tab=tab_1 (accessed on 12 December 2021).
2. Schultz, M.G.; Akimoto, H.; Bottenheim, J.; Buchmann, B.; Galbally, I.E.; Gilge, S.; Helmig, D.; Koide, H.; Lewis, A.C.; Novelli, P.C.; et al. The Global Atmosphere Watch reactive gases measurement network. *Elem. Sci. Anth.* **2015**, *3*. [CrossRef]

3. Schultz, M.G.; Schröder, S.; Lyapina, O.; Cooper, O.; Galbally, I.; Petropavlovskikh, I.; Von Schneidemesser, E.; Tanimoto, H.; Elshorbany, Y.; Naja, M.; et al. Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations. *Elem. Sci. Anth.* **2017**, *5*, 58. [CrossRef]
4. Gaudel, A.; Cooper, O.R.; Ancellet, G.; Barret, B.; Boynard, A.; Burrows, J.P.; Clerbaux, C.; Coheur, P.F.; Cuesta, J.; Cuevas, E.; et al. Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation. *Elem. Sci. Anth.* **2018**, *6*, 39. [CrossRef]
5. Rao, S.T.; Galmarini, S.; Puckett, K. Air Quality Model Evaluation International Initiative (AQMEII) advancing the state of the science in regional photochemical modeling and its applications. *Bull. Am. Meteorol. Soc.* **2011**, *92*, 23–30. [CrossRef]
6. Schultz, M.G.; Stadtler, S.; Schröder, S.; Taraborrelli, D.; Franco, B.; Krefting, J.; Henrot, A.; Ferrachat, S.; Lohmann, U.; Neubauer, D.; et al. The chemistry–climate model ECHAM6.3-HAM2.3-MOZ1.0. *Geosci. Model Dev.* **2018**, *11*, 1695–1723. [CrossRef]
7. Wagner, A.; Bennouna, Y.; Blechschmidt, A.M.; Brasseur, G.; Chabrillat, S.; Christophe, Y.; Errera, Q.; Eskes, H.; Flemming, J.; Hansen, K.; et al. Comprehensive evaluation of the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis against independent observations: Reactive gases. *Elem. Sci. Anth.* **2021**, *9*, 00171. [CrossRef]
8. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.* **2019**, *119*, 285–304. [CrossRef]
9. Kleinert, F.; Leufen, L.H.; Schultz, M.G. IntelliO3-ts v1.0: A neural network approach to predict near-surface ozone concentrations in Germany. *Geosci. Model Dev.* **2021**, *14*, 1–25. [CrossRef]
10. Stirnberg, R.; Cermak, J.; Kotthaus, S.; Haeffelin, M.; Andersen, H.; Fuchs, J.; Kim, M.; Petit, J.E.; Favez, O. Meteorology-driven variability of air pollution (PM1) revealed with explainable machine learning. *Atmos. Chem. Phys. Discuss.* **2020**, *2020*, 1–35. [CrossRef]
11. Betancourt, C.; Stomberg, T.; Roscher, R.; Schultz, M.G.; Stadtler, S. AQ-Bench: A benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data* **2021**, *13*, 3013–3033. [CrossRef]
12. Gu, J.; Yang, B.; Brauer, M.; Zhang, K.M. Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models. *Atmos. Environ.* **2021**, *246*, 118125. [CrossRef]
13. Betancourt, C.; Stomberg, T.T.; Edrich, A.-K.; Patnala, A.; Schultz, M.G.; Roscher, R.; Kowalski, J.; Stadtler, S. Global, high-resolution mapping of tropospheric ozone—Explainable machine learning and impact of uncertainties. *Geosci. Model Dev. Discuss.* **2022**, (*in preparation*).
14. Tuia, D.; Roscher, R.; Wegner, J.D.; Jacobs, N.; Zhu, X.; Camps-Valls, G. Toward a Collective Agenda on AI for Earth Science Data Analysis. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 88–104. [CrossRef]
15. Brokamp, C.; Jandarov, R.; Rao, M.; LeMasters, G.; Ryan, P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmos. Environ.* **2017**, *151*, 1–11. [CrossRef]
16. Mallet, M.D. Meteorological normalisation of PM10 using machine learning reveals distinct increases of nearby source emissions in the Australian mining town of Moranbah. *Atmos. Pollut. Res.* **2021**, *12*, 23–35. [CrossRef]
17. AlThuwaynee, O.F.; Kim, S.W.; Najemaden, M.A.; Aydda, A.; Balogun, A.L.; Fayyadh, M.M.; Park, H.J. Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. *Environ. Sci. Pollut. Res.* **2021**, *28*, 1–23. [CrossRef]
18. Tian, Y.; Yao, X.A.; Mu, L.; Fan, Q.; Liu, Y. Integrating meteorological factors for better understanding of the urban form-air quality relationship. *Landsc. Ecol.* **2020**, *35*, 2357–2373. [CrossRef]
19. Lu, H.; Xie, M.; Liu, X.; Liu, B.; Jiang, M.; Gao, Y.; Zhao, X. Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China. *Atmos. Pollut. Res.* **2021**, *12*, 101066. [CrossRef]
20. Alimissis, A.; Philippopoulos, K.; Tzanis, C.; Deligiorgi, D. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmos. Environ.* **2018**, *191*, 205–213. [CrossRef]
21. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099. [CrossRef]
22. Sayeed, A.; Choi, Y.; Eslami, E.; Jung, J.; Lops, Y.; Salman, A.K.; Lee, J.B.; Park, H.J.; Choi, M.H. A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Sci. Rep.* **2021**, *11*, 10891. [CrossRef] [PubMed]
23. McGovern, A.; Lagerquist, R.; Gagne, D. Using machine learning and model interpretation and visualization techniques to gain physical insights in atmospheric science. In Proceedings of the ICLR AI for Earth Sciences Workshop, Online, 29 April 2020.
24. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
25. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115.
26. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*; Citeseer: Princeton, NJ, USA, 2014.
27. Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. Visualizing higher-layer features of a deep network. *Univ. Montr.* **2009**, *1341*, 1.

28. Yan, X.; Zang, Z.; Luo, N.; Jiang, Y.; Li, Z. New interpretable deep learning model to monitor real-time PM2. 5 concentrations from satellite data. *Environ. Int.* **2020**, *144*, 106060. [CrossRef]
29. Bennett, A.; Nijssen, B. *Explainable AI Uncovers How Neural Networks Learn to Regionalize in Simulations of Turbulent Heat Fluxes at FluxNet Sites*; Earth and Space Science Open Archive ESSOAr: Washington, DC, USA, 2021.
30. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
31. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **2020**, *8*, 42200–42216. [CrossRef]
32. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017 Proceedings)*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; NeurIPS: Long Beach, CA, USA, 2017; pp. 4765–4774.
33. Toms, B.A.; Barnes, E.A.; Hurrell, J.W. Assessing Decadal Predictability in an Earth-System Model Using Explainable Neural Networks. *Geophys. Res. Lett.* **2021**, e2021GL093842. [CrossRef]
34. Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H.G.; Mahlein, A.K.; Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2020**, *2*, 476–486. [CrossRef]
35. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **2019**, *10*, 1096. [CrossRef]
36. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
37. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*, 1st ed.; MIT Press Cambridge: Cambridge, UK, 2016.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
40. Bilgin, Z.; Gunestas, M. Explaining Inaccurate Predictions of Models through k-Nearest Neighbors. In Proceedings of the International Conference on Agents and Artificial Intelligence, Online, 4–6 February 2021; pp. 228–236.
41. Meyer, H.; Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [CrossRef]
42. Sofen, E.; Bowdalo, D.; Evans, M. How to most effectively expand the global surface ozone observing network. *Atmos. Chem. Phys.* **2016**, *16*, 1445–1457. [CrossRef]
43. Petermann, E.; Meyer, H.; Nussbaum, M.; Bossew, P. Mapping the geogenic radon potential for Germany by machine learning. *Sci. Total Environ.* **2021**, *754*, 142291. [CrossRef]

D.3 Third paper (Betancourt et al., 2022)

Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties

Clara Betancourt, Timo T. Stomberg, Ann-Kathrin Edrich, Ankit Patnala, Martin G. Schultz, Ribana Roscher, Julia Kowalski, and Scarlet Stadtler

Journal: Geoscientific Model Development (2022), Vol. 15, No. 11
Status: published (June 2022)
DOI: 10.5194/gmd-15-4331-2022

Geosci. Model Dev., 15, 4331–4354, 2022
https://doi.org/10.5194/gmd-15-4331-2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development
Open Access
EGU

Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties

Clara Betancourt[1], Timo T. Stomberg[2], Ann-Kathrin Edrich[3,5], Ankit Patnala[1], Martin G. Schultz[1], Ribana Roscher[2,4], Julia Kowalski[5], and Scarlet Stadtler[1]

[1]Jülich Supercomputing Centre, Jülich Research Centre, Wilhelm-Johnen-Straße, 52425 Jülich, Germany
[2]Institute of Geodesy and Geoinformation, University of Bonn, Niebuhrstraße 1a, 53113 Bonn, Germany
[3]Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Schinkelstrasse 2a, 52062 Aachen, Germany
[4]Data Science in Earth Observation, Technical University of Munich, Lise-Meitner-Str. 9, 85521 Ottobrunn, Germany
[5]Methods for Model-based Development in Computational Engineering, RWTH Aachen University, Eilfschornsteinstr. 18, 52062 Aachen, Germany

Correspondence: Scarlet Stadtler (s.stadtler@fz-juelich.de)

Received: 5 January 2022 – Discussion started: 19 January 2022
Revised: 14 April 2022 – Accepted: 11 May 2022 – Published: 3 June 2022

Abstract. Tropospheric ozone is a toxic greenhouse gas with a highly variable spatial distribution which is challenging to map on a global scale. Here, we present a data-driven ozone-mapping workflow generating a transparent and reliable product. We map the global distribution of tropospheric ozone from sparse, irregularly placed measurement stations to a high-resolution regular grid using machine learning methods. The produced map contains the average tropospheric ozone concentration of the years 2010–2014 with a resolution of $0.1° \times 0.1°$. The machine learning model is trained on AQ-Bench ("air quality benchmark dataset"), a pre-compiled benchmark dataset consisting of multi-year ground-based ozone measurements combined with an abundance of high-resolution geospatial data.

Going beyond standard mapping methods, this work focuses on two key aspects to increase the integrity of the produced map. Using explainable machine learning methods, we ensure that the trained machine learning model is consistent with commonly accepted knowledge about tropospheric ozone. To assess the impact of data and model uncertainties on our ozone map, we show that the machine learning model is robust against typical fluctuations in ozone values and geospatial data. By inspecting the input features, we ensure that the model is only applied in regions where it is reliable.

We provide a rationale for the tools we use to conduct a thorough global analysis. The methods presented here can thus be easily transferred to other mapping applications to ensure the transparency and reliability of the maps produced.

1 Introduction

Tropospheric ozone is a toxic trace gas and a short-lived climate forcer (Gaudel et al., 2018). Contrary to stratospheric ozone which protects humans and plants from ultraviolet radiation, tropospheric ozone causes substantial health impairments to humans because it destroys lung tissue (Fleming et al., 2018). It is also the cause of major crop loss, as it damages plant cells and leads to reduced growth and seed production (Mills et al., 2018). Tropospheric ozone is a secondary pollutant with no direct sources but with formation cycles depending on photochemistry and precursor emissions. It is typically formed downwind of precursor sources from traffic, industry, vegetation, and agriculture, under the influence of solar radiation. Ozone patterns are also influenced by the local topography causing specific flow patterns (Monks et al., 2015; Brasseur et al., 1999). Depending on the on-site conditions, ozone can be destroyed in a matter of minutes or have a lifetime of several weeks with advection from source regions to remote areas (Wallace and Hobbs, 2006). The interrelation

of these factors of ozone formation, destruction, and transport is not fully understood (Schultz et al., 2017). This makes ozone both difficult to quantify and to control. Public authorities recognize ozone-related problems. They install air quality monitoring networks to quantify ozone (Schultz et al., 2015, 2017). Furthermore, they enforce maximum exposure rules to mitigate ozone health and vegetation impacts (e.g., European Union, 2008).

Currently, there is increased use of machine learning methods in tropospheric ozone research. Such "intelligent" algorithms can learn nonlinear relationships of ozone processes and connect them to environmental conditions, even if their interrelations are not well understood through process-oriented research. Kleinert et al. (2021) and Sayeed et al. (2021) used convolutional neural networks to forecast ozone at several hundred measurement stations, based on meteorological and air quality data. Large training datasets allowed them to train deep neural networks, resulting in a significant improvement over the first machine learning attempts to forecast ozone (Comrie, 1997; Cobourn et al., 2000). Machine learning is also used to calibrate low-cost ozone monitors that complement existing ozone monitoring networks (Schmitz et al., 2021; Wang et al., 2021). Furthermore, compute-intensive chemical reactions schemes for numerical ozone modeling can be emulated using machine learning (Keller et al., 2017; Keller and Evans, 2019). Ozone datasets which are used as training data for machine learning models are increasingly made available as FAIR (Wilkinson et al., 2016) and open data. AQ-Bench ("air quality benchmark dataset", Betancourt et al., 2021b), for example, is a dataset for machine learning on global ozone metrics and serves as training data for this mapping study.

We refer to mapping as a data-driven method for spatial predictions of environmental target variables. For mapping, a model is fit to observations of the target variable at measurement sites, which might even be sparse and irregularly placed. Environmental features are used as proxies for the target variable to fit the model. A map of the target variable is produced by applying the model to the spatially continuous features in the mapping domain. Mapping for environmental applications has been performed since the 1990s (Mattson and Godfrey, 1994; Briggs et al., 1997). It was deployed for air pollution as an improvement over spatial interpolation and dispersion modeling, which suffer from performance issues due to sparse measurements, and lack of detailed source description (Briggs et al., 1997). Hoek et al. (2008) describe these early mapping studies as "linear models with little attention to mapping outside the study area". In contrast, modern machine learning algorithms are often trained on thousands of samples for mapping (Petermann et al., 2021; Heuvelink et al., 2020). Several studies (e.g., Li et al., 2019; Ren et al., 2020) have shown that mapping using machine learning methods is superior to other geostatistical methods such as Kriging because it can capture nonlinear relationships and makes ideal use of environmental features by exploiting similarities between distant sites. In contrast to traditional interpolation techniques, mapping allows to extend the domain to the global scale, because it can predict the variable of interest based on environmental features, even in regions without measurements (Lary et al., 2014; Bastin et al., 2019; Hoogen et al., 2019). Recently, it is questioned whether machine learning methods are the most suitable to "map the world" (Meyer, 2020): Meyer et al. (2018) and Ploton et al. (2020) point out that some studies may be overconfident because they validate their maps on data that are not statistically independent from the training data. This occurs when a random data split is used on data with spatiotemporal (auto)correlations. There are also concerns when the mapping models are applied to areas that have completely different properties from the measurement locations (Meyer and Pebesma, 2021). A model trained on certain input feature combinations can only be applied to similar feature combinations. Furthermore, uncertainty estimates of the produced maps are important as they are often used as a basis for further research.

In this study, we produce the first fully data-driven global map of tropospheric ozone, aggregated in time over the years 2010–2014. This study builds upon Betancourt et al. (2021b) who proved that ozone metrics can be predicted using static geospatial data. We provide the map as a product and combine it with uncertainty estimates and explanations to ensure the trustworthiness of our results. We justify the choice of methods and clarify why they are necessary for a thorough global analysis. Section 2 contains a description of the data and machine learning methods, including explainable machine learning and uncertainty estimation. Section 3 contains the results, which are discussed in Sect. 4. We conclude in Sect. 5.

2 Data and methods

2.1 Data description

In this section, we present the datasets used in this study. Technical details on these data are given in Appendix A.

2.1.1 AQ-Bench dataset

We fit our machine learning model on the AQ-Bench dataset ("air quality benchmark dataset", Betancourt et al., 2021b). The AQ-Bench dataset is a machine learning benchmark dataset that allows to relate ozone statistics at air quality measurement stations to easy-access geospatial data. It contains aggregated ozone statistics of the years 2010–2014 at 5577 stations around the globe, compiled from the database of the Tropospheric Ozone Assessment report (TOAR, Schultz et al., 2017). The AQ-Bench dataset considers ozone concentrations on a climatological time scale instead of day-to-day air quality data. The scope of this dataset is to discover purely spatial relations. Machine

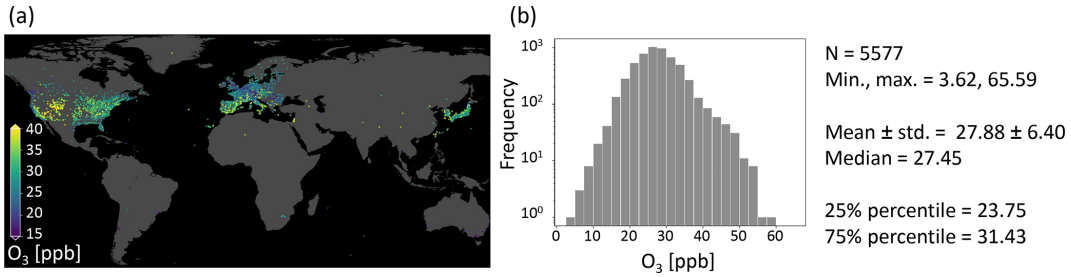


Figure 1. Average ozone statistic of the AQ-Bench dataset. The values at 5577 measurement stations are aggregated over the years 2010–2014. **(a)** Values on a map projection. **(b)** Histogram and summary statistics.

learning models trained on this dataset will output aggregated statistics over the years 2010–2014 and will not be able to capture temporal variances. This is beneficial if the required final data products are also aggregated statistics. The majority of the stations are located in North America, Europe, and East Asia. The dataset contains different kinds of ozone statistics such as percentiles or health-related metrics. This study focuses on the average ozone statistic as the target (Fig. 1).

The features in the AQ-Bench dataset characterize the measurement site and are proxies for ozone formation, destruction, and transport processes. For example, the "altitude" and "relative altitude" of the station are important proxies for local flow patterns and ozone sinks. "Population density" in different radii around every station are proxies for human activity and thus ozone precursor emissions. "Latitude" is a proxy for ozone formation through photochemistry, as radiation and heat generally increase towards the Equator. The land cover variables are proxies for precursor emissions and deposition. The full list of features and their relation to ozone processes are documented by Betancourt et al. (2021b). Figure 2 shows predictions of a machine learning model on the test set of AQ-Bench. Table 1 lists all features used in this study.

2.1.2 Gridded data

Features are needed on a regular grid (i.e., as raster data) over the entire mapping domain to map the target average ozone. The original gridded data used here (Appendix Sects. A and B) has a resolution of $0.1° \times 0.1°$ or finer. Since our target resolution is $0.1° \times 0.1°$, the gridded data are downscaled to that resolution if the original resolution is finer. The "land cover", "population", and "light pollution" features of the AQ-Bench dataset are spatial aggregates in a certain radius around the station (see Table 1). To prepare gridded fields of these features, the area around each individual grid point is considered, and the required radius aggregation is written to that grid point. The gridded dataset is available under the DOI https://doi.org/10.23728/b2share.9e88bc269c4f4dbc95b3c3b7f3e8512c (Betancourt et al., 2021c).

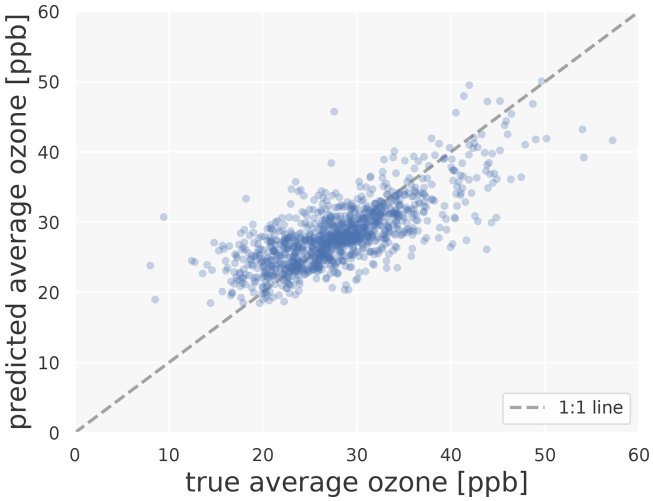


Figure 2. Predicted ozone values versus measurement values of the test set of the AQ-Bench dataset. See Sect. 3.3.1 for the specifications of the used machine learning model.

2.2 Explainable machine learning workflow

We apply a standard mapping workflow and extend it with explainable machine learning methods as described in this section. Together with the uncertainty assessment methods described in Sect. 2.3, they allow for a thorough analysis of our machine learning model. A random forest (Breiman, 2001) is fit on the AQ-Bench dataset to predict average ozone for given features. A random forest is an ensemble of regression trees that is created by bootstrapping the training dataset to increase generalizability. We choose random forest because tree-based models are the state of the art for structured data (Lundberg et al., 2020). Random forest was also shown to outperform linear regression and a shallow neural network in predicting average ozone on the AQ-Bench dataset (Betancourt et al., 2021b). In addition, this algorithm has been proven to be suitable for mapping in several studies (Petermann et al., 2021; Nussbaum et al., 2018; Ren et al., 2020). We use the Python framework SciKit-learn (Pedregosa et al., 2011) for machine learning and hyperactive (Blanke, 2021) for hyperparameter tuning.

Table 1. Features selected from the AQ-Bench dataset.

	Feature	Unit
General	Climatic zone	–
	Latitude	°
	Altitude	m
	Relative altitude	m
Land cover	Water in 25 km area	%
	Evergreen needleleaf forest in 25 km area	%
	Evergreen broadleaf forest in 25 km area	%
	Deciduous needleleaf forest in 25 km area	%
	Deciduous broadleaf forest in 25 km area	%
	Mixed forest in 25 km area	%
	Closed shrublands in 25 km area	%
	Open shrublands in 25 km area	%
	Woody savannas in 25 km area	%
	Savannas in 25 km area	%
	Grasslands in 25 km area	%
	Permanent wetlands in 25 km area	%
	Croplands in 25 km area	%
	Urban and built-up in 25 km area	%
	Cropland/natural vegetation mosaic in 25 km area	%
	Snow and ice in 25 km area	%
	Barren or sparsely vegetated in 25 km area	%
Agriculture	Wheat production	$1000\,\mathrm{t\,yr^{-1}}$
	Rice production	$1000\,\mathrm{t\,yr^{-1}}$
Ozone precursors	NO_x emissions	$\mathrm{g\,m^{-2}\,yr^{-1}}$
	NO_2 column	$10^5\,\mathrm{molec\,cm^{-2}}$
Population	Population density	$\mathrm{person\,km^{-2}}$
	Maximum population density in 5 km area	$\mathrm{person\,km^{-2}}$
	Maximum population density in 25 km area	$\mathrm{person\,km^{-2}}$
Light pollution	Nightlight in 1 km area	brightness index
	Nightlight in 5 km area	brightness index
	Maximum nightlight in 25 km area	brightness index

A proper validation strategy is crucial for spatial prediction models because both environmental conditions and target variables are often correlated in space. When tested on spatially correlated and thus statistically dependent samples, mapping results may be overconfident (Meyer et al., 2018; Ploton et al., 2020). We use the independent spatial data split provided with the AQ-Bench dataset to validate spatial generalizability. Details on our validation strategy are given in Sect. 2.2.1.

As an extension of the standard mapping workflow described in Sect. 1, we perform experiments to increase interpretability, test robustness, and explain the model. The extended workflow is summarized in Table 2 and further justified in the following.

The use of redundant features in mapping applications can favor spatial overfitting. We thus remove counterproductive features by forward feature selection as proposed by Meyer et al. (2018). Additionally, we apply basic feature engineering to increase the interpretability of the model. Details on feature engineering and feature selection are described in Sect. 2.2.2. In order to make our mapping model trustworthy, we verify its robustness and ability to generalize to unseen locations, and to explore the limits of its predictive capabilities. Noise in the AQ-Bench dataset causes problems if the model is not robust. Additionally, limited availability of ozone measurements in regions like central and southeast Asia, Central and South America, and Africa poses a problem as it is unclear whether our model will generalize to these regions. We address the issues of robustness and generalizability using the spatial cross-validation strategy described in Sect. 2.2.3.

We also aim to explain how the model arrives at its predictions and check consistency with common ozone process understanding by using SHAP (SHapley Additive exPlanations, Lundberg and Lee, 2017), a post hoc explainable

Table 2. Machine learning experiments as an addition to the standard mapping method. For details on the methods, refer to the given sections.

Section	Method	Goal
2.2.2	Feature engineering	Make features easier to interpret
	Forward feature selection	Remove counterproductive features which favor overfitting
2.2.3	Spatial cross validation	Check model spatial robustness
	Cross validation on world regions	Evaluate model generalizability
2.2.4	Calculate SHAP values	Explain model predictions

machine learning method. It is a game-theoretic approach based on Shapley values (Shapley, 1953). SHAP identifies the importance of the individual features to a model prediction (Sect. 2.2.4).

2.2.1 Evaluation scores

We rely on the independent 60 %–20 %–20 % data split of AQ-Bench as provided by Betancourt et al. (2021b). Here, stations with a distance of more than 50 km are considered independent of each other.

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

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

where m denotes a sample index, M the total number of samples, $\hat{y}_m$ a predicted target value, and y_m a reference target value. R^2 measures the proportion of variance in the output values that the model predicts. Thus, a larger R^2 represents a better model and the largest possible value is 1. We also evaluate the root mean square error (RMSE) in ppb:

$$\text{RMSE} = \sqrt{\sum_{m=1}^{M} \frac{(y_m - \hat{y}_m)^2}{M}}. \quad (2)$$

2.2.2 Feature engineering and feature selection

We perform basic feature engineering to improve the interpretability of our model. Different types of savanna, shrublands, and forests are given individually in AQ-Bench (Table 1). We merge them into "savanna", "forest", and "shrubland" because a high number of features with similar properties would make the model interpretation more difficult. Instead of "latitude", we train on the "absolute latitude", since radiation and temperature decrease when moving away from the Equator, regardless of whether one moves south or north. Compared to experiments performed without feature engineering, we did not see any change in evaluation scores.

We use the forward feature selection method for spatial prediction models by Meyer et al. (2018). The model is initially trained on all two-feature pairs. The pair with the highest evaluation score is kept. The model is then trained on each remaining feature along with the already selected features. The additional feature with the best evaluation score is appended to the existing list of features. This iterative approach is continued until the R^2 value drops, which indicates that a feature leads to overfitting. The selected features are presented in Sect. 3.1.1.

2.2.3 Spatial cross validation

We apply cross validation to prove the robustness of our model. We split the test and training set into four independent cross-validation folds of 20 % each. Like Betancourt et al. (2021b), we assume that air quality measurement stations with a distance of at least 50 km are independent of each other. We, therefore, produce the cross-validation folds with a two-step approach. First, we cluster the data based on the spatial location of the measurement sites using the density-based clustering algorithm DBSCAN (Ester et al., 1996). The maximum distance between clusters is set to 50 km so stations closer than that distance are assigned to the same cluster. Small clusters are randomly assigned to the cross-validation folds. In the second step, larger clusters ($n > 50$) are split with k-means clustering (Duda et al., 2001) to ensure the same statistical distribution of all cross-validation folds. The resulting smaller clusters are again randomly assigned to the cross-validation folds. Figure 3a shows this data split.

We extend our spatial cross-validation experiment to evaluate the generalizability of our predictions to world regions with few measurements. Here, we divide the data into the three world regions: North America, Europe, and East Asia (Fig. 3b). A random forest is fit and evaluated on two of the three regions and also evaluated on the third region for comparison. For example, it is fit and evaluated on data of Europe and North America and additionally evaluated in East Asia. The difference in the resulting evaluation scores shows the spatial generalizability of the model. The results are presented in Sect. 3.1.2.

2.2.4 SHapley Additive exPlanations

SHAP (Lundberg and Lee, 2017) provides detailed explanations for individual predictions by quantifying how each feature contributes to the result. The contribution refers to the average model output (or base value) over the train-

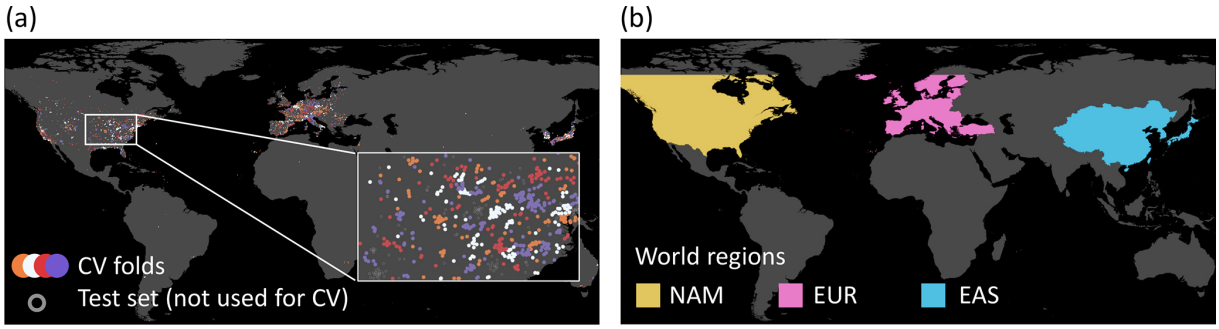


Figure 3. Data splits for the spatial cross validation. **(a)** Station clusters are randomly assigned to four cross-validation (CV) folds. **(b)** The data are divided by the world regions North America (NAM), Europe (EUR), and East Asia (EAS).

ing set: a feature with the SHAP value x causes the model to predict x more than the base value. We use the TreeShap module (Lundberg et al., 2018) of the Python package SHAP (Lundberg and Lee, 2017) to calculate SHAP values. Global feature importance is obtained by adding up all local contributions to the predictions. Features with high absolute contributions are considered more important. The SHAP values of our model are presented in Sect. 3.1.3.

2.3 Methods to assess the impact of uncertainties

Uncertainty assessment increases the trustworthiness of our machine learning approach and final ozone map. In general, the predictions of machine learning models have two kinds of uncertainties (Gawlikowski et al., 2021): first, model uncertainty, which results from the trained machine learning model itself, and second, data uncertainty which stems from the uncertainty inherent in the data. It is common to treat these uncertainties separately. Developing an uncertainty assessment strategy for our mapping approach is challenging because different uncertainties arise at different stages of the mapping process. Every ozone measurement, every preprocessing step, and every model prediction is a potential source of error. It would be infeasible to investigate the impacts of every error. We, therefore, identify the most important error sources and analyze the uncertainty induced in our produced map only for these. The decision on which aspects to analyze specifically is based on expert knowledge and the results of our machine learning experiments, i.e., robustness analysis (Sect. 2.2.3) and SHAP values (Sect. 2.2.4). We develop a formalized approach which is summarized in Table 3 and further elaborated in the following.

The model error is caused by the uncertainty of the trainable parameters of the model. It becomes visible, for example, when different results are obtained if the model is initialized with different random seeds before training (Petermann et al., 2021). To rule out this training instability, we retrained our models several times with different random seeds and monitored the results. We found negligible variations and thus rule out this kind of uncertainty. Apart from uncertainty through training instability, the model uncertainty is usually high for predictions in areas of the feature space where training data are sparse (Lee et al., 2017; Meyer and Pebesma, 2021). For example, a model that was not trained on data from very high mountains or deserts is not expected to produce reliable results in areas with these characteristics. We apply the concept of "area of applicability" by Meyer and Pebesma (2021) to limit our mapping to regions where our model is expected to produce reliable results. The details are described in Sect. 2.3.1.

The target variable "average ozone" is the first choice for assessment of data errors. Fluctuations and random measurement errors introduce uncertainty into the ozone measurements. We evaluate the uncertainty introduced by these influences in the map using a simple error model. The error model is used to perturb the training data, to check how the map changes when the model is trained on perturbed data instead of original data. The error model is described in Sect. 2.3.2.

Additional data uncertainty stems from the features. For example, geospatial data derived from satellite products are sensitive to retrieval errors. Based on the sources and documentation of our geospatial data (Appendix A), we expect such errors to have a small impact in this study. However, we inspect the subgrid features in the geospatial data and their effect on the model results. We limit ourselves to the "altitude" because our SHAP analysis (Sect. 3.1.3) has shown that it is the most important feature besides "latitude" which does not have critical subgrid variations. Subgrid variations of the altitude might influence our final map, especially if a feature like a cliff or a high mountain is present in the respective grid cell. We evaluate the influence of subgrid variations in altitude on the final map by propagating higher resolution altitudes through the final model as described in Sect. 2.3.2.

2.3.1 Area of applicability method

We adopt the area of applicability method from Meyer and Pebesma (2021). The method is based on considering the distance of a prediction sample to training samples in the feature space. This concept is illustrated in Fig. 4, where it can

Table 3. Uncertainty assessment for our mapping method. For details on the methods, refer to the given sections.

Section	Method	Goal
2.3.1	Define area of applicability	Ensure the model is only applied where it is reliable
2.3.2	Modeling of ozone fluctuations	Evaluate the impact of ozone fluctuations on produced map
2.3.3	Propagate subgrid altitude variation through model	Evaluate uncertainty introduced by altitude variation

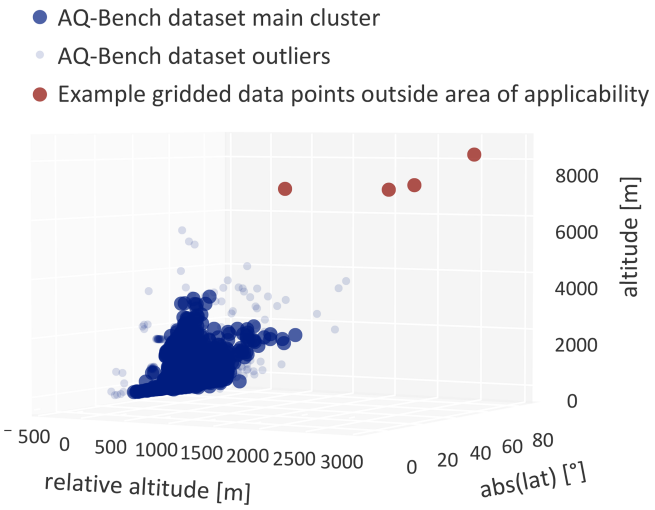


Figure 4. Principle of the area of applicability. The plot displays the distribution of all AQ-Bench samples along the three most important feature axes "absolute latitude", "altitude", and "relative altitude". It is clearly visible that the AQ-Bench samples form a cluster, and that some feature combinations in the gridded data are far away from this cluster.

be clearly seen that the AQ-Bench dataset forms a cluster in the feature space, but that our mapping domain contains feature combinations that do not belong to this cluster. Predictions made on these feature combinations suffer from high uncertainty. Consequently, we mark data points with a great distance to the training data cluster as "not predictable".

After we normalized the features, we scaled them accordingly to their global feature importance (Sect. 2.2.4) to increase their respective relevance. We use the cross-validation sets described in Sect. 2.2.3 to find a threshold distance for non-predictable samples. In detail, we calculate the distance from every training data point to the closest data point in a different cross-validation set. The threshold distance for "non-predictable" data is the upper whisker of all the cross-validation distances. Since the model is trained on land surface data only, we also remove the oceans from the area of applicability. The result of this experiment is shown in Sect. 3.2.1.

2.3.2 Modeling ozone fluctuations

Here, we describe our error model for evaluating the uncertainty introduced by typical ozone biases in the produced map. Such biases may arise from measurement uncertainties, local geographic effects, or an "unusual" environment with respect to precursor emission sources. We consider all of these effects as ozone measurement uncertainties although it would be more precise to say that they are uncertainties in the determination of ozone concentrations at the scale of our grid boxes.

Quantification of these uncertainties is challenging, as we typically lack the necessary local information. We, therefore, assume the local ozone values are subject to a Gaussian error of mean 0 ppb and variance 5 ppb (Sect. 4, Schultz et al., 2017). We randomly perturb a subset of the training ozone values with this Gaussian error and monitor resulting variances in the final map. Assuming only one-quarter of the measurement values are biased, 25 % of the training ozone values are either increased or decreased by random values in this Gaussian distribution. We use multiple realizations of this error model to perturb the training data, each realization perturbing a different subset with different values. One example error model realization is shown in Appendix C.

We train on the randomly perturbed data, obtain a "perturbed model", and then create "perturbed maps". If the perturbations of the resulting ozone maps are less or equal to the initial perturbations, the resulting uncertainty in the map is acceptable. If completely different maps would be produced, this would point to a model lacking robustness. The process of perturbing, training, and comparing maps is repeated until the standard deviation of all perturbed maps converges. The error model converged fully after 100 realizations (Appendix D). The result of this experiment is presented in Sect. 3.2.2.

2.3.3 Propagating subgrid altitude variation through model

In contrast to perturbing the targets and retraining the machine learning model, here we sample inputs from a finer resolution grid and propagate them through the existing trained model. For every grid cell of our final map with 0.1° resolution, we propagate all "altitude" values of the original finer resolution digital elevation model (DEM, resolution 1′, Appendix A) through our random forest model while leaving the other variables unchanged. For each coarse 0.1° resolution grid cell, we find 36 altitude values of the fine grid cells and can thus make 36 predictions. We monitor the deviation

of these predictions from the reference prediction in that cell. The results of these experiments are presented in Sect. 3.2.3.

3 Results

The results of our explainable machine learning mapping workflow (Sect. 2.2, Table 2) are presented in Sect. 3.1. The impact of uncertainties (Sect. 2.3, Table 3) are presented in Sect. 3.2. The final ozone map that is generated based on the knowledge gained from all experiments is presented in Sect. 3.3.

3.1 Explainable machine learning model

3.1.1 Selected hyperparameters and features

We choose the following standard hyperparameters for our random forest model: 100 trees are fit on bootstrapped versions of the AQ-Bench dataset with a mean square error (MSE) loss function and unlimited depth. The evaluation scores found to be insensitive to the choice of hyperparameters. Therefore, the standard hyperparameters are used to fit the model in all experiments of this study.

Based on the forward feature selection (Sect. 2.2.2), the following variables are used to build the model:

- climatic zone,
- absolute latitude,
- altitude,
- relative altitude,
- water in 25 km area,
- forest in 25 km area,
- shrublands in 25 km area,
- savannas in 25 km area,
- grasslands in 25 km area,
- permanent wetlands in 25 km area,
- croplands in 25 km area,
- rice production,
- NO_x emissions,
- NO_2 column,
- population density,
- maximum population density in 5 km area,
- maximum population density in 25 km area,
- nightlight in 1 km area,
- nightlight in 5 km area, and
- maximum nightlight in 25 km area.

Table 4. Four-fold cross-validation results.

Fold	R^2	RMSE [ppb]
1	0.64	3.83
2	0.58	4.03
3	0.61	4.04
4	0.61	3.97
∅	0.61 ± 0.02	3.97 ± 0.08

The following features are discarded because the validation R^2 score decreases when they are used to train the model: "urban and built-up in 25 km area", "cropland/natural vegetation mosaic in 25 km area", "snow and ice in 25 km area", "barren or sparsely vegetated in 25 km area", and "wheat production". A discussion of why these features are counterproductive follows in Sect. 4.1.

3.1.2 Spatial cross validation reveals limits in the model generalizability

The four-fold cross validation from Sect. 2.2.3 results in R^2 values in the range of 0.58 to 0.64 and RMSEs in the range of 3.83 to 4.04 ppb (Table 4). These evaluation scores show that all models are useful despite the variance in evaluation scores. The mean R^2 score is 0.61 and the mean RMSE is 3.97 ppb. Putting this RMSE value into perspective, 5 ppb is a conservative estimate for the ozone measurement error (Schultz et al., 2017). It is also lower than the 6.40 ppb standard deviation of the true ozone values of the training dataset (Fig. 1). Although the evaluation scores of all folds are in an acceptable range, the evaluation scores depend on the data split to some extend.

If our model is validated on a different region than it has been trained on, we observe a drop of the R^2 value by 0.13 to 0.49, while the RMSE increases for two of the three training regions (Table 5). One reason for the change in evaluation scores when training and validating in different world regions could be different feature combinations of the different world regions. We ruled out this reason by inspecting the feature space (similar to Sect. 2.3.1; not shown). The only other possible reason for the decrease in R^2 is that the relationship between features and ozone is not the same in different world regions. Therefore, the expected evaluation scores of our map vary not only with the feature combinations (as described in Sect. 2.3.1) but also spatially. We differentiate between the two issues and their influence on the model applicability in Sect. 3.2.1.

Table 5. Cross validation on the world regions Europe (EUR), East Asia (EAS), and North America (NAM). We give the difference in R^2 values and RMSEs when validating the model in another world region than the training region.

Training region	Validation region		R^2	RMSE [ppb]
EUR + EAS	EUR + EAS		0.57	3.54
	NAM		0.34	5.01
		diff.	−0.23	+1.47
EAS + NAM	EAS + NAM		0.52	3.76
	EUR		0.39	4.64
		diff.	−0.13	+0.88
NAM + EUR	NAM + EUR		0.63	3.92
	EAS		0.14	3.78
		diff.	−0.49	−0.14

3.1.3 SHAP values quantify the influence of the features on the model results

SHAP was used to determine the feature importance of the random forest model as described in Sect. 2.2.4. Figure 5 contains a summary plot with the global feature importance (left side) and SHAP values of all features on the test set (right side). The global importance of the features "absolute latitude", "altitude", "relative altitude", and "nightlight in 5 km area" are highest with a contribution of at least 10 %. The remaining features have a weaker influence on the model output. For example, the influence of the "climatic zone" is often negligible. The local SHAP values in Fig. 5 reveal the contribution of features to the predictions. A lower "absolute latitude" value leads to an increased ozone value prediction. Likewise, higher "altitude" and "relative altitude" increase predicted ozone values. High "nightlight in 5 km area" values lead to lower predicted ozone concentrations. These tendencies are in line with domain knowledge on the atmospheric chemistry of ozone. Appendix E shows SHAP values of two individual predictions. We discuss the physical consistency of the model based on the SHAP values in Sect. 4.1.

3.2 Evaluating the impact of uncertainties

3.2.1 Applicability and uncertainty of the model depend on both features and location

As described in Sect. 2.3.1, predictions of our model are valid if the feature combinations are similar to those of the training dataset. Additionally, the results of the spatial cross validation (Sect. 3.1.2) have shown that the spatial proximity to the training locations has an influence on the model performance and uncertainty. Two cases were examined in this section: firstly, the cross-validation sets which are close to each other (RMSE in the range of 4 ppb, as seen in Table 4), and secondly, the cross validation on different world regions (RMSE values of up to 5 ppb, as seen in Table 5). In our uncertainty assessment, we therefore combine findings from both the area of applicability (for matching features) and the spatial cross-validation methods (for spatial proximity).

Analogously to the approach of the area of applicability (Sect. 3.2.1), we analyze the distances between measurement stations in the geographical space. To quantify spatial proximity, we calculate the mean distance of a measurement station and its closest neighboring station in a different cross-validation set. Disregarding stations that are too far away from the others, we identified the distance of approximately 182 km (upper whisker), within which we expect a comparable RMSE as shown in Table 4. We assume a higher RMSE for locations that are more than 182 km away from their closest neighboring measurement station. Figure 6 shows the area of applicability of our model including this spatial distinction.

The majority of the regions with good coverage of measurement stations (North America, Europe, and parts of East Asia) are well predictable. In these regions, only some areas in the high north and high mountains are not predictable. Conversely, large areas in South and Central America, Africa, far northern regions, and Oceania have feature combinations different from the training data and therefore are not predictable. There are some regions in the Baltic area, South America, Africa, and south Australia where feature combinations can be predicted by the model, but they are far away from the AQ-Bench stations. A broader discussion of the global applicability of our machine learning model follows in Sect. 4.3.

3.2.2 Uncertainty due to ozone fluctuations is within an acceptable range

The error model for ozone uncertainties is described in Sect. 2.3.2. The R^2 values of the perturbed models varied between 0.50 and 0.58. Figure 7 shows the resulting standard deviation in the mapped ozone. The assumed ozone fluctuations have a higher impact in areas with sparse training data. We conclude that our error model does not tend to amplify

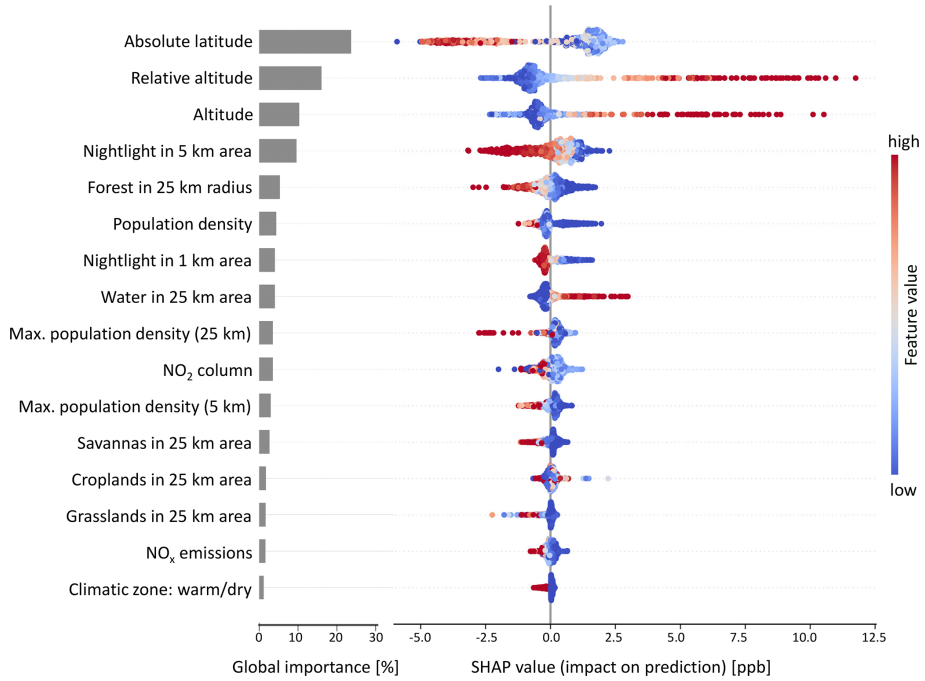


Figure 5. SHAP summary plot. The global importance on the left side is calculated from the averaged sum of the absolute SHAP values. The dots in the beeswarm plots on the right side show the SHAP values of single predictions. The color indicates the respective feature value. This plot shows only features with more than 1 % global importance.

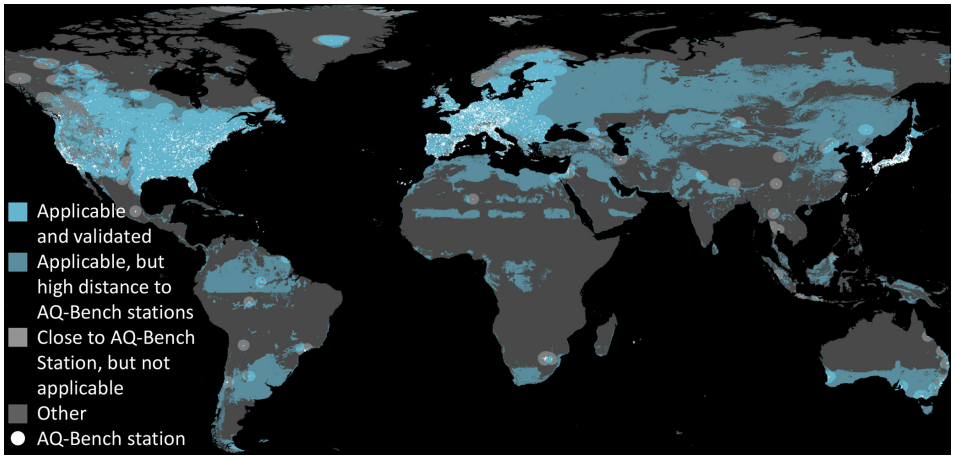


Figure 6. Area of applicability with restrictions in the feature space and spatial restrictions. The bright turquoise areas fulfill all prerequisites to be predictable: they have similar features as the AQ-Bench dataset and they are close to stations for validation. The darker shade of turquoise indicates similar predictions but no proximity to stations for validation. Light gray areas indicate the proximity of a station but no applicability of the model. The locations of all measurement stations are plotted in white.

the effects of perturbed training data. This means that the machine learning algorithm smoothes out noise during training. This is explained by the core functioning of the random forest which uses bootstrapping during training.

Figure 7 also shows that regions with poor spatial coverage by measurement stations (darker shade of turquoise in Fig. 6) are more sensitive to noisy training data. Example regions are the patches in Greenland, Africa, Australia, and South America. This is because the model relies its predictions on a few samples and is thus sensitive to perturbations of these few measurements.

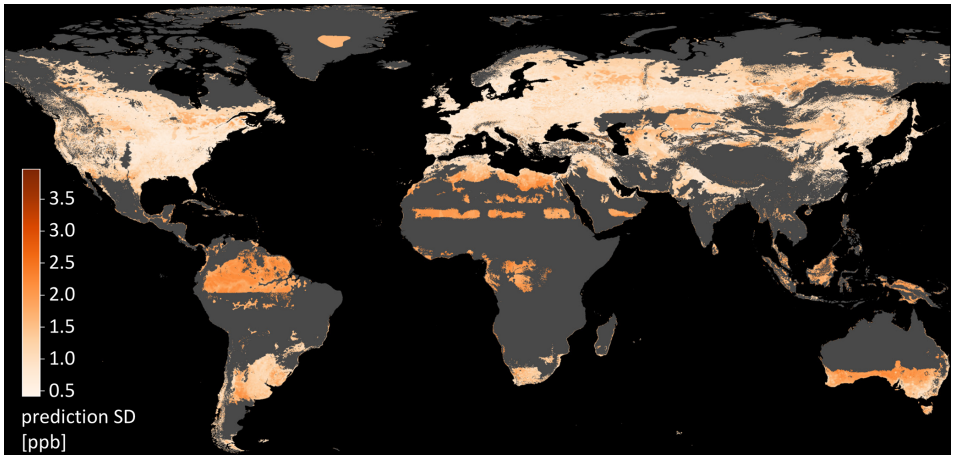


Figure 7. Standard deviation of the ozone predictions under perturbations. This map was created by stacking the maps of 100 error model realizations along the z axis and then calculating the grid point-wise standard deviation along the z axis.

3.2.3 Uncertainty through subgrid DEM variation is within an acceptable range

This method was described in Sect. 2.3.3. In most regions of the world, subgrid DEM variations around mean altitude are below 50 m (Fig. 8a), e.g., in the central and eastern United States and in Europe except for the Alps. There are regions with higher variances such as the Rocky Mountains and their surroundings, the Alps, and large parts of Japan outside Tokyo. In Fig. 8b, it can be seen how these variations influence the predicted ozone values. In the flat regions, the variance is below 0.5 ppb, and even in the high-variance regions, the deviation is seldom above 2 ppb. This means the model is robust against these variances. Few exceptions are present at the border of the area of applicability (Sect. 3.2.1), e.g., in the Alps. But even in these regions, the deviation is well below 5 ppb. A discussion of implications for general subgrid variances can be found in Sect. 4.1.

3.3 The final ozone map

3.3.1 Production of the final map

All selected features listed in Sect. 3.1.1 are used to fit the final model. In contrast to the experiments in the previous sections, we train the model on 80 % of the AQ-Bench dataset and test it on the remaining 20 % of the independent test set. Figure 2 shows the predictions on the test set vs. the true values. The R^2 value of this model is 0.55 and the RMSE is 4.4 ppb. There is a spread around the 1 : 1 line; furthermore, extremes are not captured as well as values closer to the mean. True values of less than 20 ppb or more than 40 ppb are predicted with high bias, which is expected since random forests tend to predict extremes less accurately than values closer to the mean.

3.3.2 Visual analysis

The final map is shown in Fig. 9 (data available under https://doi.org/10.23728/b2share.a05f33b5527f408a99faeaeea033fcdc, Betancourt et al., 2021d). Predictions are in a range between 9.4 and 56.5 ppb. There are some characteristics that are visible at first sight, e.g., higher values in mountain areas, like in the western US. The global importance of "absolute latitude" shows through a latitudinal stratification and a clear north–south gradient in Europe, the US, and East Asia. Sometimes the borders of climatic zones are visible, like in the north of North America, and across Asia. This shows that even if the climatic zones are not important globally, they can be locally important. There are larger areas with low ozone variation in Greenland, Africa, and South America.

In Fig. 10, a detailed look at three selected areas is given, and the predictions are compared to the true values. In Fig. 10a, a uniform, low ozone concentration is predicted over the peninsula of Florida. Figure 10b shows low ozone values in the Po Valley, a densely populated plane. Towards the mountains which surround the valley, higher values are predicted, and for the higher mountains, no predictions can be made. Figure 10c shows the city of Tokyo, which is covered with ozone measurements and where ozone values are relatively low. At the coasts of Japan, the values are lower. The spatial ozone patterns described here can also be found in ozone maps generated by traditional chemical models such as the fusion products by DeLang et al. (2021). We discuss the prospects of global ozone mapping more thoroughly in Sect. 4.4.

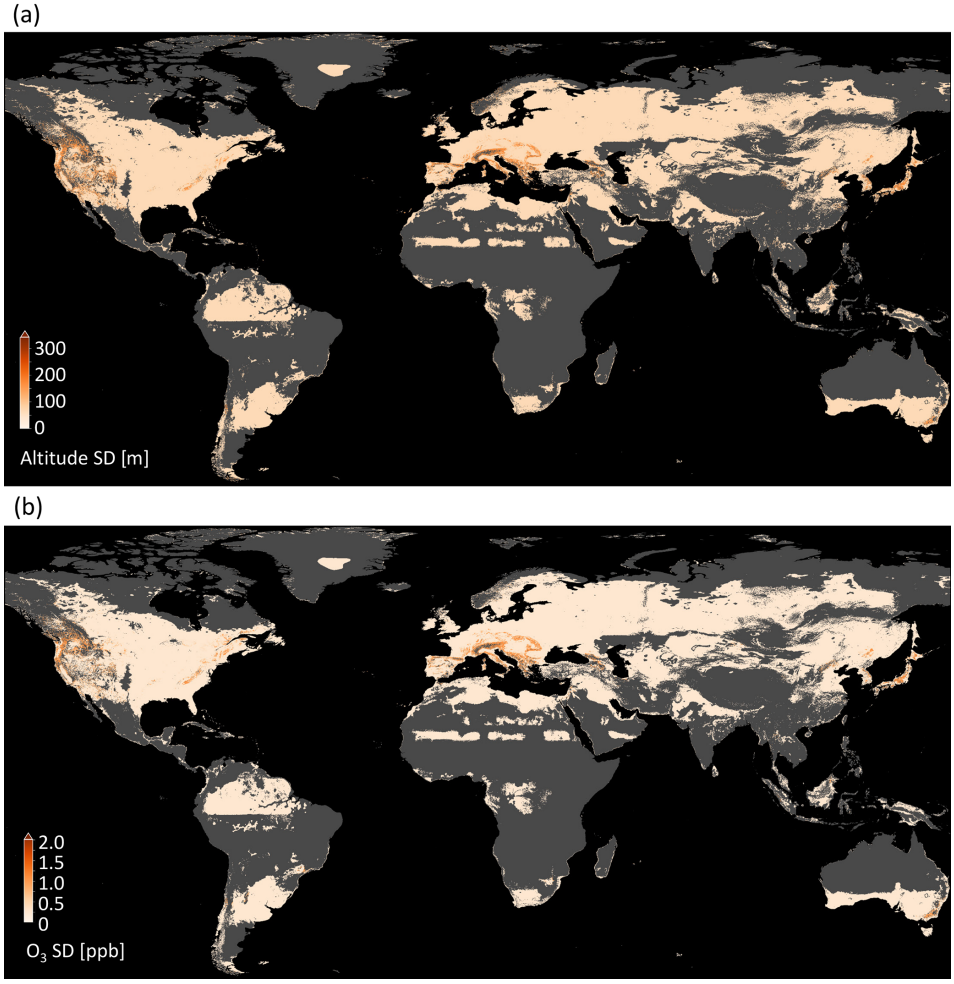


Figure 8. Results of propagating subgrid DEM variations through the model. **(a)** Spread of subgrid DEM data. **(b)** Spread of ozone values.

4 Discussion

4.1 Robustness

Based on Hamon et al. (2020), we define robustness as follows: *The model and map are considered robust if they do not change substantially under noise or perturbations that could realistically occur.* We define a 5 ppb change in RMSE score or predicted ozone values as significant (Schultz et al., 2017). Methods to assess the robustness are part of both the explainable machine learning workflow (Table 2) and the uncertainty assessments (Table 3). Regarding the robustness of the training process, the cross-validation results in Table 4 show that the model performance depended on the data split. This was already noted by Betancourt et al. (2021b) and is regarded as an inherent limitation of a noisy dataset.

We tested the robustness regarding typical variances in the ozone and geospatial data. The results from Sect. 3.2.2 and 3.2.3 show that the produced ozone map is robust against these fluctuations. The variances are never above the initial perturbations, and variances in the map do not exceed our limit of 5 ppb. Limits in the robustness were only shown through variances above 3 ppb at the borders of the area of applicability, and in regions with sparse training data (gray and dark turquoise areas in Figs. 7 and 8). This outcome shows that the issues of applicability (Sect. 4.3) and robustness are interconnected. In areas where the model is applicable, it is also more robust and uncertainties are lower.

In order to make the robustness assessment with respect to data feasible, we strongly reduced the dimensionality of our error model by using expert knowledge. We conducted two experiments where we modify training data and model inputs (Sect. 2.3.2 and 2.3.3). These experimental setups were chosen because they are expected to generalize well. The combined robustness experiments have shown that our produced maps are robust.

4.2 Scientific consistency

We discuss the scientific consistency of our model by assessing the results of the explainable machine learning work-

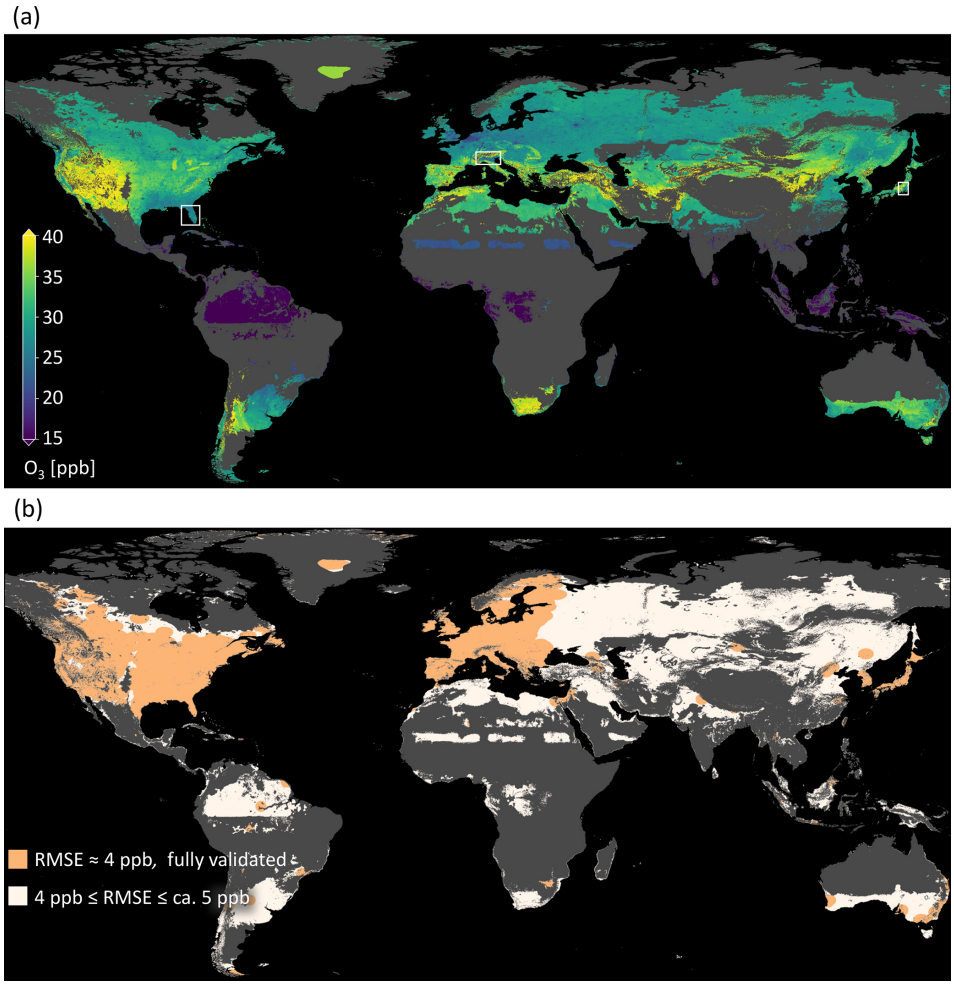


Figure 9. The final ozone map as produced in this study. Panel **(a)** shows the ozone values; **(b)** shows the uncertainty estimates. The areas shown in Fig. 10 are highlighted by white boxes.

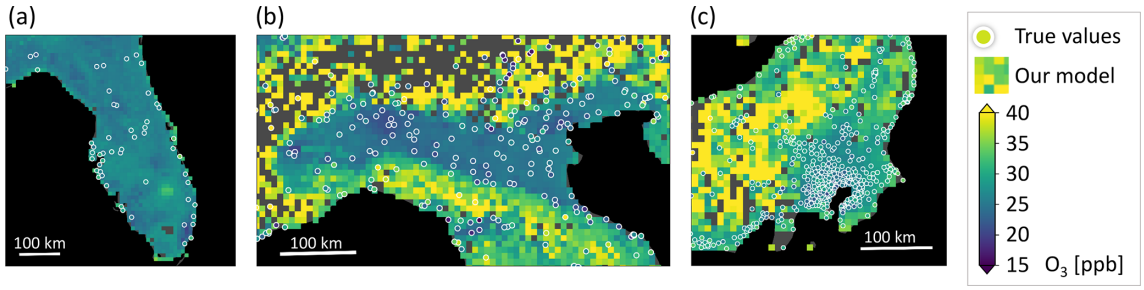


Figure 10. Map details with true values are given as white circles. **(a)** The Florida peninsula, USA. **(b)** The Po Valley in northern Italy. **(c)** Tokyo, Japan, and its surroundings.

flow (Table 2). We interpret the selected features, their importance, and their influence on the model predictions. The features are proxies to ozone processes, which makes it challenging to interpret the underlying chemical processes. Nevertheless, the connections between the features can be discussed, if they are plausible and consistent with respect to our understanding of ozone processes. This is a pure a posteriori approach, meaning we did not in any way enforce scientific consistency during the training process.

Regarding the global feature importance of SHAP (Fig. 5), it might be counterintuitive that the model focuses more on geographical features such as "absolute latitude" and "alti-

tude" than chemical factors such as the "NO_2 column", and "NO_x emissions". Geographic features are proxies for flow patterns and heat, not for ozone chemistry, which would be expected to be more important. This contradiction is due to the fact that the model provides an as-is view of ozone concentration and is not process oriented in any way. Many features such as "nightlight" and "population density" are correlated, so retraining the model might swap dependence in the SHAP values as noted by Lundberg et al. (2020).

The beeswarm plot in Fig. 5 shows the physical consistency of our model. The effect of "absolute latitude" on predictions is consistent with known ozone formation processes; i.e., ozone production generally increases when more sunlight is available. This is also evident in the latitudinally stratified ozone overview plots in global measurement-based studies such as TOAR health and TOAR vegetation (Fleming et al., 2018; Mills et al., 2018). Ozone is affected by meteorology (temperature, radiation) and precursor emissions (Sect. 1). The fact that there is no continuous increase of ozone towards tropical latitudes shows that the mapping model at least qualitatively captures the influence of low precursor emissions in the tropics. The importance of "absolute latitude" also indicates that the model can be improved by including temperature and radiation features from meteorological data. High "relative altitude" and "altitude" both increase the predicted ozone. These relations are consistent with Chevalier et al. (2007). There are relatively important chemistry-related features. We see that high values of "nightlight in 5 km area" reduce the predicted ozone. This is consistent with NO titration (Monks et al., 2015). Nightlights are a proxy for human activity, generally in the context of fossil fuel combustion, which leads to elevated NO_x concentrations. NO destroys ozone, and especially during the night time this leads to ozone levels close to zero ppb. High "forests in 25 km area" values lead to lower ozone predictions. This is plausible because there is little human activity in forested areas and thus no combustion-related precursor emissions occur. Quantification of either influence is not possible because, for example, it is unclear to what extent the different forests emit volatile organic compounds which are also ozone precursors. A city with "nightlight in 5 km area" equal to 50 cannot be directly quantified in terms of precursor emissions either. It is also not expected that the machine learning model learns the ozone-related processes described above because it is not process based. Instead, it learns the effects of processes if they are reflected in the training data.

The forward feature selection (Sects. 2.2.2 and 3.1.1) can also be discussed in terms of plausibility. Features selected by this method favor a generalizable model. Discarded features may help to characterize the locations, but their addition to the training data does not lead to a more generalizable model. "Urban and built-up in 25 km area" was not selected presumably because urban areas are often localized. This feature is therefore not as meaningful as the variables "nightlight" and "population density", which are also proxies for human activity, but are available at higher resolution. Similarly, the feature "cropland/natural vegetation mosaic in 25 km area" was discarded because ozone is affected differently by croplands and natural vegetation. Together with the large area considered, this feature becomes obsolete. We suspect the features "snow and ice in 25 km area", "barren or sparsely vegetated in 25 km area", and "wheat production" did not contribute to the model generalizability because they are simply not represented well in the training data. A feature may be an important proxy for ozone, but if the relationship is not expressed in the training data, it cannot be learned by a machine learning model. This feature can become more important if other training locations are included. This shows that the placing of measurement locations is crucial.

4.3 Mapping the global domain

The model has to generalize to unseen locations for global mapping. Two prerequisites are (1) the model must have seen the feature combination during training; (2) the connection between features and the target, ozone, must be the same. The two conditions are only fulfilled in a strictly constrained space, as shown in Fig. 6. We combined cross validation with an inspection of the feature space to ensure matching feature combinations. Then, based on the cross validation on different world regions, we point out regions with sparse or no training data, where higher model errors are expected (Sect. 3.2.1). We also conducted spatial cross validation with a shallow neural network (as in the baseline experiments of Betancourt et al., 2021b). The neural network had similar evaluation scores on the test set but did not generalize to other world regions, even showing negative R^2 values when evaluated in other world regions. We decided to discard the neural network architecture, because our main goal is global generalizability.

We can confidently map Europe, large parts of the US and East Asia, where the majority of the measurement stations are located. Those are industrialized countries in the northern hemisphere. The cross-validation results (Sect. 3.1.2), the area of applicability (Sect. 3.2.1), and expert knowledge confirm that uncertainties increase when a model trained on the AQ-Bench dataset is applied to other world regions. However, the cross validation in connection with the area of applicability technique shows that the model can be used in other world regions with acceptable uncertainties. That is promising for future global mapping approaches. One idea to solve these problems of different connections between features and ozone in different world regions is to train localized models, and apply them wherever possible. Localized models could not only yield more accurate predictions but in connection with SHAP values (Sect. 2.2.4), they could also rule out the governing factors of ozone in the respective regions and be easier to interpret.

With regard to the spatial domain, we can also discuss the resolution. The model was trained on point data of the "ab-

solute latitude", "altitude", and "relative altitude", and one could produce more fine-grained maps if the gridded data are present in higher resolution. However, one may need to reconsider some assumptions made here in terms of regional representativity of the measurements and the relation between geographic features and ozone on a different scale.

4.4 Prospects for ozone mapping

We mapped average tropospheric ozone from the stations in the AQ-Bench dataset to a global domain. For this, we fused different auxiliary geospatial datasets and gridded data with machine learning. We used features that are known proxies for ozone processes, and that were already proven to enable a prediction of ozone concentrations (Betancourt et al., 2021b). Our choice of data and algorithms is well justified and transparent. Errors did not exceed 5 ppb, which is an acceptable uncertainty. The R^2 value of the final model is 0.55, which is a good value for properly validated mapping. The maps produced show known patterns of ozone such as lower levels in metropolitan areas and higher levels in Mediterranean or mountainous regions. However, extremes (Fig. 2) are predicted with higher bias. This can be considered as a general problem of machine learning (Guth and Sapsis, 2019) but was also noted in other ozone modeling studies (Young et al., 2018).

For this first approach, we limited ourselves to the static mapping of aggregated mean ozone. An advantage of this approach is that the model result is directly the ozone metric of interest (in this case, average ozone). Since the AQ-Bench dataset contains other ozone metrics, they could be mapped as well. For example, vegetation- or health-related ozone metrics can be mapped with the same workflow as described here. Another advantage is that we used a multitude of inputs that could not be used in a traditional model because their connection to ozone is unknown. This means we exploit two benefits of machine learning: first, obtaining a bias-free estimate of the target directly, and second, using a multitude of inputs with unknown direct impact on the target.

Our model is only valid for the training data period (2010–2014), and it is not suitable to predict ozone values in other years. Our data product is a map that is aggregated in time. This could be a limitation as sometimes the data product of interest is a seasonal aggregate or even maps of daily or hourly air pollutant concentrations. The use of meteorological data as static or non-static inputs can be beneficial to further increase model performance and allow time-resolved mapping. We applied a completely data-driven approach, relying heavily on geospatial data. The other side of the spectrum is DeLang et al. (2021), who fused chemical transport model output to observations without exploiting the connection to other features. A possible direction to go from here is described by Irrgang et al. (2021), who propose the fusion of models and machine learning to benefit from both methods.

5 Conclusions

In this study, we developed a completely data-driven, machine-learning-based global mapping approach for tropospheric ozone. We mapped from the 5577 irregularly placed measurement stations of the AQ-Bench dataset (Betancourt et al., 2021b) to a regular $0.1° \times 0.1°$ grid. We used a multitude of geospatial datasets as input features. To our knowledge, this is the first completely data-driven approach to global ozone mapping. We combined this mapping with an end-to-end approach for explainable machine learning and uncertainty estimation. This allowed us to assess the robustness, scientific consistency, and global applicability of the model. We linked interpretation tools with domain knowledge to obtain application-specific explanations, which is in line with Roscher et al. (2020). The methods are interconnected; e.g., forward feature selection made the model easier to interpret. Likewise, the area of applicability was shown to match the model's robustness. We justified the choice of tools and detailed how they provided us with the results to make a comprehensive global analysis. The combination of explainable machine learning and uncertainty quantification makes the model and outputs trustworthy. Therefore, the map we produced provides information on global ozone distribution and is a transparent and reliable data product.

We explained the outcome and the model, which can lead to new scientific insights. Mapping studies like ours could also contribute to studies like Sofen et al. (2016) that propose locations for new air quality measurement sites to extend the observation network. Here, the inspection of the feature space helps to cover not only spatial world regions but also air quality regimes and areas with diverse geographic characteristics. Building locations can also be proposed based on their contribution to maximizing the area of applicability (Stadtler et al., 2022). The map as a data product can also be used to refine studies like TOAR (Fleming et al., 2018; Mills et al., 2018) because it enables analyzing locations with no measurement stations.

It would be beneficial to add time-resolved input features to the training data to improve evaluation scores and increase the temporal resolution of the map. Adding training data from regions like East Asia, or new data sources such as OpenAQ (https://openaq.org/, last access: 2 November 2021), would close the gaps in the global ozone map.

Appendix A: Technical details on the data

Table A1. Technical details on the data used in this work. For more information on the station location data, refer to Betancourt et al. (2021b). Please note that "land use in 25 km area" comprises all the different land cover features.

Variable	Data source and technical info	Reference
Ozone average values	Aggregated average ozone measurements of the stations in the AQ-Bench dataset from the years 2010–2014. The original data source is the database of the Tropospheric Ozone Assessment Report (TOAR).	Betancourt et al. (2021b), Schultz et al. (2017)
Climatic zone	12 classes of the IPCC 2006 classification scheme for default climate regions with a resolution of $5'$. Stations were attributed to the climatic zone in the respective grid cell. To prepare the gridded field, downscaling to 0.1° resolution was done by nearest-neighbor interpolation.	https://esdac.jrc.ec.europa.eu/projects/RenewableEnergy/ (last access: 23 March 2021)
Geographic location	The geographical location of the stations (longitude and latitude) was reported by the data providers and quality controlled by the TOAR database administrators. A gridded field of 0.1° resolution was generated within this study.	Schultz et al. (2017)
Altitude	The station altitude was reported by the data providers and quality controlled by the TOAR database administrators. The gridded field of 0.1° resolution was produced by linear 2-D interpolation of the ETOPO1 digital elevation model with an original resolution of $1'$.	Schultz et al. (2017), Amante and Eakins (2009)
Relative altitude	Derived at stations from the ETOPO1 digital elevation model and the station altitude. To generate a gridded field, the relative altitude was determined for every pixel from ETOPO1 data.	Amante and Eakins (2009)
Land cover in 25 km area	Derived from yearly land cover type L3 from the MODIS MD12C1 collection with an original resolution of 0.05°. The year 2012 and the International Geosphere-Biosphere Programme (IGBP) classification scheme with 17 classes were used. For the data at station locations, land cover data in the area of 25 km around each station was considered. Similarly, for the gridded fields, the 25 km area around each pixel was considered.	https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD12C1/ (last access: 23 March 2021)
Wheat/rice production	Annual wheat/rice production of the year 2000 according to the global agroecological zone data, version 3 with an original resolution of $5'$. The stations were attributed with data of the respective pixel. The gridded field of 0.1° was produced by linear 2-D interpolation.	https://www.fao.org/ (last access: 23 March 2021)
NO_x emissions	Annual NO_x emissions of the year 2010 from Emissions Database for Global Atmospheric Research – Hemispheric Transport of Air Pollution (EDGAR HTAP) inventory V2 with an original resolution of 0.1°. The stations were attributed with data of the respective pixel. The gridded field of 0.1° was produced by linear 2-D interpolation.	Janssens-Maenhout et al. (2015)
NO_2 full column	5-year average (2011–2015) tropospheric NO_2 column value from the Ozone Monitoring Instrument (OMI) on NASA AURA with an original resolution of 0.1°. The stations were attributed with data of the respective pixel.	Krotkov et al. (2016)
Population density	GPWv3 population density of the year 2010 with an original resolution of $2.5'$. For the data at station locations, data were aggregated in 1, 5, and 25 km around the station location. Similarly, for the gridded fields, data were aggregated in these radii around each pixel.	CIESIN (2005)
Nightlight	Stable nighttime lights of the year 2013 extracted from the NOAA Defense Meteorological Satellite Program (DMSP) product with an original resolution of 0.925 km. For the data at station locations, data were aggregated in 1, 5, and 25 km around the station location. Similarly, for the gridded fields, data were aggregated in these radii around each pixel.	https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html (last access: 23 March 2021)

Appendix B: Plots of gridded fields used as inputs for mapping model

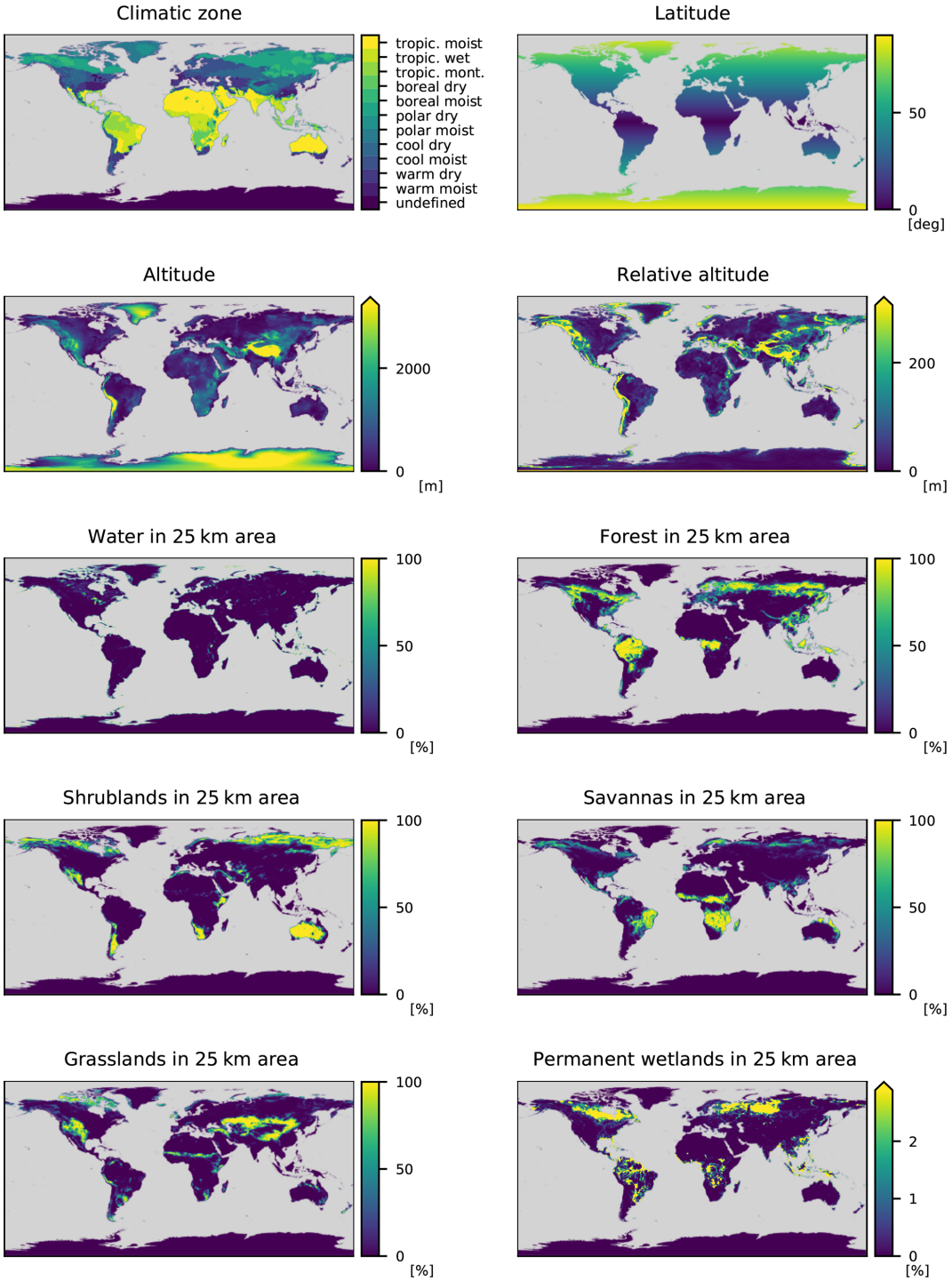


Figure B1. Gridded fields used for the final map production. Please note that the feature engineering was done as described in Sect. 2.2.2.

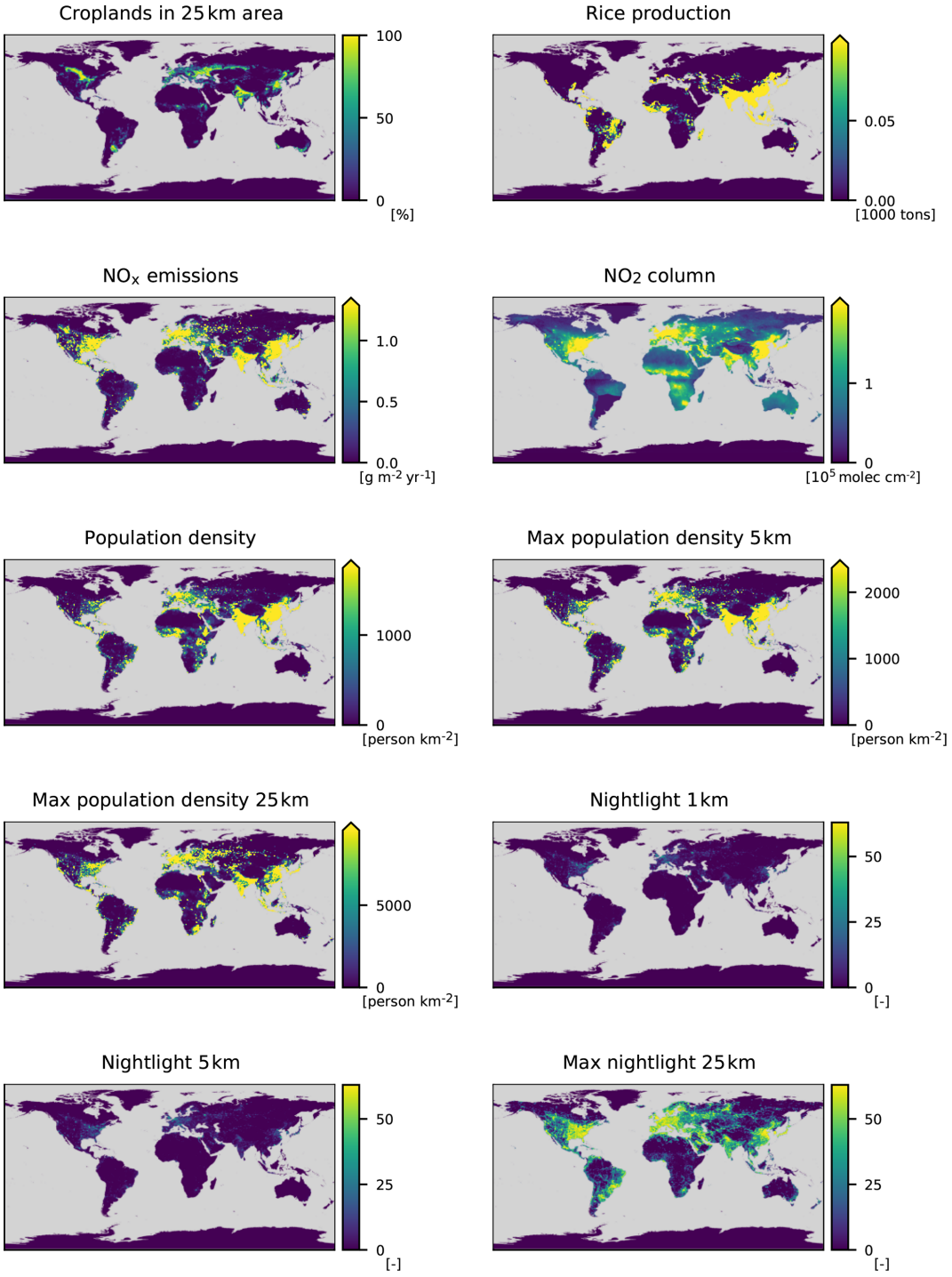


Figure B2. Gridded fields used for the final map production. Please note that the feature engineering was done as described in Sect. 2.2.2.

Appendix C: Example realization of error model

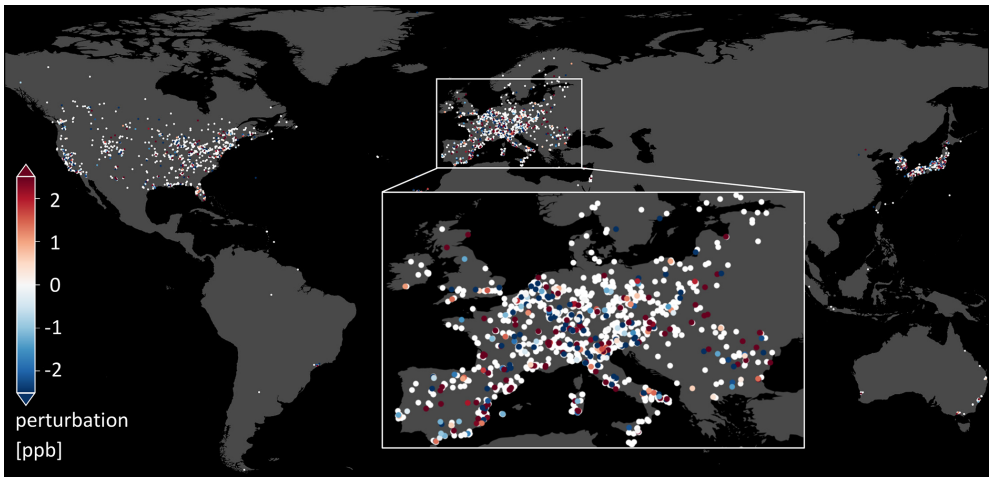


Figure C1. Example realization of the error model for ozone uncertainties as described in Sect. 2.3.2. A random subset of 25 % of the ozone values in the training set is perturbed with values sampled from a Gaussian distribution with 0 ppb mean and 5 ppb variance.

Appendix D: Convergence of the error model

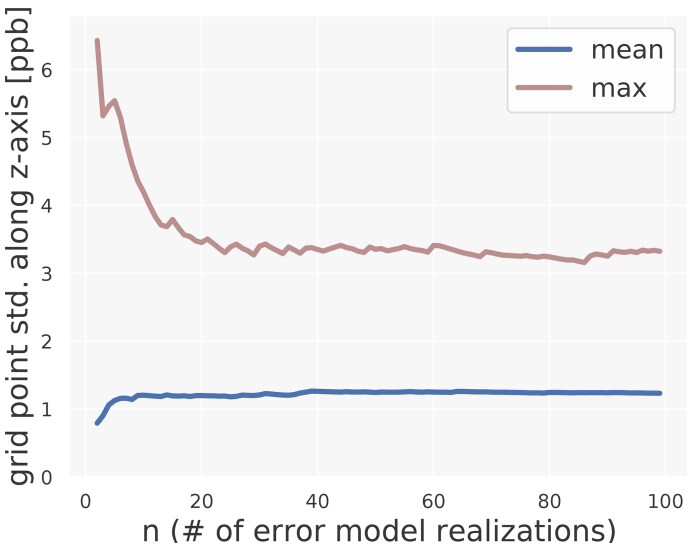


Figure D1. This plot justifies the use of 100 error model realizations in Sect. 3.2.2. We have stacked n perturbed maps along the z axis and monitored the grid point wise standard deviation along the z axis. The mean standard deviation over the whole map stabilizes after approximately 40 realizations. The maximum standard deviation exceeds 3.5 ppb for less than 20 realizations. This can be explained by the fact that some grid points base their predictions on single, differently perturbed stations when the number of realizations is low. This effect smoothes out after 20 realizations. Even though the maximum is not as stable as the mean (which is expected), convergence can be assumed after 100 realizations.

Appendix E: SHAP values of single predictions

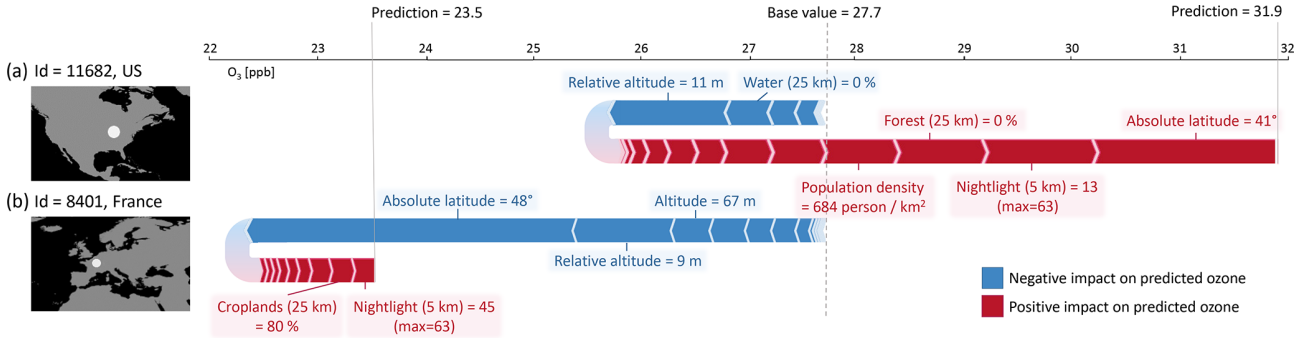


Figure E1. SHAP force plots for two example low-bias (< 1 ppb) predictions at **(a)** a rural station in the US and **(b)** an urban station in France, in addition to SHAP results from Sect. 3.1.3. Starting from the base value (27.7 ppb), a feature can increase or decrease the predicted ozone (red and blue arrows). The final predictions (23.5 and 31.9 ppb, respectively) result from adding all SHAP values to the base value. The most contributing features are labeled and their values are given. The high ozone station **(a)** is located in a rural area in the US with many agricultural fields and a smaller city nearby. The average ozone at this location is predicted to be high because the model uses the absence of forests, the low "nightlight in 5 km area" value, and the "absolute latitude" as features leading to high ozone values. This is consistent with Fig. 5, where it can be seen that a lower "absolute latitude" often increases the ozone value. The French station **(b)** is an urban background station surrounded by fields. The location is further in the north than the US station which leads to a strong decrease in the predicted ozone value. The low "(relative) altitude" further decreases the predicted ozone.

Code and data availability. The code which was used to generate the results published here is available under https://doi.org/10.34730/af084443e1c444feb12d83a93a65fa33 (Betancourt et al., 2022) under MIT License. The current version of the code is available under https://gitlab.jsc.fz-juelich.de/esde/machine-learning/ozone-mapping (last access: 13 December 2021) under MIT License. The AQ-Bench dataset (Betancourt et al., 2021b) is available under https://doi.org/10.23728/b2share.30d42b5a87344e82855a486bf2123e9f (Betancourt et al., 2020). The gridded data are available under https://doi.org/10.23728/b2share.9e88bc269c4f4dbc95b3c3b7f3e8512c (Betancourt et al., 2021c). The data products generated in this study, namely the ozone map and the area of applicability, are available under https://doi.org/10.23728/b2share.a05f33b5527f408a99faeaeea033fcdc (Betancourt et al., 2021d). All datasets are published under the CC-BY license.

Author contributions. All authors jointly developed the concept of the project under the leadership of CB and MGS. CB and ScS coordinated the project. MGS, RR, and JK supervised the project. CB, TTS, AE, AP, and ScS developed the code, conducted the experiments, and prepared the initial manuscript draft. MGS, RR, and JK reviewed and edited the manuscript. All authors read and approved the manuscript.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Earth System Science Data*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Parts of this research were presented in oral and display format at the conference "EGU General Assembly 2021" (Betancourt et al., 2021a).

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We are thankful to the TOAR community and several international agencies and institutions for making air quality and geospatial data available. We thank Hanna Meyer and Hu Zhao for helpful discussions. Clara Betancourt and Scarlet Stadtler acknowledge funding from the European Research Council, H2020 Research Infrastructures (IntelliAQ (grant no. ERC-2017-ADG#787576)). Timo T. Stomberg, Ann-Kathrin Edrich, Ankit Patnala, and Scarlet Stadtler acknowledge funding from the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety under grant no. 67KI2043 (KISTE). Ribana Roscher acknowledges funding by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant no. 01DD20001). The authors gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Krause, 2019) at the Jülich Supercomputing Centre (JSC). Open-access publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – grant no. 491111487. We thank the two anonymous reviewers for their suggestions to improve this work.

Financial support. This research has been supported by the European Research Council, H2020 European Research Council (IntelliAQ (grant no. 787576)), the Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (grant no. 67KI2043), the Initiative and Networking Fund of the Helmholtz Association (grant no. DB001549), Deutsche Forschungsgemeinschaft (grant no. 491111487), and the Bundesministerium für Bildung und Forschung (grant no. 01DD20001).

The article processing charges for this open-access publication were covered by the Forschungszentrum Jülich.

Review statement. This paper was edited by Fiona O'Connor and reviewed by two anonymous referees.

References

Amante, C. and Eakins, B. W.: ETOPO1 arc-minute global relief model: procedures, data sources and analysis, Tech. rep., NOAA National Geophysical Data Center, Boulder, Colorado, https://doi.org/10.7289/V5C8276M, 2009.

Bastin, J.-F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., Zohner, C. M., and Crowther, T. W.: The global tree restoration potential, Science, 365, 76–79, https://doi.org/10.1126/science.aax0848, 2019.

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

Betancourt, C., Stadtler, S., Stomberg, T., Edrich, A.-K., Patnala, A., Roscher, R., Kowalski, J., and Schultz, M. G.: Global fine resolution mapping of ozone metrics through explainable machine learning, EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-7596, https://doi.org/10.5194/egusphere-egu21-7596, 2021.

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

Betancourt, C., Edrich, A.-K., and Schultz, M. G.: Gridded data for the AQ-Bench dataset, B2SHARE [data set], https://doi.org/10.23728/b2share.9e88bc269c4f4dbc95b3c3b7f3e8512c, 2021c.

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., and Stadtler, S.: Global average ozone map 2010–2014, B2SHARE [data set], https://doi.org/10.23728/b2share.a05f33b5527f408a99faeaeea033fcdc, 2021d.

Betancourt, C., Stomberg, T., Edrich, A.-K., Patnala, A., and Stadtler, S.: Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties – Source Code, B2SHARE [code], https://doi.org/10.34730/af084443e1c444feb12d83a93a65fa33, 2022.

Blanke, S.: Hyperactive: An optimization and data collection toolbox for convenient and fast prototyping of computationally expensive models, v2.3.0, GitHub [code], https://github.com/SimonBlanke/Hyperactive, last access: 4 December 2021.

Brasseur, G., Orlando, J. J., and Tyndall, G. S. (Eds.): Atmospheric chemistry and global change, Oxford University Press, New York, US, 1st Edn., ISBN-10 0195105214, 1999.

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

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K., and Van Der Veen, A.: Mapping urban air pollution using GIS: a regression-based approach, Int. J. Geogr. Inf. Sci., 11, 699–718, https://doi.org/10.1080/136588197242158, 1997.

Chevalier, A., Gheusi, F., Delmas, R., Ordóñez, C., Sarrat, C., Zbinden, R., Thouret, V., Athier, G., and Cousin, J.-M.: Influence of altitude on ozone levels and variability in the lower troposphere: a ground-based study for western Europe over the period 2001–2004, Atmos. Chem. Phys., 7, 4311–4326, https://doi.org/10.5194/acp-7-4311-2007, 2007.

CIESIN: Gridded Population of the World, Version 3 (GPWv3): Population Count Grid, Center for International Earth Science Information Network – CIESIN – Columbia University, United Nations Food and Agriculture Programme – FAO, and Centro Internacional de Agricultura Tropical – CIAT, CIAT, Palisades, NY, NASA Socioeconomic Data and Applications Center (SEDAC), https://doi.org/10.7927/H4639MPP, 2005.

Cobourn, W. G., Dolcine, L., French, M., and Hubbard, M. C.: A Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting, J. Air. Waste Manage., 50, 1999–2009, https://doi.org/10.1080/10473289.2000.10464228, 2000.

Comrie, A. C.: Comparing Neural Networks and Regression Models for Ozone Forecasting, J. Air. Waste Manage., 47, 653–663, https://doi.org/10.1080/10473289.1997.10463925, 1997.

DeLang, M. N., Becker, J. S., Chang, K.-L., Serre, M. L., Cooper, O. R., Schultz, M. G., Schro¨der, S., Lu, X., Zhang, L., Deushi, M., Josse, B., Keller, C. A., Lamarque, J.-F., Lin, M., Liu, J., Marécal, V., Strode, S. A., Sudo, K., Tilmes, S., Zhang, L., Cleland, S. E., Collins, E. L., Brauer, M., and West, J. J.: Mapping Yearly Fine Resolution Global Surface Ozone through the Bayesian Maximum Entropy Data Fusion of Observations and Model Output for 1990–2017, Environ. Sci. Technol., 55, 4389–4398, https://doi.org/10.1021/acs.est.0c07742, 2021.

Duda, R. O., Hart, P. E., and Stork, D. G.: Pattern Classification, chap. 10, John Wiley & Sons, Inc., New York, US, 2nd Edn., ISBN-10 0471056693, 2001.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, in: KDD-96 Proceedings, Portland, OR, US, second International Conference on Knowledge Discovery and Data Mining (KDD), 2–4 August 1996, 34, 226–231, 1996.

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 (last access: 31 May 2022), 2008.

Fleming, Z. L., Doherty, R. M., Von Schneidemesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., Lefohn, A. S., Hamad, S., Moolla, R., Solberg, S., and Feng, Z.: Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health, 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.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X.: A Survey of Uncertainty in Deep Neural Networks, arXiv [preprint], arXiv:2107.03342v1, 2021.

Guth, S. and Sapsis, T. P.: Machine Learning Predictors of Extreme Events Occurring in Complex Dynamical Systems, Entropy, 21, 925, https://doi.org/10.3390/e21100925, 2019.

Hamon, R., Junklewitz, H., and Sanchez, I.: Robustness and explainability of artificial intelligence, Tech. Rep. JRC119336, Publications Office of the European Union, Luxembourg, Luxembourg, https://doi.org/10.2760/57493, 2020.

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., and Sanderman, J.: Machine learning in space and time for modelling soil organic carbon change, Eur. J. Soil Sci., 72, 1607–1623, https://doi.org/10.1111/ejss.12998, 2020.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D.: A review of land-use regression models to assess spatial variation of outdoor air pollution, Atmos. Environ., 42, 7561–7578, https://doi.org/10.1016/j.atmosenv.2008.05.057, 2008.

Hoogen, J. V. D., Geisen, S., Routh, D., Ferris, H., Traunspurger, W., Wardle, D. A., de Goede, R. G. M., Adams, B. J., Ahmad, W., Andriuzzi, W. S., Bardgett, R. D., Bonkowski, M., Campos-Herrera, R., Cares, J. E., Caruso, T., de Brito Caixeta, L., Chen, X., Costa, S. R., Creamer, R., Mauro da Cunha Castro, J., Dam, M., Djigal, D., Escuer, M., Griffiths, B. S., Gutiérrez, C., Hohberg, K., Kalinkina, D., Kardol, P., Kergunteuil, A., Korthals, G., Krashevska, V., Kudrin, A. A., Li, Q., Liang, W., Magilton, M., Marais, M., Martín, J. A. R., Matveeva, E., Mayad, E. H., Mulder, C., Mullin, P., Neilson, R., Nguyen, T. A. D., Nielsen, U. N., Okada, H., Rius, J. E. P., Pan, K., Peneva, V., Pellissier, L., Carlos Pereira da Silva, J., Pitteloud, C., Powers, T. O., Powers, K., Quist, C. W., Rasmann, S., Moreno, S. S., Scheu, S., Setälä, H., Sushchuk, A., Tiunov, A. V., Trap, J., van der Putten, W., Vestergård, M., Villenave, C., Waeyenberge, L., Wall, D. H., Wilschut, R., Wright, D. G., Yang, J.-I., and Crowther, T. W.: Soil nematode abundance and functional group composition at a global scale, Nature, 572, 194–198, https://doi.org/10.1038/s41586-019-1418-6, 2019.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science, Nat. Mach. Intell., 3, 667–674, https://doi.org/10.1038/s42256-021-00374-3, 2021.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M.: HTAP_v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmos. Chem. Phys., 15, 11411–11432, https://doi.org/10.5194/acp-15-11411-2015, 2015.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geosci. Model Dev., 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Keller, C. A., Evans, M. J., Kutz, J. N., and Pawson, S.: Machine learning and air quality modeling, in: Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), IEEE, Boston, MA, USA, 4570–4576, https://doi.org/10.1109/BigData.2017.8258500, 2017.

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, Geosci. Model Dev., 14, 1–25, https://doi.org/10.5194/gmd-14-1-2021, 2021.

Krause, D.: JUWELS: Modular Tier-0/1 Supercomputer at Jülich Supercomputing Centre, Journal of large-scale research facilities (JLSRF), 5, 1–8, https://doi.org/10.17815/jlsrf-5-171, 2019.

Krotkov, N. A., McLinden, C. A., Li, C., Lamsal, L. N., Celarier, E. A., Marchenko, S. V., Swartz, W. H., Bucsela, E. J., Joiner, J., Duncan, B. N., Boersma, K. F., Veefkind, J. P., Levelt, P. F., Fioletov, V. E., Dickerson, R. R., He, H., Lu, Z., and Streets, D. G.: Aura OMI observations of regional SO_2 and NO_2 pollution changes from 2005 to 2015, Atmos. Chem. Phys., 16, 4605–4629, https://doi.org/10.5194/acp-16-4605-2016, 2016.

Lary, D. J., Faruque, F. S., Malakar, N., Moore, A., Roscoe, B., Adams, Z. L., and Eggelston, Y.: Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$), Geospatial Health, 8, S611–S630, https://doi.org/10.4081/gh.2014.292, 2014.

Lee, K., Lee, H., Lee, K., and Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples, arXiv [preprint], arXiv:1711.09325, 2017.

Li, J., Siwabessy, J., Huang, Z., and Nichol, S.: Developing an Optimal Spatial Predictive Model for Seabed Sand Content Using Machine Learning, Geostatistics, and Their Hybrid Methods, Geosciences, 9, 4, https://doi.org/10.3390/geosciences9040180, 2019.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30 (NeurIPS 2017 proceedings), edited by: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., 4765–4774, Long Beach, CA, USA, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (last access: 31 May 2022), 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent individualized feature attribution for tree ensembles, arXiv [preprint], arXiv:1802.03888, 2018.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-

I.: From local explanations to global understanding with explainable AI for trees, Nature machine intelligence, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Mattson, M. D. and Godfrey, P. J.: Identification of road salt contamination using multiple regression and GIS, Environ. Manage., 18, 767–773, https://doi.org/10.1007/BF02394639, 1994.

Meyer, H.: Machine learning as a tool to "map the world"? On remote sensing and predictive modelling for environmental monitoring, 17th Biodiversity Exploratories Assembly, Wernigerode, Germany [keynote], 4 March 2020.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods Ecol. Evol., 12, 1620–1633, https://doi.org/10.1111/2041-210X.13650, 2021.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environ. Modell. Softw., 101, 1–9, https://doi.org/10.1016/j.envsoft.2017.12.001, 2018.

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, Atmos. Chem. Phys., 15, 8889–8973, https://doi.org/10.5194/acp-15-8889-2015, 2015.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4, 1–22, https://doi.org/10.5194/soil-4-1-2018, 2018.

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, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Petermann, E., Meyer, H., Nussbaum, M., and Bossew, P.: Mapping the geogenic radon potential for Germany by machine learning, Sci. Total Environ., 754, 142291, https://doi.org/10.1016/j.scitotenv.2020.142291, 2021.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models, Nat. Commun., 11, 1–11, https://doi.org/10.1038/s41467-020-18321-y, 2020.

Ren, X., Mi, Z., and Georgopoulos, P. G.: Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States, Environ. Int., 142, 105827, https://doi.org/10.1016/j.envint.2020.105827, 2020.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J.: Explainable Machine Learning for Scientific Insights and Discoveries, IEEE Access, 8, 42200–42216, https://doi.org/10.1109/ACCESS.2020.2976199, 2020.

Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., Lee, J.-B., Park, H.-J., and Choi, M.-H.: A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance, Sci. Rep., 11, 1–8, https://doi.org/10.1038/s41598-021-90446-6, 2021.

Schmitz, S., Towers, S., Villena, G., Caseiro, A., Wegener, R., Klemp, D., Langer, I., Meier, F., and von Schneidemesser, E.: Unravelling a black box: an open-source methodology for the field calibration of small air quality sensors, Atmos. Meas. Tech., 14, 7221–7241, https://doi.org/10.5194/amt-14-7221-2021, 2021.

Schultz, M. G., Akimoto, H., Bottenheim, J., Buchmann, B., Galbally, I. E., Gilge, S., Helmig, D., Koide, H., Lewis, A. C., Novelli, P. C., Plass-Dülmer, C., Ryerson, T. B., Steinbacher, M., Steinbrecher, R., Tarasova, O., Tørseth, K., Thouret, V., and Zellweger, C.: The Global Atmosphere Watch reactive gases measurement network, Elem. Sci. Anth., 3, 000067, https://doi.org/10.12952/journal.elementa.000067, 2015.

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

Shapley, L.: A Value for n-Person Games, vol. II of Contributions to the Theory of Games, Princeton University Press, Princeton, UK, chap. 17, 307–318, https://doi.org/10.1515/9781400881970-018, 1953.

Sofen, E. D., Bowdalo, D., and Evans, M. J.: How to most effectively expand the global surface ozone observing network, Atmos. Chem. Phys., 16, 1445–1457, https://doi.org/10.5194/acp-16-1445-2016, 2016.

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

Wallace, J. and Hobbs, P.: Atmospheric Science: An Introductory Survey, vol. 92 of International Geophysics Series, Elsevier Academic Press, Burlington, MA, USA, 2nd Edn., https://doi.org/10.1016/C2009-0-00034-8, 2006.

Wang, S., Ma, Y., Wang, Z., Wang, L., Chi, X., Ding, A., Yao, M., Li, Y., Li, Q., Wu, M., Zhang, L., Xiao, Y., and Zhang, Y.: Mobile monitoring of urban air quality at high spatial resolution by low-cost sensors: impacts of COVID-19 pandemic lockdown, Atmos. Chem. Phys., 21, 7199–7215, https://doi.org/10.5194/acp-21-7199-2021, 2021.

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, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

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.

D.4 Fourth paper (Betancourt et al., 2023)

Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data

Clara Betancourt, Cathy W. Y. Li, Felix Kleinert, and Martin G. Schultz

Journal: Environmental Science and Technology
Status: published (September 2023)
DOI: 10.1021/acs.est.3c05104

Environmental Science & Technology

pubs.acs.org/est

Open Access

This article is licensed under CC-BY 4.0

Article

Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data

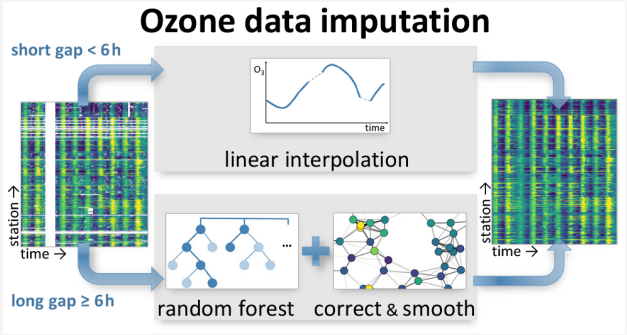
Clara Betancourt, Cathy W. Y. Li, Felix Kleinert, and Martin G. Schultz*

Cite This: *Environ. Sci. Technol.* 2023, 57, 18246−18258

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Gaps in the measurement series of atmospheric pollutants can impede the reliable assessment of their impacts and trends. We propose a new method for missing data imputation of the air pollutant tropospheric ozone by using the graph machine learning algorithm "correct and smooth". This algorithm uses auxiliary data that characterize the measurement location and, in addition, ozone observations at neighboring sites to improve the imputations of simple statistical and machine learning models. We apply our method to data from 278 stations of the year 2011 of the German Environment Agency (Umweltbundesamt − UBA) monitoring network. The preliminary version of these data exhibits three gap patterns: shorter gaps in the range of hours, longer gaps of up to several months in length, and gaps occurring at multiple stations at once. For short gaps of up to 5 h, linear interpolation is most accurate. Longer gaps at single stations are most effectively imputed by a random forest in connection with the correct and smooth. For longer gaps at multiple stations, the correct and smooth algorithm improved the random forest despite a lack of data in the neighborhood of the missing values. We therefore suggest a hybrid of linear interpolation and graph machine learning for the imputation of tropospheric ozone time series.



KEYWORDS: *graph signal processing, graph machine learning, missing data imputation, air quality, tropospheric ozone*

Downloaded via 87.79.82.246 on November 30, 2023 at 08:49:03 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Tropospheric ozone is a toxic air pollutant and a short-lived climate forcer.[1,2] While stratospheric ozone protects life on earth from harmful ultraviolet radiation, tropospheric ozone is a health hazard[3−5] and a substantial threat to global food security through the destruction of crops.[6−8] Its surplus radiative forcing is estimated to be 0.39 W m^{-2}, which is about a quarter of the radiative forcing of carbon dioxide.[9,10] As a secondary air pollutant, ozone is formed by a cascade of (photo-)chemical processes in the atmosphere, which include precursors such as nitrogen oxides (NO_x) and volatile organic compounds (VOCs).[11,12] The interplay of chemistry, transport, and deposition induces daily and seasonal cycles in the distribution of ozone concentrations, which are superimposed with variances from all spatiotemporal scales.[1,12,13] As such, ozone concentrations can change substantially in a matter of hours and on spatial scales of kilometers. It is therefore difficult to quantify the regional distribution of tropospheric ozone, and a fine-resolution monitoring network is required to obtain reasonably precise estimates of this distribution.[14,15] The measurements, which are typically reported as hourly averages, are used to determine whether thresholds of ozone statistics (or ozone "metrics") are exceeded[13,14,16] and hence to assess the impacts of ozone at various times and locations.

Like any air quality monitoring time series, ozone measurements suffer from missing data. These can occur due to sensor malfunctioning, calibration procedures, issues with data transfer, or the stations going out of operation. Missing data reduces the robustness of statistical analyses.[13,17] For example, if an ozone metric counts concentration threshold exceedances on a yearly basis and a sensor fails on a day with an exceedance, then a yearly statistic can be corrupted. Missing data also impede the usefulness of such data in other contexts. For example, machine learning models to forecast ozone concentrations[18−21] require a gap-free time series as input to make predictions. It is therefore necessary to impute the gaps in the ozone concentrations.

Ozone metrics for air quality assessments are usually aggregated hourly measurements of longer time periods, e.g., one year. Often, missing data within the aggregation period are compensated by imputing the average concentration over this period for each missing value.[13,22] This approach is often

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: June 30, 2023
Revised: August 24, 2023
Accepted: August 24, 2023
Published: September 4, 2023

ACS Publications

© 2023 The Authors. Published by American Chemical Society

applied implicitly when an ozone metric is calculated on the basis of the available fraction of the data set. In order to ensure a certain level of robustness of the metric, this simple imputation method is generally only applied to time series with a maximum fraction of missing values of 25% or less.[14,22] More advanced missing data imputation techniques for missing air pollutant data were developed during the past years.[23,24] Univariate interpolation methods, e.g., linear interpolation and spline interpolation, depend on the available data at time steps before and after a gap and are therefore suitable for shorter gaps in the range of hours. In contrast, multivariate methods, which include linear regression and machine learning algorithms such as neural network and random forest, make use of auxiliary data or covariates such as meteorological data and are therefore suitable for longer gaps. The imputation performance depends not only on the amount of missing data but also on the manner in which data are missing, i.e., the "missingness". Missing data patterns can be classified into three types according to the dependency of the missingness on the variable of interest and the auxiliary data:[25−27] (1) missing completely at random (MCAR), where the data missingness is independent of the variable of interest and any other external influences; (2) missing at random (MAR), where the data missingness is independent of the variable of interest but the missing data pattern can be related to auxiliary data; and (3) not missing at random (NMAR), where the data missingness depends on the variable of interest. The missing data patterns in air quality monitoring are generally MCAR or MAR.[23,25,28] In that case, the reasons why the data are missing can be ignored in the analysis of the data, and hence the methods used for missing data imputation can be simplified.[27]

Univariate and multivariate methods, or combinations of them, were successfully applied for missing air quality data imputation.[23,25,29,30] However, even sophisticated machine learning methods fail to efficiently utilize available data at monitoring stations in the neighborhood of a missing measurement. Challenges in using these data arise because stations are irregularly placed and neighboring measurements may not be available for all time steps. One simple approach to include neighboring data to predict or impute air quality data is to consider spatial distances or correlations between the stations.[31−33] A more advanced solution to this is graph machine learning,[34,35] a subfield of graph signal processing[36,37] which allows machine learning on irregularly structured data such as a monitoring network. Graph-based methods have been adopted for air quality-related tasks, such as outlier detection, postprocessing of low-cost sensor data, or high-resolution forecasting.[38−44] Graph machine learning was shown to be suitable for the imputation of different data sets,[45−47] yet, to the best of our knowledge, they have not yet been used to impute missing air quality data.

In this study, we develop a strategy to use graph machine learning to improve the imputations achieved by other existing methods. As a case study, we use a data set of hourly observations from 278 stations of the German Environment Agency (Umweltbundesamt − UBA) air quality monitoring network in the year 2011. Figure 1 shows the station locations and their relations in the graph that is built according to the procedures described in the next section. We combine the available observations with geospatial metadata, meteorological, and reanalysis data to allow the different regression and machine learning approaches to exploit relationships between these data and the measured ozone time series. For the analysis

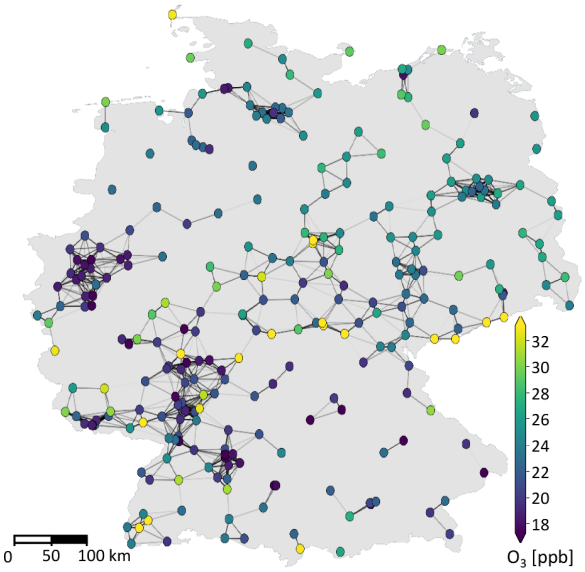


Figure 1. Illustration of the graph structure defined on the stations of the UBA monitoring network. The stations are nodes in the graph. Nodes of 50 km distance or less are connected by edges, which allow a graph machine learning algorithm to pass messages between them. In this figure, the nodes are labeled with the average ozone concentration over 2011, omitting temporal variances for clarity.

of the performance of these approaches, we identify three types of gaps that frequently occur: (1) shorter isolated gaps in the range of hours, e.g., when an instrument is offline for 1 h during calibration; (2) longer gaps in the range of months including multiple daily cycles and even changes in seasons; (3) gaps occurring at all stations of the network at the same time. The assessment of the three types of gaps suggests optimal imputation strategies for each gap type. We compare the performance of our method with published baseline statistical, numerical, and machine learning methods. Besides the code and input data, we also provide the final imputed version of the data set.

DATA AND METHODS

Ozone Data. Ozone data used in this study are from the German Environment Agency (Umweltbundesamt − UBA). The UBA collects and provides air quality data for Germany. We extracted the data from the Tropospheric Ozone Assessment Report (TOAR) database[13] at the Jülich Supercomputing Centre. The TOAR database receives a copy of all German ozone data from the UBA in near-real time. We selected hourly data from 278 stations across Germany in 2011 because there was an exceptionally large number of missing values in these data. It should be noted that UBA itself provides a final validated data set of ozone concentrations in the following year, which has fewer gaps than the data we have worked with. However, to develop our method and demonstrate its potential, we have chosen the preliminary data set with more frequent and larger gaps, and we use the final validated data set to crosscheck our results.

The selected data set contains over 2.4 million data points for training, testing, and evaluating the different imputation methods. The location of the stations and their mean ozone concentrations are shown in Figure 1. Compared to the theoretically available maximum number of hourly values, 15% of the data are missing. The missing data are generally

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

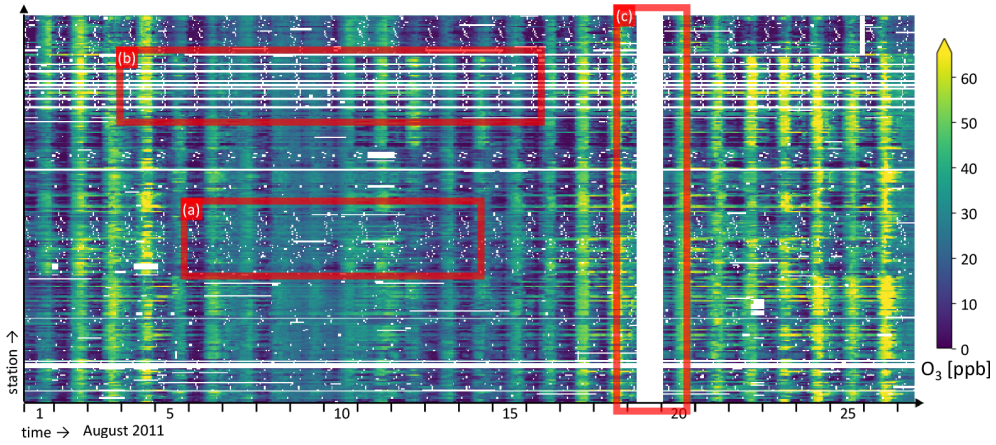


Figure 2. Measured ozone concentrations of August 1−26, 2011, at the selected UBA stations. Examples of three cases are marked: (a) short isolated gaps, (b) longer isolated gaps, and (c) gaps occurring at all stations simultaneously.

completely random (MCAR), except for a few cases where sensors are offline during the night and thus missing randomly (MAR). Seventeen % of data gaps occur at single stations during short periods of up to 5 h length. 57% occur as longer periods at single stations, and 26% of the data gaps occur at all stations simultaneously. This last category contains several short gaps of 3−4 h and three longer gaps with 18−43 h length starting from August 19, October 8, and December 20, 2011, respectively. The latter gaps could be traced back to data transmission gaps between the UBA and the TOAR database and are not part of the original UBA data set. Figure 2 shows an excerpt of these data, including examples of the three missing data patterns. Section S1 of the Supporting Information contains the summary statistics of the data. A detailed overview of gap lengths is given in section S2.

Auxiliary Data. We selected the following auxiliary data as features for multivariate imputation because they have been shown suitable to predict ozone in previous machine learning studies:[18,20,48]

- Datetime features: hour of the day, day of the week, and day of the year;
- Meteorological data: temperature, relative humidity, cloud cover, planetary boundary layer height, and wind components u and v;
- Atmospheric composition reanalysis data: concentrations of ozone (O_3), nitrogen monoxide (NO), and nitrogen dioxide (NO_2);
- Emission data: nitrogen oxides (NO_x); and
- Static station metadata: altitude, relative altitude, population density, nightlight intensity, station type, and type of area.

Meteorological data were extracted from the 6 km hourly resolution COSMO reanalysis[49] (COSMO-REA6). Atmospheric composition reanalysis data were extracted from the surface level of the ECMWF Atmospheric Composition Reanalysis 4 (EAC4) data set,[50] which combines observations with model data from a global chemical transport model (CTM). The emission data were extracted from the CAMS Global anthropogenic emissions (version 5.3)[51] with monthly resolution. Station metadata are taken from the TOAR database.[13] Here the relative altitude is the difference in elevation to the lowest point 5 km around the station.

Missing Data Imputation with Mean Values. As a statistical baseline method (B), we impute the spatiotemporal mean (stm) over all available data to all gaps:

$$\hat{y}_{B,stm} = \langle y \rangle \quad \text{with} \quad \langle y \rangle = \frac{1}{N} \sum_{x_i, t_j} y(x_i, t_j) \tag{1}$$

$\hat{y}$ denotes an imputed value, y is a measurement, N is the total number of available measurements, x_i is a station with index i, and t_j is a time step with index j. As a variant of this method, we impute the time-dependent spatial mean (sm), which is the mean over all available measurements at a specific time step:

$$\hat{y}_{B,sm}(t_j) = \langle y(t_j) \rangle \quad \text{with} \quad \langle y(t_j) \rangle = \frac{1}{N(t_j)} \sum_{x_i} y(x_i, t_j) \tag{2}$$

If no data are available for a given time step (which is the case for 352 of 8760 time steps), we impute the mean ozone concentration from EAC4 of that time step. This method captures daily, weekly, and seasonal cycles inherent in the available data without regard for extra station information or meteorology.

Missing Data Imputation with the Nearest-Neighbor Hybrid Method. A second statistical baseline method is the hybrid of linear interpolation (lin) for short gaps and multivariate nearest-neighbor (nn) interpolation for longer gaps. This method has been shown to be effective for missing data imputation of air pollution data.[23] The authors called it the nearest-neighbor hybrid (nnh), and we adopt their naming convention. According to this method, shorter gaps with a length L shorter than a threshold length L_t are imputed by fitting a straight line between the two end points t_1 and t_2 of the gap and calculating the missing values for any time $t_1 < t_j < t_2$ from this line equation. L_t is a tunable hyperparameter and varies according to the air pollutant in question. Longer gaps are imputed by multivariate nearest-neighbor interpolation as follows: The auxiliary data (features) of a data point are considered to be points $\vec{f}$ in the multidimensional feature space. In this study, we use 19 features, so this space has 19 dimensions. For every missing ozone value with index k, the nearest-neighbor sample with index k' with an available ozone measurement is searched in the feature space. Thereby, all features are standardized to zero mean and unit variance, so features covering different scales are treated with equal

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

importance. The distance measure is the Euclidean distance. Thus, in effect, the imputed ozone value of a missing data point is calculated according to eqs 3−5:

$$\hat{y}_{B,lin}(x_i, t_j) = y(x_i, t_1) + \frac{t_j - t_1}{t_2 - t_1}(y(x_i, t_2) - y(x_i, t_1)) \tag{3}$$

$$\hat{y}_{B,nn}(\vec{f}_k) = y(\vec{f}_{k'})|_{d_{k,k'}=d_{\min}} \quad \text{with} \quad d_{k,k'} = |(\vec{f}_k - \vec{f}_{k'})| \tag{4}$$

$$\hat{y}(x_i, t_j, \vec{f}_k)_{B,nnh} = \begin{cases} \hat{y}(x_i, t_j)_{B,lin} & \text{if } L < L_t \\ \hat{y}(\vec{f}_k)_{B,nn} & \text{if } L \geq L_t \end{cases} \tag{5}$$

An arrow ($\vec{\cdot}$) above a variable denotes a vector in this and all following equations.

Missing Data Imputation with Atmospheric Reanalysis. Imputation with ozone values from atmospheric reanalyses (here EAC4) is another baseline method against which machine learning models can be compared. To obtain the EAC4 reanalysis, observations from multiple satellites were assimilated with ECMWF's Integrated Forecasting System[50] (IFS). The model's prior estimates are optimized through minimizing the cost function, which measures the difference between modeled and observed fields to produce an improved estimate over the reanalysis period. Without the time constraint of issuing timely forecasts, the quality of reanalysis products benefits from the improvement of the quality and availability of observations. The EAC4 data are available in gridded format with 80 km spatial resolution and 3 h temporal resolution. We impute the ozone concentration from the EAC4 data set of the nearest-neighbor grid cell to all gaps:

$$\hat{y}_{B,EAC4}(x_i, t_j) = y(x_i, t_j)_{EAC4} \tag{6}$$

We point out that the imputation of measurements with gridded data is not ideal due to the representation mismatch of points and grid boxes. Furthermore, this method is prone to model biases that cannot be completely removed with statistical bias correction methods.

Random Forest for Missing Data Imputation. Random forest is a tree-based machine learning algorithm developed by Breiman in 2001.[52] Tree-based models were proven to excel in particular on tabular style data like the auxiliary data of this study.[53] A random forest is an ensemble of decision trees for classification or (in our case) regression. Decision trees iteratively partition the training data by finding logical rules associated with the input features to minimize a cost function such as squared loss. Individual decision trees have a low bias but are prone to overfitting. A random forest improves this problem through the resampling of the available training data. It is obtained by fitting many, usually several hundred, decision trees on bootstrapped training data sets. We chose random forest because in preliminary experiments it outperformed other machine learning models. In particular, gradient boosted trees[54] performed slightly worse than random forest on our data set, presumably because they are more prone to overfitting on noisy data with many variables such as ours. We rejected linear models because they failed to capture the ozone cycles and nonlinear relationships with the input features in our preliminary experiments.

In this study, a random forest (*rf*) is fitted on the features as inputs and the available measurements as output. The features are the auxiliary data introduced earlier. This random forest predicts an estimate of ozone concentration for every missing ozone measurement, based on the features of that data point:

$$\hat{y}_{B,rf}(\vec{f}_k) = rf(\vec{f}_k) \tag{7}$$

Defining a Graph Structure on an Air Quality Monitoring Network. Graphs are a "general language for describing and analyzing entities with relations or interactions".[55] Machine learning on graphs has gained success in the past years because it can solve complex tasks on data of irregular structure, such as protein folding, traffic prediction, or action recognition in computer vision.[56−58] From a graph theoretical perspective, the task in this study is to provide labels for unlabeled nodes (in our case, data points with missing ozone values).

We define the graph structure of our data in the following way: Each data point at station x and time step t is a node; therefore, there are ca. 2.4 million nodes in total. If there is a measurement y available for that data point, then the node is labeled with that measurement. If not, it is unlabeled. So in our case, 15% of the nodes are unlabeled. Every node has features $\vec{f}$, namely, the 19 auxiliary data values described above. An edge exists between the nodes k and k' if two conditions are fulfilled: first, they are 50 km or closer in spatial distance, and second, the time difference between them is 6 h or less. We chose these thresholds because the areas of influence of two measuring stations overlap at a distance of 50 km or less[14] and because ozone varies on hourly scales. The edge allows node k' to receive information from node k. The total number of edges obtained in this way is about 240 million, so each node receives information from about 100 nodes on average. The edges are weighted according to the spatial and temporal distances Δx, Δt:

$$w_{k\to k'} = \frac{50\text{ km} - |\Delta x_{i,i'}|}{50\text{ km}} \cdot \frac{6\text{ h} - |\Delta t_{j,j'}|}{6\text{ h}} \tag{8}$$

Figure 1 illustrates the graph, omitting the time component and self-loops for clarity. An isolated node in this figure has neighbors only in the temporal domain, so message passing will only be possible along the temporal axis. For more information on graph theory, the reader is referred, for example, to the book by Hamilton.[34]

Graph Machine Learning To Improve Missing Data Imputation. Graphs are routinely used in semisupervised missing data imputation, where information from both labeled and unlabeled data are used.[59−61] In particular, the correct and smooth algorithm by Huang et al. has proven effective in such tasks.[60] Correct and smooth is a graph machine learning method, since there is an iterative improvement of predictions based on message passing within the graph. It is about 100 times faster to fit than a graph neural network.[58,60]

We apply this algorithm to improve the baseline imputation methods described above. As the algorithm was originally designed to output class probabilities in semisupervised classification tasks, we had to make minor adjustments to apply it to the imputation of ozone concentrations, which is a regression task. The original method predicts a label score for every class and then converts all label scores to class probabilities by applying a softmax function. We modified the method to have only one output as we impute only one variable (ozone). We also removed the softmax function, which is unnecessary for regression problems. To the best of

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

our knowledge, this is the first study in which the correct and smooth algorithm is used for a regression task rather than a classification task.

The correct and smooth algorithm is applied in three steps. In the first step ("estimate"), a base model B estimates the node labels $\hat{y}$ based on the node features without making use of the graph structure. In this study, the base model is any of the statistical or machine learning methods described above:

$$\hat{y}_{estimate} = B(x_i, t_j, \vec{f}_k) \tag{9}$$

In the second step ("correct"), the errors e_k^0 of this base model are calculated by comparing the predicted labels to the true labels wherever possible. These errors are then propagated iteratively L_1 times to the unlabeled nodes, and the resulting error correction is added to the base prediction. This step, also called "residual propagation", assumes that if nodes k and k' are connected by an edge, their errors e_k and e_k' of the base model are correlated.

$$e_k^0 = \begin{cases} y_k - \hat{y}_{B,k} & \text{if } k \text{ is a training node} \\ 0 & \text{else} \end{cases} \tag{10}$$

$$\vec{e}^{\,l} = \alpha_1 \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\vec{e}^{\,l-1} + (1-\alpha_1)\vec{e}^{\,l-1} \tag{11}$$

$$\hat{\vec{y}}_{correct} = \hat{\vec{y}}_{estimate} + \gamma \vec{e}^{\,L_1} \tag{12}$$

Here, $\mathbf{A}$ with $A_{k,k'} = w_{k\to k'}$ is the adjacency matrix of the graph that contains the scaled edge weights as entries. $\mathbf{D}$ is the degree matrix of the graph that contains the node degrees as diagonal entries; therefore, $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is the normalized adjacency matrix. The index l denotes an iteration step between 0 and L_1. L_1, α_1 and γ are tunable hyperparameters.

The third step ("smooth") is similar to the second step, but here the labels y and $\hat{y}_{correct}$ are propagated because it is assumed that neighboring nodes have similar labels. This assumption is valid because ozone concentrations are correlated in space and time.[62]

$$\hat{y}_k^0 = \begin{cases} y_i & \text{if } k \text{ is a training node} \\ \hat{y}_{i,correct} & \text{else} \end{cases} \tag{13}$$

$$\hat{\vec{y}}^{\,l} = \alpha_2 \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\hat{\vec{y}}^{\,l-1} + (1-\alpha_2)\hat{\vec{y}}^{\,l-1} \tag{14}$$

$$\hat{\vec{y}}_{smooth} = \vec{y}^{\,L_2} \tag{15}$$

The smoothing step therefore resembles a graph filter.[63] L_2 and α_2 are tunable hyperparameters.

Evaluation. To evaluate the different imputation methods, we must artificially mask a share of the labeled data points as missing and compare the imputed ozone concentrations to the originally reported values. In machine learning, it is common to reserve a large share of labeled data for fitting the models ("training set") and smaller shares to tune hyperparameters ("validation set") and to test the final model performance ("test set"). Therefore, we split the data as follows: 70% of the data are used as is for training. Fifteen % of the data are masked (i.e., labels are removed), and of these, half are assigned to the validation set and half to the test set. The remaining 15% of the data are unlabeled samples. These are the missing data samples described above. To realistically test the predictive performance of the different algorithms, we maintained the gap characteristics of the missing data in the masking of the validation and test sets. For every gap length found at single stations, we mask counterparts of equal length randomly in the validation and test sets. Similarly, we mask counterparts of the gaps occurring at multiple stations. See section S2 for a detailed list of gaps masked for validation and test purposes.

We used three evaluation metrics that are commonly used for missing data imputation. The coefficient of determination R^2 is unitless and measures the proportion of variance in the true values that is explained by the model. A larger R^2 denotes a better model, and the largest possible value is 1.

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{N}(y_k - \langle y\rangle)^2} \quad \text{with} \quad \langle y\rangle = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{16}$$

We also evaluate the root-mean-square error (RMSE) in ppb:

$$\text{RMSE} = \sqrt{\sum_{k=1}^{N}\frac{(y_k - \hat{y}_k)^2}{N}} \tag{17}$$

Obviously, perfect agreement would yield an RMSE of zero. The third evaluation metric is Willmott's index of agreement,[64] which measures the degree to which a model's predictions are error-free. It can point out the total discrepancies between the imputations and the observations that are not captured by the index of agreement. It is unitless, and its largest possible value is 1.

$$d = 1 - \frac{\sum_{k=1}^{N}(\hat{y}_k - y_k)^2}{\sum_{k=1}^{N}(|\hat{y}_k - \langle y\rangle| + |y_k - \langle y\rangle|)} \tag{18}$$

In eqs 16−18, k denotes a sample index, N is the total number of samples, $\hat{y}_k$ is an imputed ozone value, and y_k is a measured ozone value.

To ensure robustness of the imputation methods and hyperparameters, we iteratively generate ten versions of the aforementioned data splits and compare their evaluation results. We also produce an imputed data set with the best method and hyperparameters and crosscheck this imputation with the final validated data set UBA provides.

RESULTS

This section is organized as follows: First, we describe hyperparameter tuning and model fitting. Then follows the evaluations of three distinct missing data cases: short gaps of up to 5 h length, longer gaps, and gaps at multiple stations. We then consolidate the findings for the three cases into a combined imputation. Lastly, we describe the production of the final imputed data set.

Hyperparameter Tuning and Model Fitting. To tune the hyperparameters for the nearest-neighbor hybrid (*nnh*), the random forest (*rf*), and the correct and smooth postprocessing, the models were fit on the training set and evaluated on the validation set. The *nnh* model (eqs 3−5) has only the parameter L_t. We tuned this parameter by starting with a threshold length of 1 h and increasing it in steps of 1 h. The best evaluation metrics were found for a threshold length of L_t = 6 h. For the random forest (*rf*, eq 7), 500 trees were initially trained with unlimited depths. To avoid overfitting, the maximum depth was then diminished, until the training and

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

Table 1. Evaluation Results for Short Gaps

gap length	model		R^2	RMSE [ppb]	d
1 h	spatiotemporal mean		0.00	15.82	0.02
		+ correct and smooth	0.70	8.67	0.91
	spatial mean		0.63	9.60	0.88
		+ correct and smooth	0.82	6.62	0.95
	nearest-neighbor hybrid		**0.97**	**2.43**	**0.99**
		+ correct and smooth	0.96	3.07	0.99
	EAC4 reanalyses		0.53	10.84	0.86
		+ correct and smooth	0.81	6.91	0.94
	random forest		0.85	6.15	0.96
		+ correct and smooth	0.90	4.94	0.97
2 h	spatiotemporal mean		0.00	15.82	0.02
		+ correct and smooth	0.65	9.36	0.89
	spatial mean		0.61	9.92	0.87
		+ correct and smooth	0.79	7.17	0.94
	nearest-neighbor hybrid		**0.96**	**3.33**	**0.99**
		+ correct and smooth	0.94	3.91	0.98
	EAC4 reanalyses		0.50	11.22	0.85
		+ correct and smooth	0.78	7.50	0.93
	random forest		0.84	6.26	0.95
		+ correct and smooth	0.89	5.34	0.97
3−5 h	spatiotemporal mean		0.00	15.12	0.02
		+ correct and smooth	0.60	9.54	0.87
	spatial mean		0.64	9.55	0.87
		+ correct and smooth	0.76	7.27	0.93
	nearest-neighbor hybrid		**0.91**	**4.44**	**0.98**
		+ correct and smooth	0.90	4.82	0.97
	EAC4 reanalyses		0.49	10.84	0.85
		+ correct and smooth	0.74	7.72	0.92
	random forest		0.81	6.64	0.94
		+ correct and smooth	0.86	5.68	0.96

Table 2. Evaluation Results for Long Gaps

gap length	model		R^2	RMSE [ppb]	d
6−23 h	spatiotemporal mean		0.00	15.78	0.00
		+ correct and smooth	0.55	10.54	0.84
	spatial mean		0.64	9.38	0.88
		+ correct and smooth	0.75	7.88	0.92
	nearest-neighbor hybrid		0.75	7.85	0.93
		+ correct and smooth	0.79	7.13	0.94
	EAC4 reanalyses		0.56	10.47	0.87
		+ correct and smooth	0.73	8.22	0.92
	random forest		0.84	6.34	0.95
		+ correct and smooth	**0.87**	**5.65**	**0.96**
1−6 days	spatiotemporal mean		0.00	14.93	0.03
		+ correct and smooth	0.52	10.32	0.82
	spatial mean		0.59	9.56	0.87
		+ correct and smooth	0.71	8.00	0.91
	nearest-neighbor hybrid		0.72	7.85	0.93
		+ correct and smooth	0.78	7.07	0.94
	EAC4 reanalyses		0.47	10.84	0.85
		+ correct and smooth	0.70	8.22	0.91
	random forest		0.81	6.40	0.95
		+ correct and smooth	**0.86**	**5.64**	**0.96**
≥7 days	spatiotemporal mean		0.00	16.25	0.01
		+ correct and smooth	0.57	10.62	0.84
	spatial mean		0.67	9.27	0.89
		+ correct and smooth	0.77	7.76	0.93
	nearest-neighbor hybrid		0.69	9.00	0.92
		+ correct and smooth	0.75	8.12	0.93
	EAC4 reanalyses		0.56	10.81	0.87
		+ correct and smooth	0.75	8.17	0.92
	random forest		0.83	6.75	0.95
		+ correct and smooth	**0.86**	**6.18**	**0.96**

validation errors were the same. This resulted in a depth of 15. Features for the random forest were selected by forward feature selection.[65] As a result, all features were selected except the reanalyzed NO concentration. The parameters $\alpha_{1,2}$, $L_{1,2}$, and γ of correct and smooth (eqs 9−15) were tuned by grid search. Details and results of the hyperparameter tuning can be found in section S3.

To fit the final models, which are analyzed in the following, the optimal hyperparameters were used and models were fitted on both the training and validation sets. The models were then evaluated on the test set.

Imputation of Short Gaps. Table 1 shows the evaluation metrics of the imputation results of short gaps up to a length of 5 h. The nearest-neighbor hybrid (*nnh*), which carries out a linear interpolation (*lin*) for these gaps, performs best. Its R^2 values are between 0.91−0.97, RMSEs are between 2.43−4.44 ppb, and d is ≥0.98. This agrees with the results of Junninen et al.,[23] who found linear interpolation to be most effective for short gaps. As expected, the performance of the linear interpolation drops with the length of the gap as this method does not consider auxiliary variables or the daily cycle of ozone concentrations.

Imputation of Longer Gaps. Table 2 shows the evaluation metrics of the imputation of gaps that are 6 h or longer. The random forest in connection with correct and smooth performs best for these gaps, with R^2 values of 0.86−

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

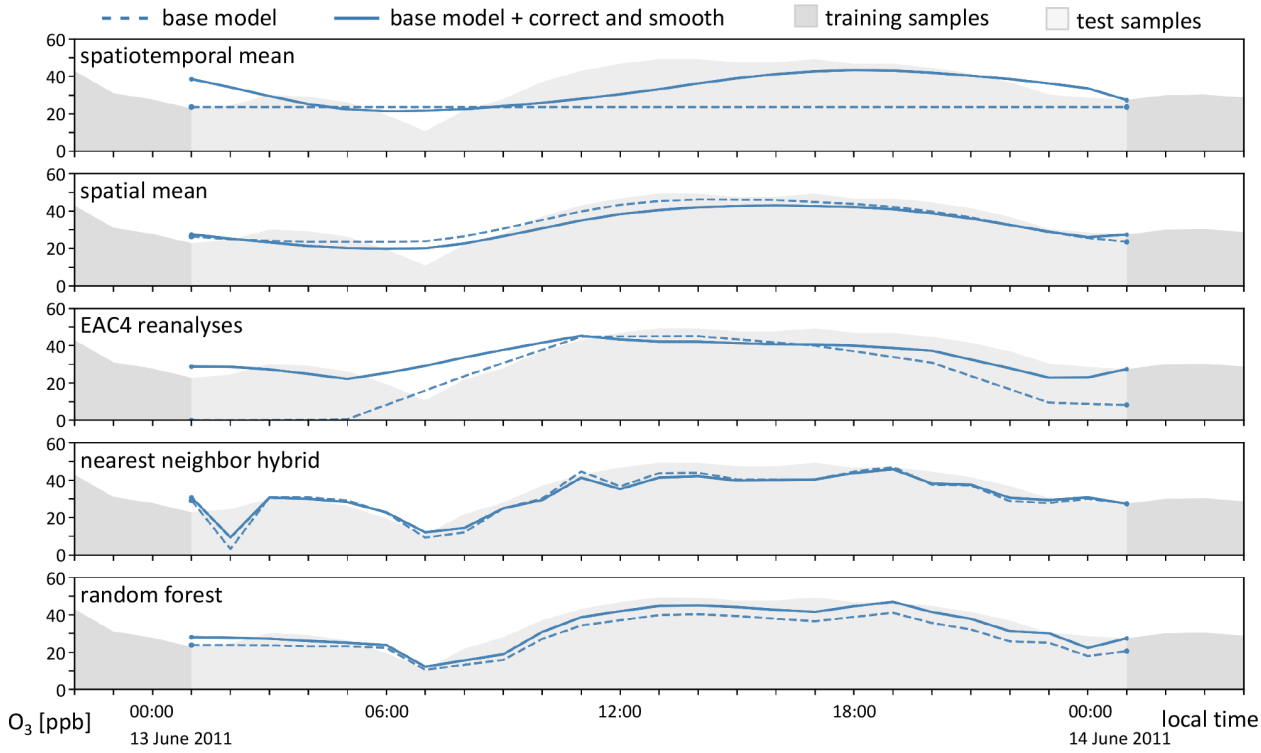


Figure 3. Example imputations of an isolated 24 h gap at station 'DEHE001' in the city of Darmstadt. The dashed lines are the imputations from the base models. The solid lines are the correct and smooth imputation postprocessed base models.

0.87, RMSEs of 5.64−6.18 ppb, and $d = 0.96$. Correct and smooth postprocessing decreased the RMSE of the random forest by 0.57−0.76 ppb. With R^2 values of 0.69−0.75, the nearest-neighbor interpolation is a suitable statistical method for missing data imputation but is consistently outperformed by the random forest.

Table 2 also shows how correct and smooth, which relies on available data at neighboring stations for long gaps, improves the base models. Its effectiveness shows best with base models of low complexity. One example is the spatiotemporal mean which imputes the same constant to all gaps. The R^2 value of this method alone is zero, because there is no variance in the imputations. Correct and smooth postprocessing increased the R^2 values of the spatiotemporal mean by 0.52−0.57. This improvement is achieved only by passing information from neighboring stations across the graph edges defined in the given monitoring network. Although the correct and smooth algorithm is iterative, information on the same station from distant time steps is not propagated into longer gaps because the autoscale option of the algorithm reduces the influence of training nodes on unlabeled nodes with the number of "hops". We therefore neglect autocorrelation of ozone values for times longer than the diurnal cycle.

Figure 3 shows the imputed concentrations of the different methods using a 24 h gap at an urban background station in the city of Darmstadt (UBA id 'DEHE001', TOAR id 3443) as an example. There are 18 stations in the radius of 50 km around this station with distances of 11.8−49.9 km, and it can receive information from these stations across the defined graph edges. In the case of spatiotemporal mean, correct and smooth postprocessing could introduce a daily cycle. It also improved the other base models, even though they already predicted the daily cycle. The random forest has low errors but is improved slightly by being correct and smooth.

Imputation of Gaps at Multiple Stations. Table 3 shows evaluation metrics of gaps occurring at all stations simultaneously. Similar to the gaps occurring at single stations, the nearest-neighbor hybrid (which carries out a linear interpolation for short gaps) reaches the best evaluation metrics for gaps of up to 5 h length. The longer gaps are still imputed best by the random forest in combination with correct and smooth, yet correct and smooth improved the RMSE by only 0.07 ppb in this situation. This can be explained by the fact that no neighboring data are available. Hence, the imputation has to rely on the features alone, which generally results in lower evaluation metrics.[48,66]

Combined Imputation. According to the results presented in Tables 1−3, we created a combined imputation to evaluate our developed method. We imputed all short gaps with a length of up to 5 h with linear interpolation and all longer gaps with random forest and correct and smooth. We did not differentiate between gaps at a single station or at multiple stations since these methods are shown to be most effective, regardless of whether a gap occurs at one station or at multiple stations. The evaluation metrics of the complete test set and the iteratively generated data splits are shown in Table 4. They indicate the robustness of the imputation method. Figure 4 shows heatmaps of true and imputed concentrations, with differentiation between short and long gaps.

Figure 5 shows a summary of how gap characteristics affect the evaluation metric R^2 for the different base models in combination with correct and smooth. The R^2 value generally decreases with an increasing gap length. Furthermore, there is a weak trend in improved R^2 when more neighboring stations are available. Both trends are more apparent for the simple base

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

Table 3. Results for the Gaps at Multiple Stations[a]

gap length	model		R^2	RMSE [ppb]	d
3−5 h	spatiotemporal mean		−0.01	15.39	0.11
		+ correct and smooth	0.63	9.32	0.89
	spatial mean		0.42	11.67	0.77
		+ correct and smooth	0.67	8.75	0.89
	nearest-neighbor hybrid		**0.92**	**4.45**	**0.98**
		+ correct and smooth	0.90	4.64	0.97
	EAC4 reanalyses		0.46	11.32	0.81
		+ correct and smooth	0.72	8.06	0.91
	random forest		0.78	7.06	0.93
		+ correct and smooth	0.84	5.95	0.96
1−6 days	spatiotemporal mean		−0.04	13.07	0.24
		+ correct and smooth	−0.13	13.64	0.31
	spatial mean		0.37	10.19	0.80
		+ correct and smooth	0.43	9.60	0.82
	nearest-neighbor hybrid		0.51	8.94	0.87
		+ correct and smooth	0.59	8.18	0.88
	EAC4 reanalyses		0.42	9.77	0.85
		+ correct and smooth	0.55	8.57	0.87
	random forest		0.78	5.95	0.93
		+ correct and smooth	**0.79**	**5.88**	**0.94**

[a]The nearest-neighbor hybrid method is a linear interpolation for 3−5 h gaps and a nearest-neighbor interpolation for longer gaps.

Table 4. Evaluation Metrics of the Test Set and Spread in Iterative Data Splits

evaluation metric	test set	ten iterative data splits
R^2	0.89	$0.89 \leq R^2 \leq 0.90$
RMSE [ppb]	5.13	$5.52 \geq \text{RMSE} \geq 5.12$
d	0.97	$0.97 \leq d \leq 0.97$

models, such as the spatial mean and the spatiotemporal mean. The random forest in connection with correct and smooth, which has the best evaluation metrics, is also least affected by variations of the gap characteristics.

Imputed Data Set. The imputed data set, which was produced within the scope of this study, is available under the DOI 10.23728/b2share.04821864a81f40af89c7633889f147cb. To produce this data set, we imputed all missing ozone data using the combined imputation and the trained random forest model. Note that data points that were masked for validation and testing were unmasked again in this final output data set; i.e., for these samples, the original measured ozone values are reported. About 180,000 samples that are missing in the preliminary UBA data set which we used to develop our method are present in the final data set which UBA provided in the following year. This is approximately 7.3% of the theoretically available samples. Cross-checking these with our imputations yields an R^2 of 0.83, an RMSE of 5.63 ppb, and an index of agreement of 0.95. The evaluation metrics are slightly inferior to those reported in Table 4.

As mentioned in the Introduction, the number of exceedances of ozone concentration thresholds is an important indicator of the assessment of air quality. One example is the number of exceedances of daily maximum 8 h values greater than 70 ppb during the summer (nvgt70 summer).[22] As a proof of concept, we count the number of additional threshold exceedances that the imputed data set contains (Figure 6). Of the total number of about 3.6×10^5 imputed values, about 10^4 samples yield ozone values above 50 ppb. Regarding the nvgt70 metric, 512 samples were imputed to the data set which exceed the threshold of 70 ppb. This shows that data imputation with our method can improve the robustness of air quality assessments.

As a second proof of concept, we imputed ozone data at station locations where no data were reported at all (Figure 7). We expect the evaluation metrics of these longer modeled ozone time series to be similar to longer gaps at single stations (Table 2), even though validation is impossible. The modeled time series is less variable than those measured at neighboring stations. This is because any (ozone) model, machine learning or otherwise, has problems predicting extremes.[67] Dips and peaks in the measured time series can sometimes be attributed to noise due to short-term or small-scale effects on ozone that are not resolved in the auxiliary data and therefore are not

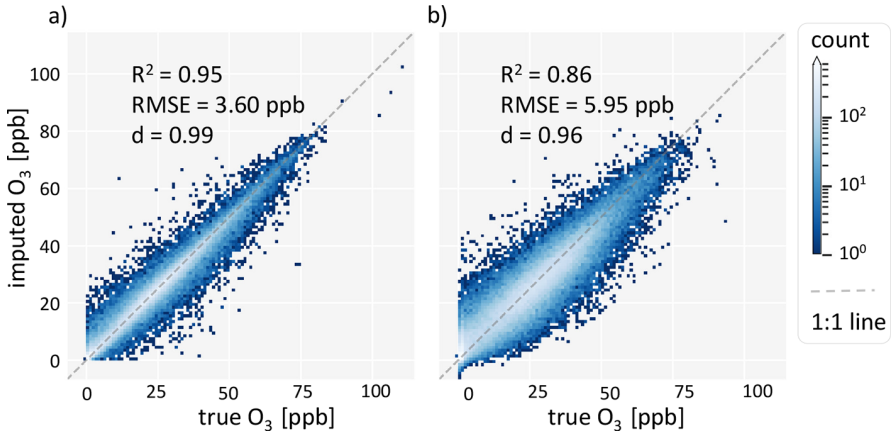


Figure 4. Heatmap of true versus imputed concentrations. (a) Short gaps of up to 5 h length, imputed by linear interpolation, and (b) random forest + correct and smooth for long gaps. This figure does not differentiate between isolated gaps and gaps at all stations.

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

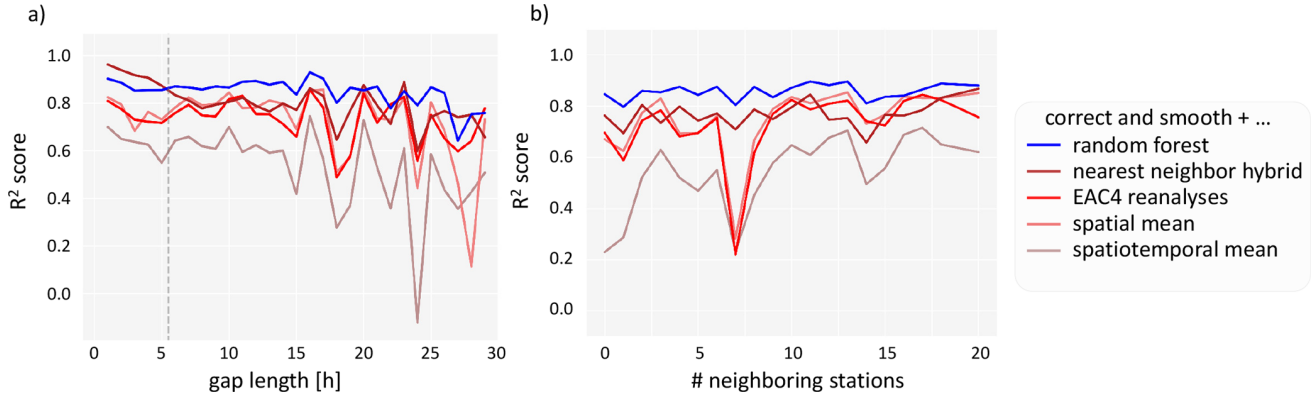


Figure 5. R^2 evaluation metric vs gap characteristics. (a) Different gap lengths up to 30 h. The dashed line marks the gap length 5 h. This is when the final imputation model changes from linear interpolation to random forest + correct and smooth. (b) Number of neighbors in a radius of 50 km around the station. This plot only contains data from isolated gaps longer than 5 h.

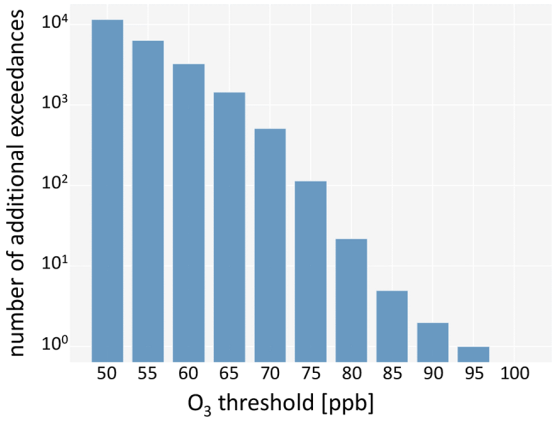


Figure 6. Number of additional exceedances of ozone thresholds contained in the data set after imputation of the data.

represented in the model. Some of this is improved by correct and smooth: If all neighboring stations have a peak where the base model does not, then it is well corrected. An example can be seen in panel (c) of Figure 7 at time step 26.

DISCUSSION

Imputing Missing Data in the UBA Data Set. The goal of this work was to impute missing ozone data at 278 stations of the UBA network in the year 2011. By fusing a variety of auxiliary data and available measurements using (graph) machine learning, high-accuracy imputations can be achieved. We applied other published methods as baseline methods to the UBA data set to compare our method with them. A direct comparison with the evaluation metrics reported in other studies may be misleading because they use different data sets, gap characteristics, and evaluation metrics than we do. We have chosen common baseline methods, namely, (1) imputation with mean values as is often implicitly done in

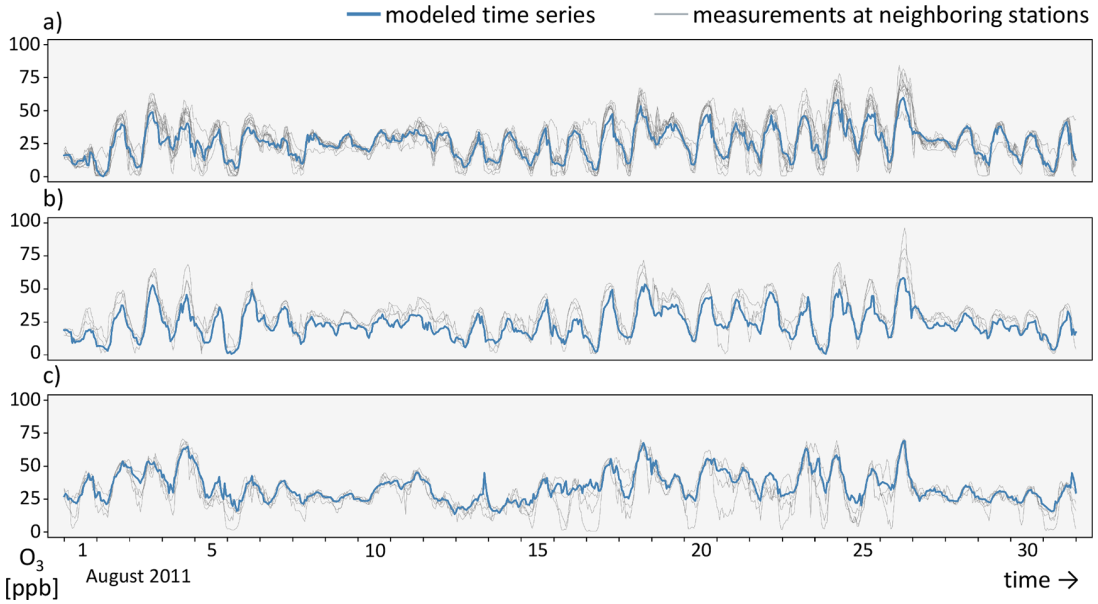


Figure 7. One-month-long excerpt of simulated ozone time series at three locations in Germany. They were modeled using our random forest and were correct and smooth. For comparison, the available measurements at stations within a radius of 50 km around the modeled locations are given. (a) Urban traffic location in the city of Borna, Sachsen, with 10 neighboring stations. (b) Urban traffic location in the city of Magdeburg with 4 neighboring stations. (c) Modeled time series in a rural background area west of Kassel with 5 neighboring stations.

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

the calculation of ozone metrics,[13,22] (2) nearest-neighbor hybrid which is the best statistical method found by Junninen et al.,[23] (3) EAC4 atmospheric reanalysis,[50] and (4) random forest, which is a state-of-the-art machine learning method for structured data. Our method achieves equal or higher imputation accuracy than these other methods, depending on the gap characteristics. Also, our method is robust, with reasonable variations in the evaluation metrics given different numbers of neighbors, the iterative data splitting, and the length of the gaps.

Unlike approaches such as physics-guided machine learning,[68] our method relies on the geospatial and statistical properties of the ozone data and auxiliary data without considering the physical or chemical processes mentioned in the Introduction. A strength of the correct and smooth method is that the correction step accounts for influences that the base models cannot predict without specifying those influences. Instead, it corrects the prediction by assuming that neighboring data points are subject to the same unknown influences; i.e., their base model errors are correlated. Smoothing ozone values across the graph structure defined on the monitoring network, as performed in the third step of correct and smooth, is a strongly simplified implementation of the ozone transport and diffusion processes. It does not consider wind speed or direction. Even though this works reasonably well, it should be improved in future models. Considering the spatiotemporal inhomogeneity of ozone and of air pollution in general, we have considered the local to regional differences in ozone levels by including both precursor emissions and meteorological parameters in our base models. We have furthermore used measurements of monitoring stations in the radius of 50 km around a missing value wherever available to better account for local to regional variances in the pollution.

The described method is suitable for near-real-time operational settings such as an imputation application for the TOAR data analysis services. Such a service is useful considering that the final validated UBA data set will not be available until the following year. Linear interpolation, random forest, and correct and smooth are comparably cheap algorithms that only take seconds to minutes to execute. Therefore, a near-real-time imputation of data can be potentially achieved by using these algorithms.

Prospects for Graph Machine Learning in Air Quality Research. We showed that graph machine learning is suitable to be used with ozone data of the UBA monitoring network due to the irregular structure of the available data. We expect our findings to apply to other air quality data as well, although further studies would be needed to assess the imputation results for variables with different statistical properties, such as nitrogen oxide or particulate matter concentrations. One advantage of correct and smooth is that it can be used with any other feature-based method and for bias correction of numerical models.

With the definition of one data point as one node and the basing of the edge definition on the spatiotemporal distance between the nodes, the graph definition we used is relatively simple. More sophisticated approaches that should be explored in the future include time-resolved graphs[69] for spatiotemporal machine learning or transformer architectures,[70] which can learn to attend to the most helpful features in unstructured data. These architectures could be trained to take transport and advection of air pollutants into account by incorporating wind directions.[19] One promising approach is also to infer the graph from the underlying data set.[47,71] To further explore how the graph structure affects the results and what parameters are most crucial, sensitivity studies are necessary.

Many studies impute missing concentrations of multiple pollutants simultaneously and with varying input data available.[23,24] From a graph perspective, this would require an algorithm that could handle different kinds of nodes with different kinds of labels. An algorithm like this would be especially interesting when real measurements of auxiliary data are used instead of reanalyses, because air quality measurements and measurements of meteorological parameters are often not reported from the same stations or they may have gaps themselves.

Further Applications. The study presented here works on a spatially (Germany) and temporally (year 2011) limited domain. The only prerequisites to using this method in a different domain would be a similar spatial coverage of measurement stations and the availability of similar auxiliary data. Reasonably dense station networks exist in large parts of Europe, the United States, and East Asia but not in other world regions such as South America and Africa.[13] Besides the lack of neighboring air quality stations, there may also be larger biases in the auxiliary data as documented, for example, with respect to the CAMS emissions and reanalyses.[50,51]

Besides missing data imputation, the method developed here could also be adapted for other questions posed in air quality research. One example is quality control—a common problem of graph theory is to flag untrustworthy nodes, such as untrustworthy Web sites or untrustworthy transactions.[72] Similarly, untrustworthy measurements could be flagged with our method. This study showed that the method could predict meaningful ozone concentrations at places or time steps without measurements (Figure 7). Technically it would also be possible to predict the ozone time series at all grid points of a regular grid and therefore provide gridded ozone fields. This would be a logical extension of the mapping study by Betancourt et al.[48]

■ ASSOCIATED CONTENT

Data Availability Statement

The training data, including the graph data set, can be found under, if needed, 10.23728/b2share.59281340dd37485eb2c6a08de3587c13. The imputed data set can be found under 10.23728/b2share.04821864a81f40af89c7633889f147cb. The code which can be used to reproduce the experiments presented in this study can be found under https://gitlab.jsc.fz-juelich.de/esde/machine-learning/ozone-imputation.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c05104.

More information on the ozone data used in this study and the hyperparameters for the correct and smooth methods (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Martin G. Schultz − *Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany;* orcid.org/0000-0003-3455-774X; Email: m.schultz@fz-juelich.de

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

Authors

Clara Betancourt − *Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany;* orcid.org/0000-0002-1347-5297

Cathy W. Y. Li − *Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany; Max-Planck-Institut für Meteorologie, 20146 Hamburg, Germany*

Felix Kleinert − *Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c05104

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the German Environment Agency (Umweltbundesamt − UBA) and the German state authorities for the collection and provision of air quality data and for helpful information concerning the specific dataset we used. We gratefully acknowledge the efforts of Sabine Schröder and her colleagues for maintaining the database of the Tropospheric Ozone Assessment Report (TOAR) and helping to access and interpret the data. The authors thank Qian Huang for answering questions about the correct and smooth method. We thank Franca Hoffmann and Ankit Patnala for helpful discussions. We thank the anonymous reviewers for the constructive review and encouraging comments. C.B. and F.K. acknowledge funding from the European Research Council, H2020 Research Infrastructures (IntelliAQ (Grant no. ERC-2017-ADG#787576)). C.W.Y.L. has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement no. 870301 and from the Helmholtz Information & Data Science Academy (HIDA), enabling a short-term research stay at Jülich Supercomputing Centre for this collaborated work. The authors gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS[73] at the Jülich Supercomputing Centre (JSC). We thank three anonymous reviewers for their suggestions to improve this work.

■ REFERENCES

(1) 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.; Williams, M. L. Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmos. Chem. Phys.* **2015**, *15*, 8889−8973.

(2) Gaudel, A.; Cooper, O. R.; Ancellet, G.; Barret, B.; Boynard, A.; Burrows, J. P.; Clerbaux, C.; Coheur, P. F.; Cuesta, J.; Cuevas, E.; Doniki, S.; Dufour, G.; Ebojie, F.; Foret, G.; Garcia, O.; Granados-Muñoz, M. J.; Hannigan, J. W.; Hase, F.; 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.; 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.* **2018**, *6*, 39.

(3) Henschel, S.; Chan, G.; World Health Organization. *Regional Office for Europe, Health risks of air pollution in Europe - HRAPIE project: New emerging risks to health from air pollution - results from the survey of experts*; World Health Organization: 2013.

(4) World Health Organization. *Review of evidence on health aspects of air pollution: REVIHAAP project*; World Health Organization: 2021.

(5) 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.; Feng, Z. Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. *Elem. Sci. Anth.* **2018**, *6*, 12.

(6) Van Dingenen, R.; Dentener, F. J.; Raes, F.; Krol, M. C.; Emberson, L.; Cofala, J. The global impact of ozone on agricultural crop yields under current and future air quality legislation. *Atmos. Environ.* **2009**, *43*, 604−618.

(7) 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.; Xu, X. Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation. *Elem. Sci. Anth.* **2018**, *6*, 47.

(8) Mills, G.; Sharps, K.; Simpson, D.; Pleijel, H.; Broberg, M.; Uddling, J.; Jaramillo, F.; Davies, W. J.; Dentener, F.; Van den Berg, M.; et al. Ozone pollution will compromise efforts to increase global wheat production. *Global change biology* **2018**, *24*, 3560−3574.

(9) IPCC Climate. Chapter 8: Anthropogenic and Natural Radiative Forcing. In *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*; IPCC: 2013; pp 659−740.

(10) Skeie, R. B.; Myhre, G.; Hodnebrog, Ø.; Cameron-Smith, P. J.; Deushi, M.; Hegglin, M. I.; Horowitz, L. W.; Kramer, R. J.; Michou, M.; Mills, M. J.; et al. Historical total ozone radiative forcing derived from CMIP6 simulations. *npj Climate and Atmospheric Science* **2020**, *3*, 32.

(11) Wallace, J.; Hobbs, P. *Atmospheric Science: An Introductory Survey*, 2nd ed.; International Geophysics Series; Elsevier Academic Press: Burlington, MA, 2006; Vol. *92*.

(12) Brasseur, G., Orlando, J. J., Tyndall, G. S., Eds. *Atmospheric chemistry and global change*, 1st ed.; Oxford University Press: New York, 1999.

(13) 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.; Feigenspan, S.; Fiebig, M.; Hjellbrekke, A.-G.; Hong, Y.-D.; Kjeld, P. C.; 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-Agullo, E.; Cupeiro, M.; Degorska, A.; Ding, A.; Fröhlich, M.; Frolova, M.; Gadhavi, H.; Gheusi, F.; Gilge, S.; Gonzalez, M. Y.; Gros, V.; Hamad, S. H.; Helmig, D.; Henriques, D.; Hermansen, O.; Holla, R.; Huber, J.; Im, U.; Jaffe, D. A.; Komala, N.; Kubistin, D.; Lam, K.-S.; Laurila, T.; Lee, H.; Levy, I.; Mazzoleni, C.; Mazzoleni, L.; McClure-Begley, A.; Mohamad, M.; Murovec, 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.; Xiaobin, X.; Xue, L.; Zhiqiang, M. Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations. *Elementa Science of the Anthropocene* **2017**, *5*, 58.

(14) 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* **2008**, 1−44.

(15) Schultz, M. G.; Akimoto, H.; Bottenheim, J.; Buchmann, B.; Galbally, I. E.; Gilge, S.; Helmig, D.; Koide, H.; Lewis, A. C.; Novelli, P. C. The Global Atmosphere Watch reactive gases measurement network. *Elem. Sci. Anth.* **2015**, *3*, 000067.
(16) World Health Organization. *WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide*; World Health Organization: Geneva, 2005; pp 173−186.
(17) Davison, A.; Hemphill, M. On the statistical analysis of ambient ozone data when measurements are missing. *Atmospheric Environment (1967)* **1987**, *21*, 629−639.
(18) Kleinert, F.; Leufen, L. H.; Schultz, M. G. IntelliO3-ts v1.0: A neural network approach to predict near-surface ozone concentrations in Germany. *Geosci. Model Dev.* **2021**, *14*, 1−25.
(19) Kleinert, F.; Leufen, L. H.; Lupascu, A.; Butler, T.; Schultz, M. G. Representing chemical history in ozone time-series predictions − a model experiment study building on the MLAir (v1.5) deep learning framework. *Geoscientific Model Development* **2022**, *15*, 8913−8930.
(20) Sayeed, A.; Choi, Y.; Eslami, E.; Jung, J.; Lops, Y.; Salman, A. K.; Lee, J.-B.; Park, H.-J.; Choi, M.-H. A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Sci. Rep.* **2021**, *11*, 10891.
(21) Leufen, L. H.; Kleinert, F.; Schultz, M. G. Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-h average ozone prediction. *Environmental Data Science* **2022**, *1*, No. e10.
(22) 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.; Gerosa, G. Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research. *Elementa: Science of the Anthropocene* **2018**, *6*, 27.
(23) Junninen, H.; Niska, H.; Tuppurainen, K.; Ruuskanen, J.; Kolehmainen, M. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* **2004**, *38*, 2895−2907.
(24) Gómez-Carracedo, M.; Andrade, J.; López-Mahía, P.; Muniategui, S.; Prada, D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems* **2014**, *134*, 23−33.
(25) Arroyo, Á.; Herrero, Á.; Tricio, V.; Corchado, E.; Woźniak, M. Neural models for imputation of missing ozone data in air-quality datasets. *Complexity* **2018**, *2018*, 7238015.
(26) Little, R. J.; Rubin, D. B. *Statistical analysis with missing data*; John Wiley & Sons: 2019; Vol. *793*.
(27) García-Laencina, P. J.; Sancho-Gómez, J.-L.; Figueiras-Vidal, A. R. Pattern classification with missing data: a review. *Neural Computing and Applications* **2010**, *19*, 263−282.
(28) Rubin, D. B. Inference and missing data. *Biometrika* **1976**, *63*, 581−592.
(29) Junger, W.; Ponce de Leon, A. Imputation of missing data in time series for air pollutants. *Atmos. Environ.* **2015**, *102*, 96−104.
(30) Alsaber, A. R.; Pan, J.; Al-Hurban, A. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). *International Journal of Environmental Research and Public Health* **2021**, *18*, 1333.
(31) Zheng, Y.; Yi, X.; Li, M.; Li, R.; Shan, Z.; Chang, E.; Li, T. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*; 2015; pp 2267−2276.
(32) Liu, X.; Wang, X.; Zou, L.; Xia, J.; Pang, W. Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environ. Int.* **2020**, *139*, 105713.
(33) Gryech, I.; Ben-Aboud, Y.; Ghogho, M.; Kobbane, A. On spatial prediction of urban air pollution. In *2021 17th International Conference on Intelligent Environments (IE)*; 2021; pp 1−6.
(34) Hamilton, W. L. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* **2020**, *14*, 1−159.
(35) Barabási, A.-L. *Network science*; Cambridge University Press: Cambridge, 2016.
(36) Ortega, A.; Frossard, P.; Kovačević, J.; Moura, J. M. F.; Vandergheynst, P. Graph Signal Processing: Overview, Challenges, and Applications. *Proceedings of the IEEE* **2018**, *106*, 808−828.
(37) Dong, X.; Thanou, D.; Toni, L.; Bronstein, M.; Frossard, P. Graph Signal Processing for Machine Learning: A Review and New Perspectives. *IEEE Signal Processing Magazine* **2020**, *37*, 117−127.
(38) Carmona-Cabezas, R.; Gómez-Gómez, J.; Ariza-Villaverde, A. B.; Gutiérrez de Ravé, E. G.; Jiménez-Hornero, F. J. Can complex networks describe the urban and rural tropospheric O_3 dynamics? *Chemosphere* **2019**, *230*, 59−66.
(39) Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Sci. Total Environ.* **2019**, *664*, 1−10.
(40) Ge, L.; Wu, K.; Zeng, Y.; Chang, F.; Wang, Y.; Li, S. Multi-scale spatiotemporal graph convolution network for air quality prediction. *Applied Intelligence* **2021**, *51*, 3491−3505.
(41) Ferrer-Cid, P.; Barcelo-Ordinas, J. M.; Garcia-Vidal, J. Data reconstruction applications for IoT air pollution sensor networks using graph signal processing. *Journal of Network and Computer Applications* **2022**, *205*, 103434.
(42) Ferrer-Cid, P.; Barcelo-Ordinas, J. M.; Garcia-Vidal, J. Volterra Graph-Based Outlier Detection for Air Pollution Sensor Networks. *IEEE Transactions on Network Science and Engineering* **2022**, *9*, 2759−2771.
(43) Han, J.; Liu, H.; Xiong, H.; Yang, J. Semi-Supervised Air Quality Forecasting via Self-Supervised Hierarchical Graph Neural Network. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *35*, 5230−5243.
(44) Pinder, T.; Turnbull, K.; Nemeth, C.; Leslie, D. *AI for Earth and Space Science: Street-level air pollution modelling with graph Gaussian processes*; ICLR: 2022.
(45) You, J.; Ma, X.; Ding, Y.; Kochenderfer, M. J.; Leskovec, J. Handling Missing Data with Graph Representation Learning. *Advances in Neural Information Processing Systems* **2020**, 19075−19087.
(46) Spinelli, I.; Scardapane, S.; Uncini, A. Missing data imputation with adversarially-trained graph convolutional networks. *Neural Networks* **2020**, *129*, 249−260.
(47) Jiang, X.; Tian, Z.; Li, K. A Graph-Based Approach for Missing Sensor Data Imputation. *IEEE Sensors Journal* **2021**, *21*, 23133−23144.
(48) Betancourt, C.; Stomberg, T. T.; Edrich, A.-K.; Patnala, A.; Schultz, M. G.; Roscher, R.; Kowalski, J.; Stadtler, S. Global, high-resolution mapping of tropospheric ozone − explainable machine learning and impact of uncertainties. *Geoscientific Model Development* **2022**, *15*, 4331−4354.
(49) Bollmeyer, C.; Keller, J. D.; Ohlwein, C.; Wahl, S.; Crewell, S.; Friederichs, P.; Hense, A.; Keune, J.; Kneifel, S.; Pscheidt, I.; Redl, S.; Steinke, S. Towards a high-resolution regional reanalysis for the European CORDEX domain. *Quarterly Journal of the Royal Meteorological Society* **2015**, *141*, 1−15.
(50) Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.-M.; Dominguez, J. J.; Engelen, R.; Eskes, H.; Flemming, J.; Huijnen, V.; Jones, L.; Kipling, Z.; Massart, S.; Parrington, M.; Peuch, V.-H.; Razinger, M.; Remy, S.; Schulz, M.; Suttie, M. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics* **2019**, *19*, 3515−3556.
(51) Granier, C.; Darras, S.; van der Gon, H. D.; Doubalova, J.; Elguindi, N.; Galle, B.; Gauss, M.; Guevara, M.; Jalkanen, J.; Kuenen, J. *The Copernicus Atmosphere Monitoring Service global and regional emissions (April 2019 version)*; Copernicus Atmosphere Monitoring Service: 2019; Vol. *4*; DOI: 10.24380/d0bn-kx16.
(52) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.

(53) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2020**, *2*, 56−67.
(54) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, *29*, 1189−1232.
(55) Leskovec, J. *Stanford University CS224W: Machine Learning with Graphs [lecture]*; 2021; http://web.stanford.edu/class/cs224w/ (accessed September 1, 2022).
(56) Nickel, M.; Murphy, K.; Tresp, V.; Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE* **2016**, *104*, 11−33.
(57) Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203−205.
(58) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*, 4−24.
(59) Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; Sun, Y. *Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification*; 2021; DOI: 10.48550/arXiv.2009.03509 (accessed August 23, 2023).
(60) Huang, Q.; He, H.; Singh, A.; Lim, S.-N.; Benson, A. R. *Combining Label Propagation and Simple Models Out-performs Graph Neural Networks*; 2020; DOI: 10.48550/arXiv.2010.13993 (accessed August 23, 2023).
(61) Wang, H.; Leskovec, J. *Unifying Graph Convolutional Neural Networks and Label Propagation*; 2020; DOI: 10.48550/arXiv.2002.06755 (accessed August 23, 2023).
(62) Chang, K.-L.; Schultz, M. G.; Lan, X.; McClure-Begley, A.; Petropavlovskikh, I.; Xu, X.; Ziemke, J. R. Trend detection of atmospheric time series: Incorporating appropriate uncertainty estimates and handling extreme events. *Elementa: Science of the Anthropocene* **2021**, *9*, 00035.
(63) Isufi, E.; Gama, F.; Shuman, D. I.; Segarra, S. *Graph Filters for Signal Processing and Machine Learning on Graphs*; 2022; DOI: 10.48550/arXiv.2211.08854 (accessed August 23, 2023).
(64) Willmott, C. J. On the validation of models. *Physical geography* **1981**, *2*, 184−194.
(65) Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Modell. Softw.* **2018**, *101*, 1−9.
(66) Betancourt, C.; Stomberg, T.; Roscher, R.; Schultz, M. G.; Stadtler, S. AQ-Bench: a benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data* **2021**, *13*, 3013−3033.
(67) Young, P. J.; Naik, V.; Fiore, A. M.; Gaudel, A.; Guo, J.; Lin, M.; Neu, J.; Parrish, D.; Rieder, H.; Schnell, J. Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elementa: Science of the Anthropocene* **2018**, *6*, 10.
(68) Pawar, S.; San, O.; Aksoylu, B.; Rasheed, A.; Kvamsdal, T. Physics guided machine learning using simplified theories. *Phys. Fluids* **2021**, *33*, 011701.
(69) Rozemberczki, B.; Scherer, P.; He, Y.; Panagopoulos, G.; Riedel, A.; Astefanoaei, M.; Kiss, O.; Beres, F.; López, G.; Collignon, N.; Sarkar, R. PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, New York, 2021; pp 4564−4573.
(70) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
(71) Jabłoński, I. Graph Signal Processing in Applications to Sensor Networks, Smart Grids, and Smart Cities. *IEEE Sensors Journal* **2017**, *17*, 7659−7666.
(72) Berkhin, P. A. Survey on PageRank Computing. *Internet Mathematics* **2005**, *2*, 73−120.
(73) Krause, D. JUWELS: Modular Tier-0/1 Supercomputer at Jülich Supercomputing Centre. *Journal of large-scale research facilities (JLSRF)* **2019**, *5*, 1−8.

https://doi.org/10.1021/acs.est.3c05104
Environ. Sci. Technol. 2023, 57, 18246−18258

D.5 Other papers of the author

I published the following papers and software products during my time at Jülich Supercomputing Centre. They are not directly related to this work.

- **Journal article.** Betancourt, C., Küppers, C., Piansawan, T., Sager, U., Hoyer, A. B., Kaminski, H., Rapp, G., John, A. C., Küpper, M., Quass, U., Kuhlbusch, T., Rudolph, J., Kiendler-Scharr, A., and Gensch, I. (2021b). "Firewood residential heating – local versus remote influence on the aerosol burden". In: *Atmospheric Chemistry and Physics* 21.8, pp. 5953–5964. DOI: 10.5194/acp-21-5953-2021.
  Summary. We include stable isotopes in the Lagrangian particle dispersion model FLEXPART to track the origin of firewood aerosol from domestic heating in Germany. Comparing modeled and measured stable isotopes is a new source apportionment methodology that enables quantitative understanding of the underlying atmospheric processes. We show that the investigated aerosols are of local to regional origin.

- **Journal article.** Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). "Can deep learning beat numerical weather prediction?" In: *Philosophical Transactions of the Royal Society A* 379.2194, p. 20200097. DOI: 10.1098/rsta.2020.0097.
  Summary. We discuss whether it is possible to completely replace current numerical models for weather prediction with machine learning. We review state-of-the-art machine learning models and their applicability to weather data, taking their statistical properties into account. We conclude that several more breakthroughs are needed in the field of machine learning before it makes numerical weather prediction obsolete.

- **Journal article.** Betancourt, C., Hagemeier, B., Schröder, S., and Schultz, M. G. (2021c). "Context aware benchmarking and tuning of a TByte-scale air quality database and web service". In: *Earth Science Informatics* 14.3, pp. 1597–1607. DOI: 10.1007/s12145-021-00631-4.
  Summary. We perform benchmarking and performance engineering of a mature terabyte-scale air quality database system that offers on-demand calculations of air quality metrics. We identify the calculation of these metrics outside the database and the necessary transfer of large amounts of raw data as the major performance bottleneck. In-database processing of the metrics resulted in a performance increase of up to 32 %.

- **Web service.** Betancourt, C. (2021). *webDO$_3$SE [web service]*. v1. URL: https://toar-data.fz-juelich.de/do3se/api/v1/.
  Summary. The DO$_3$SE model (Ashmore et al., 2017) calculates stomatal ozone fluxes of several crops, based on plant phenology and environmental influences. So far, the model was only used in local measurement-based studies. webDO$_3$Se is a prototype web service that links the DO3SE model to the database of the Tropospheric Ozone Assessment Report (TOAR) to enable global, long term assessments of the stomatal ozone fluxes.

- **Journal article [not peer reviewed].** Schultz, M. G., Kleinert, F., Leufen, L. H., Betancourt, C., Schröder, S., Gong, B., Stadtler, S., Langguth, M., and Mozaffari, A. (2022). "Artificial intelligence for air quality". In: *The Project Repository Journal* 12.1, pp. 70–73. DOI: 10.54050/PRJ1218384.
  Summary. This article describes up-to-date achievements of the IntelliAQ project, which leverages the potential of modern Deep Learning and Big Data processing in air quality research. IntelliAQ aims to take the analysis of global air pollution observations to a new level and is a foundation for the future development of scientifically sound air quality services.

Acknowledgements

I would like to express my gratitude to all those who have contributed to the completion of my PhD thesis. First, I would like to thank my first supervisor PD Dr. Martin G. Schultz for his guidance, support, and encouragement throughout the research process. Without his initial research idea and ongoing assistance, this work would not have been possible. I would also like to extend my thanks to Prof. Dr. Ribana Roscher, my second supervisor, for her invaluable feedback and insights throughout the course of my research. Additionally, I would like to thank Dr. Scarlet Stadtler for her scientific and personal supervision, as well as her companionship, which has been essential in helping me navigate the challenges of my PhD journey. I am grateful for the collaboration and scientific exchange with all my Co-Authors, whose contributions have significantly enhanced the quality of my research work: Timo T. Stomberg, Ann-Kathrin Edrich, Ankit Patnala, Prof. Julia Kowalski, Dr. Cathy W. Y. Li, and Felix Kleinert. Many thanks to Dr. Mark Faerber for carefully proofreading the final version of this thesis. I also want to express my appreciation to the Earth System Data Exploration group for their support and friendship during my PhD time. Last but not least, I would like to thank my family, my partner Kevin, and my son Tim for their love, encouragement, and unwavering support.