Referee comment on "Water quality, discharge and catchment attributes for large-sample studies in Germany – QUADICA" by Pia Ebeling et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-6-RC2, 2022

With this manuscript the authors describe in great detail a dataset that combines several water quality and quantity related variables which covers the extent of Germany.

In my opinion, the manuscript is very well written, follows a clear and understandable structure and contains all the necessary information for someone to comprehend and use the described dataset.

I think that the dataset has a high scientific value and as the authors state can have multiple applications in environmental sciences. In fact, I believe that the authors could emphasize the main advantages of the dataset, that are the large temporal and spatial coverage of actual measurements and the inclusion of both water quality and quantity data along with drivers, which facilitates the hypothesis testing and finding environmental relationships. Overall, I think that the manuscript is worth publishing as it will help promote the dataset. Perhaps it will inspire other researchers to compile actual field data into large datasets and make them accessible too. Below there are a few minor comments and suggestions for improving some parts of the manuscript.

L40-43: It is not very clear to me why machine learning techniques are highlighted here as a tool for finding relationships between environmental variables or defining patterns. Also I am not sure that machine learning is the best option for hypothesis testing. My point is that there are many options for data analysis. Perhaps this is relevant with a previous use of the presented dataset?

L180: It would benefit the manuscript if a few details about the methods used for the quantification of water quality parameters are included, at least in the Appendix. It could be useful for the user of the data to be able to know this information.
L185: I think it would be optimal to not exclude outlier values from the dataset. Given that these values are not errors and that fall within a possible range of values, excluding outliers could miss extreme events like very large floods. Actually it might be of interest for some researchers to identify anomalies in time series and how these changes across temporal or spatial scales.

L245: There are two different sources of water quantity data. Gauges and I guess field measurements that were taken in parallel with water samples. Is there any statistical difference of the medians between the two types of measurement?

L271: *provide* is repeated in the same sentence