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.

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. An average over such a long time would filter out a lot of important information. 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. 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.
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. For example, how many observation samples were used for the 5-year averaging at different sites? Are the emission data reference to a base year or the average of the five years?
3. The AQ-Bench dataset was validated through using three machine learning methods with a defined task. 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. 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? 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.

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