Response to the anonymous Referee #1 Comments on “AQ-Bench: A Benchmark Dataset for Machine Learning on Global Air Quality Metrics” by Betancourt et al.

> This study introduced a AQ-Bench dataset for machine learning. This dataset includes aggregated air quality data of more than 5500 air quality monitoring stations originated from Tropospheric Ozone Assessment Report (TOAR). This is an interesting dataset that contains both data of tropospheric ozone and data of many ozone-related factors, which enables a quick start of machine learning with the AQ-Bench dataset. I would suggest publication after the following issues are addressed.

Answer: We appreciate the comments and suggestions from Referee #1 and are grateful for the opportunity to improve our study. All questions and comments are addressed individually below. Since some comments cover multiple aspects, we have split them into subsections (a), (b) ...

> 1. My main concern is that this dataset is a five-year averaged ozone data and does not include any time series information of ozone in the final data product. (a) An average over such a long time would filter out a lot of important information. (b) I am afraid that such a small dataset may limit the applications for many machine learning methods to investigate the complex and nonlinear relationship between ozone and other factors. (c) Moreover, because there is no time series information, it also makes it difficult to well verify the trained model. For example, can the model built using these 5 years of data be used for another period? Therefore, even for climate-scale analysis, I recommend adding some time series data to this dataset.

Answer:
AQ-Bench has been designed as a low-barrier entry dataset for atmospheric scientists wanting to make first experiences with machine learning techniques, and machine learners who want to familiarize themselves with atmospheric data. It therefore focuses on the analysis of spatial patterns. Of course, the reviewer is right that ozone levels are influenced by time-dependent phenomena, even on longer scales, and AQ-Bench can therefore only capture a snapshot of the relations between geospatial data and multi-annual ozone metrics. However, we believe that there are enough interesting aspects in the dataset as it is, and the addition of time-varying features would complicate the analysis and add extra computational burden to the machine learning solutions. Furthermore, there are many published studies on time series analyses with machine learning techniques, and they sometimes come with easily accessible code and data (e.g. Kleinert et al., 2020 as cited in our manuscript). We are developing the air quality benchmark dataset concept further and are planning to produce a benchmark dataset for learning spatiotemporal relations on much larger scales at a later stage.
(a): Filtering out information in the time domain is a valid approach when studying environmental relationships on longer time scales. The metrics available in JOIN can be queried over the time span of one year, as it makes physical/scientific sense to consider one value that is representative for the ozone distribution at a site over this period. Averaging over five years makes the values more robust to drift and interannual changes. By filtering out the time resolved information, we emphasize robust, static features, which can be analyzed on the global scale. We added the sentence ‘Furthermore, through a longer aggregation period, we emphasize robust, static features. This

aggregation reduces the size of the dataset and makes a global coverage possible. Due to our focus on spatial relationships we consciously ignore time-resolved patterns.' in the limitations part (Section 6.2, line 331ff in the marked-up manuscript).
(b): AQ-Bench does not aim to enable the exploration of "[all] complex and nonlinear relationship(s) between ozone and other factors", but instead defines one specific research task which can be addressed with a variety of machine learning techniques. We understand that our ambitions could have been overstated a bit in the abstract and introduction and therefore added the sentence 'The purpose of this dataset is to produce estimates of various long-term ozone metrics based on time-independent local site conditions.' in the abstract, line 8f of the marked-up manuscript. We furthermore added the sentences '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 the limits part (Section 6.2, line 325ff in the marked-up manuscript).
(c): The reviewer raises the valid point of generalisability here, which is, however, not in scope of AQ-Bench. The target of AQ-Bench is to obtain the best possible coefficient of determination for various ozone metrics in relation to the predictor variables. This requires no further verification as such. Of course it would be interesting to broaden the task to include temporal changes in predictor variables, but this has to be the subject of a different study.

> 2. More efforts are needed to show the quality control processes and the error estimation of AQ-Bench dataset. Although it have been stated that the data originates from the TOAR database, I think as a data paper, it is necessary to provide more complete information about data sources, quality control and data accuracy. (a) For example, how many observation samples were used for the 5-year averaging at different sites? (b) Are the emission data reference to a base year or the average of the five years?

Answer:
Thank you very much for drawing our attention to this point. In general, the quality control of ozone data in TOAR is in the hand of the air quality networks, and data as well as metadata were also quality controlled by TOAR (Schultz et al, 2017, as cited in our paper). We have added a couple of sentences regarding quality control in TOAR in Section 4, where TOAR data products are introduced. More specifically, line 156ff in the marked-up manuscript now reads '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).'. Line 177ff reads '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.'
(a) 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 hourly ozone 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.
(b) We added a table in the Appendix A where more information on the geospatial data is given, including the year, and the data source. We reference this appendix in the part where the metrics of the datasets are described, in Section 4.1, line 202f of the marked-up manuscript.

> 3. The AQ-Bench dataset was validated through using three machine learning methods with a defined task. (a) As a machine learning dataset, I think the first step is to verify the reliability of the data itself and then its usability in machine learning. (b) For the machine learning tests, how to judge the quality of the dataset based on the training results? Is there a well-defined standard for that？(c) Additionally, more explanation are needed for the purpose and significance of the task mapping from metadata in Table 1 to the ozone metric values in Table 2.

Answer:
We assessed the issue of verifying the reliability and quality of our data carefully. We hope that our answers and changes regarding remark 2 have made this clearer.
(a) The data used in AQ-Bench is reliable because we are using TOAR data which was already verified, and has been the basis for many studies, e.g. the Tropospheric Ozone Assessment Report (TOAR, as cited in our manuscript). We hope that the reference to TOAR and our corrections in regard to remark 1 suffice as a verification for the reliability of the data. Regarding the usability, we have added a summary in the conclusion, Section 8, line 363ff of the marked-up manuscript: '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.'.
(b) For a first estimation of the suitability of the data for machine learning, it is sufficient to use a standard method and a suitable evaluation metric with a standard train-validation-test split (Section 5). The metric for success is clearly defined as it is common practice in machine learning applications. We chose to use the R2 score which has to be larger than zero. We mention up-front that the dataset is validated by the machine learning baselines: 'Baseline scores obtained from a linear regression method, a fully connected neural network and random forest are provided for reference and validation.', abstract, line 10f of the marked-up manuscript. We mention that besides these first baselines, the dataset and machine learning results can be further validated by cross-validation: 'For further evaluation of machine learning results, cross validation can be applied.', Section 5.1, line 242 of the marked-up manuscript.

(c) We added a detailed explanation in the discussion on the purpose and significance of AQ-Bench in, Section 8, line 367ff of the marked-up manuscript: ‘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.’ In this sense, we liken AQ-Bench to the famous *MNIST* dataset, which played a major role in the early stages of the development of machine learning techniques for image classification. The AQ Bench dataset is simple, but it still highlights several specific challenges of atmospheric data analysis, which differ from most other classical machine learning problems (cf. Schultz et al. (2020) as cited in our manuscript). We added one additional remark on the purpose of the dataset in the abstract (line 8f of the marked-up manuscript): ‘The purpose of this dataset is to produce estimates of various long-term ozone metrics based on time-independent local site conditions.’.

> 4. The performance of a machine learning method depends on the configured parameters of the method in the experiment. The manuscript shows the configurations (e.g., 100 trees in RF) of different method in the baseline experiments, but it does not explain the reason for adopting such configurations. At least, some discussions on this issue are needed.

Answer:
It is common to use standard machine learning algorithms (the so called ‘vanilla’ methods) with hyperparameters suitable for the given task as baselines for a benchmark dataset. This is done to show that the dataset is suitable for machine learning. The choice of machine learning methods and optimization of hyperparameters is then left open to the users. For the random forest we used 100 trees. This is a common choice producing good results on the AQ-Bench dataset. For the neural network, we chose a shallow architecture based on our empirical studies. The reported learning rate and L2 regularization parameter were established through random search. We added more justification and explanation on hyperparameters. In the baseline section (Section 5.3), we rewrote two sentences. Line 272f of the marked-up manuscript now reads: ‘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.’. Line 275f of the marked-up manuscript now reads: ‘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 optimisation criterion.’ In Appendix D, line 620f of the marked-up manuscript, we added the sentence ‘They are determined from empirical studies and random search. ‘In the conclusion (Section 8), line 361f of the marked-up manuscript, we added: ‘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.’

> 5. Table 3, please clarify if this is the result for the validation datasets?

Answer:
Table 3 shows the results on the test set. To clarify this, we changed the caption of Table 3 to: ‘R2-scores of the test set in %. [...]’.