Response to the anonymous Referee Comment #2 on "AQ-Bench: A Benchmark Dataset for Machine Learning on Global Air Quality Metrics" by Betancourt et al.

General Comments

This paper has the primary goal of presenting a dataset combining several long-term metrics related to surface ozone concentration together with explanatory parameters (metadata) related to the ozone measurement stations, which can be used as inputs to a machine learning framework to predict the metrics. It also presents a benchmark application of several simple machine learning algorithms to this problem as an illustrative example. Overall the dataset and its goal are very useful, but I believe several improvements should be made prior to its publication.

Answer: We want to thank the Referee #2 for the valuable comments and questions. We appreciate the opportunity to improve our study. In the following, we address each of the comments individually.

The goal (as summarized best, I think, at the start of Section 4) is to produce estimates of various long-term ozone metrics based on local site conditions, rather than (for example) short-term estimates based on atmospheric characteristics (temperature, wind, etc.). I think this should be more clearly and explicitly stated up-front, ideally in the abstract, so the specific goals for which this dataset has been created are more immediately clear to a reader.

Answer:
We added a sentence in the abstract (line 8f in the marked-up manuscript), to clarify the goal of this dataset: '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.'.

Although it is perhaps beyond the scope of this work, I believe that the introduction should also make mention of the use of machine learning in other areas related to atmospheric chemistry and air quality. Machine learning techniques are often used as emulators or surrogate models for more computationally complex components of atmospheric chemistry models, i.e., to replace complicated and costly atmospheric chemistry calculations with simpler machine learning surrogates to improve computational performance of the models. Second, machine learning is being extensively used in the calibration of low-cost sensors for air quality, in order to account for the many sources of interference with the measurements of these sensors and allow them to more effectively supplement the existing monitoring networks for pollutants such as ozone. Although such use of machine learning is beyond the scope of the present work, I believe it is still worthwhile to draw attention to these other areas of ongoing work combining machine learning and atmospheric science.

Answer:
Thank you. These applications of machine learning in atmospheric chemistry are certainly interesting and worth to mention. We added two sentences in the introduction pointing to studies on parameterization and low cost sensors: '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 (e.g. 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 (e.g. Schmitz et al., 2021; Wang et al., 2020).' (Section 1, line 40ff of the marked-up manuscript).

> In Table 1, sources should be provided for these metadata, especially data on NOx emissions, NOx column, and night light intensity. It should also be made more clear what time period these represent, i.e., are they the average over the entire time interval, or more representative of the situation in one particular year. I suspect that the answer will vary by dataset; for example, population density might be based on census data from a particular year, while NOx column density might be derived from a satellite, and therefore represent an average (although only for clear-sky periods during which the satellite is overhead) over many years. Details do not necessarily need to be provided for each underlying dataset, but the initial source for each data should be made clear so the reader can look into the details.

Answer:
As referee #2 assumed correctly, the answer varies by dataset. We added a table in the Appendix A where more information on the geospatial data is given, including the data sources for more background information if needed. This table also contains links to the data documentation, which provides further information for the interested reader. We reference this appendix in Section 4.1 where the metrics of the datasets are described, line 202f of the marked-up manuscript.

> In Table 2, there seems to be an inconsistent definition for data completeness applied across these metrics (e.g., the overall average requires only one valid datapoint, percentiles require 10, and daytime and nighttime averages require 75% completeness). I would suggest, at a minimum, that a "completeness" metric also be included in the data set, indicating what fraction of the expected total number of hourly measurements for each site were actually present in the dataset. This would allow users to potentially make different decisions about what values they consider to be "valid" to their applications. Furthermore, if the datasets are more complete towards the beginning or end of the time period in question, this could combine with the fact that the underlying metadata (e.g., population density) might be changing over time, and be a confounding factor in the performance of the machine learning approaches. While this is a complex issue to address, some mention should be made of this limitation on the data which a single "completeness" metric will not capture.

Answer:
We agree that more clarity is needed regarding the data entering the AQ Bench summary statistics. However, we see relatively little value in a long table with the absolute number of samples included. There are two reasons: First, we apply data capture criteria which are established in the ozone research community. When these criteria are met, the data is considered statistically robust and

thus valid. We do not think more details are needed for the users of a benchmark dataset. Second, even if we would give the data completeness of every data point, it is still an open field of research how uncertainty induced by missing data propagates through metrics calculation (Section 3.3 of Lefohn et al., 2018, as cited in our manuscript). Therefore, we are of the opinion that such a table is not needed by the user. Instead we have added a reference to the data capture criteria in Table 2, which were used in compiling our dataset in Section 4.2, line 224f of the marked-up manuscript 'A summary of all metrics and their data capture criteria is given in Table 2.', to clarify the data capture process. We have also added more details on percental data capture of *ozone_average_values* and a graph to clarify the data capture criteria of an exemplary ozone metric (*ozone_aot40*) in Appendix B. We reference this Appendix in the dataset description part, in Section 4.2, line 224f of the marked-up manuscript.
Concerning the potential source of errors for unequally sampled data we added some discussion on this in Section 6.2: '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 urbanisation.' (line 320ff in the marked-up manuscript).

> In section 5.1, a particular machine learning task and evaluation metric is introduced. While this is useful for benchmarking purposes, it may not be appropriate in all cases, depending on the application. For example, determination of whether metrics fall into certain discrete classification regimes, e.g., "healthy" or "unhealthy", may be a desired goal in certain applications. While it is impossible to account for all potential uses of these metrics, some mention should be made that this is only one of many possible machine learning goals and evaluation metrics.

Answer:
Thank you for these suggestion. We added a sentence to the Conclusions which reads: "Further applications of AQ-Bench could be developed such as a classification of ozone sites into 'healthy' or 'unhealthy'."

> **Specific Comments**
>
> Line 36: O3 and Ozone are both used; this may be redundant.

Answer:
Yes, 'O_3' is redundant. Section 1, line 37f of the marked up manuscript now reads: '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.'.

> Line 79: Starting the sentence with "I.e." seems awkward to me; I would suggest rephrasing this.

Answer:
We rephrased this sentence. Section 2.1, line 85f of the marked up manuscript now reads: 'Agricultural fields, forests, and grasslands therefore yield different magnitudes and seasonal cycles of VOC emissions (Simpson et al. 1999).'.

Lines 105-106: suggest revising to "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)."

Answer:
We applied the correction as suggested by referee #2 (Section 2.3, line 114f).

Line 110: suggest removing "exemplary".

Answer:
We removed "exemplary" (Section 2.4, line 118f).

Lines 127-128: This sentence is awkward; I am not sure what the correct phrasing should be, but the authors should consider revising this sentence.

Answer:
We rephrased the sentence and hope that it is now clear for the reader. Section 2.4, line 136ff of the marked up manuscript now reads: '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).'.

Line 132: suggest revising "ground-level ozone levels" to "ground-level ozone concentrations" to avoid the repeated word.

Answer:
We revised the sentence as suggested (Section 2.4, line 142).

Line 169: It is unclear whether "mean ozone metrics" refers to the average values of different metrics across the given time period, or simply the average ozone concentrations in different locations. This may need to be clarified.

Answer:
We rephrased the sentence as follows: Section 4, line 184f of the marked up manuscript now reads: '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'. We scanned the manuscript for other occurrences of this ambiguity, and rephrased them where necessary. Section 3.1, line 166 now reads '[...] also basic statistics such as average, median and percentiles are available [...].'.

Lines 233-236: It is unclear how these clusters are used. Are entire clusters assigned to one of the three data subsets, or are individual members of the clusters divided between subsets, such that each subset will get at least one sample from each cluster? Also, how is

the issue of sparse measurements in South America or Africa, for example, addressed using this spatial clustering approach?

Answer:
Clustering was used to prevent putting strongly correlated measurements from neighboring stations into different datasets as this would lead to overfitting and overly optimistic evaluation of the machine learning results. This is why we always sort one cluster into one of the three datasets. Stations in sparsely covered regions will generally not be clustered as their distance is usually larger than the 50 km applied as selection criterion. Section 5.2, line 252f of the marked-up manuscript: '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 (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.' We hope this explanation is sufficient to understand our approach.