Reply to comments from Referee 1 of the preprint in ESSD “Water quality, discharge and catchment attributes for large-sample studies in Germany - QUADICA” by Ebeling et al.

RC1.1: Overall, I find this manuscript and dataset to be valuable and accessible. However, I suggest some revisions and changes that I believe will improve the clarity and usability of this dataset. I present my suggestions for some revisions to the text of manuscript and to the content and presentation of the dataset below:

We thank Reviewer 1 for the positive assessment of our work and the helpful suggestions. We address the individual comments below with responses in blue for clarity.

Manuscript Comments

RC1.2: Line 44 While I absolutely agree that the compilation and dissemination of high quality, comprehensive datasets is valuable, data-driven science has always been and will always be constrained by data availability. Therefore, I find statements such as *“… harmonized and quality controlled large-sample water quality and quantity data are still not widely available”* to be subjective, difficult to evaluate, and unnecessary. I suggest instead emphasizing a more specific description of the significance of this dataset in the context of other large hydrologic data sources, which the authors do elsewhere in the introduction.

Thank you for this suggestion, we have changed the formulation to emphasize more the significance of large-sample data sets.

Line 41-46 of revised manuscript, changes in bold:

“**The** collection and availability **of water quantity and quality data are** steadily increasing **with technological advances (Rode et al. 2016), but particularly** harmonized and quality controlled large-sample **data sets of** water quality and quantity **along with catchment attributes are needed. These enable identifying and characterizing water quality and quantity response patterns and relationships with potential controls, facilitate hypothesis testing and thus advance** our understanding of the complex coupled hydrological and biogeochemical systems across larger samples and **domains** (Li et al., 2021).”

RC1.3: Line 57 Awkward language, consider rephrasing. I suggest *“… recent large-sample water quality studies have* provided a basis for increasing *our understanding of catchment functioning...”*

Thank you for this suggestion. We changed it to “provided a basis for enhancing our understanding”.

RC1.4: Line 70 In addition to their utility in addressing the questions raised here, large-sample, high quality, accessible datasets can also support uses that are un-anticipated by their authors. I think that this benefit of large-sample datasets is worth mentioning in this paragraph, and I have some suggestions below for ways to achieve this (in general, to provide a curated dataset while preserving all information, even information that may not seem useful today).

We thank the reviewer for their thoughtful remark. The text referred to gives examples of questions that still remain unanswered and might be interesting to study with this data set. We agree with the reviewer that these are not exclusive. The reviewer suggestions have been very helpful and we have incorporated them to a large degree (see below).

RC1.5: Line 137 I appreciate the clear description of the inclusion criteria used, yet I would appreciate a more detailed description of the criteria for outlier removal.

The preprocessing steps and the outlier test was described later in the manuscript, which we have now moved to this location (line 134, revised manuscript).

RC1.6: Line 145 This river network would be a valuable inclusion in the dataset. While the end user can create an approximation bu using similar parameters (100m DEM, D8, and a 10m burn in), the quality control and manual adaptations described here make a product that is unique to this analysis. Having access to this river network could support more additional analyses that are currently not possible, and that might depend on the exact alignment between the river segments, sampling stations, and catchments.

Thank you for this suggestion. We agree that the provision of the processed data sets will increase the usability of the data set for users. Therefore, we have added additional data sets to the data repository:

- DEM 100m
- Flow direction raster used for catchment delineation
- Flow accumulation raster
- Modified River network after manual adaptations used for burning into DEM
- Modified station locations consistent with the flow accumulation grid with a 100m snapping distance.

The new version of the repository can be accessed via http://www.hydroshare.org/resource/88254bd930d1466c85992a7dea6947a4 (note that DOI provided in the manuscript becomes valid only after publication). We also added a sentence to the manuscript (line 161 in the revised manuscript).

“The used DEM, flow direction and flow accumulation raster as well as the modified station locations and river network are also provided in the data repository for further use.”

RC1.7: Line 186 To my understanding, there is no ‘confidence interval’ associated with this method for excluding outliers. If the distribution of the data is correctly represented by the log-normal model, than 1 of 10000 values would be expected to exceed than the specified threshold. Given a large enough dataset (which we have here!), the presence of such values would be expected, even in the absence of any errors that would warrant the exclusion of such data points.

Further, because extreme concentrations of a solute are likely to result from uncommon mechanisms which are not likely to be accounted for in any general distribution model that describes the ‘normal’ behavior, I am skeptical of the use of such distribution models to identify outliers. For example, in my region, there is a small lake which hosts an enormous and unusual population of migratory geese for a few days a year. Whether examined across space or time, macronutrient concentration values from this circumstance appear as outliers, yet in fact they may describe this rare event accurately. The exclusion of this extreme data would appear reasonable to anyone not familiar with this particular circumstance, yet would do a disservice to future users of the dataset.

Unfortunately, I have no perfect method for separating unusual ‘outliers’ from erroneous ‘outliers’. Instead, I suggest that the complete raw dataset should be provided, to allow future users the freedom to develop their own approach to this issue, or to specifically examine the characteristics of this extreme data. If possible, this raw data could be accompanied by a ‘QC’ column indicating the result of the authors entirely reasonable but necessarily imperfect inclusion criteria.

Thank you very much for your considerations. We agree with the referee that raw data would allow most freedom in their own choices and transparency for the users. Unfortunately, we are not allowed to provide the raw data at this point in time as the individual federal states providing the data to us did not agree on handing out raw data to others. Hopefully the condition of use for the data will change in the near future to fully support open data and science. We are aware of the risks of misinterpretation of extreme versus erroneous samples arising from outlier tests. In this case, we think that the issue of outliers is not of major importance regarding the robust aggregated metric (median) that we provide. The aggregated data unfortunately does not allow analysis of such extreme events, which the raw data would. We like the referee’s idea of flagging detected outliers in the raw data, but under the given circumstances, we still provide the number of excluded data points per time series to allow users to exclude stations from certain analysis using this criterion.

The outlier test applied uses the mean concentrations and standard deviations in logarithmic space. We then use a very high confidence level (note that this is *not* a confidence interval; but a low significance level) to only exclude very extreme outliers and be tolerant towards less extreme values compared to a lognormal distribution. A confidence level of >99.99% describes exactly what the referee said: the probability of <1/10000 falling beyond the threshold of the log-normal distribution estimated from the samples. Although there are indeed many samples in the dataset (>10000), the samples of the single time series for which the test is applied are less (median number of samples around 150, see Table 2). Therefore, we think that the adopted method for the outlier detection is reasonable, and more so as we focused on providing robust metrics. We removed the explanation from line 185 (submission) referring now to line 134 (revised manuscript, see also RC1.5).

RC1.8: Line 214 I recognize the value of the WRTDS analysis, but I think that the data underlying this analysis is more valuable than the analysis itself in this context. Is all the data that underlies this analysis is present in this dataset? I believe it is, but I would like to see a clear statement to this effect.

We agree that the raw data would be of high value, but as stated above it is not possible to provide them due to existing legal concerns (see our reply to RC1.7). The input to the WRTDS models is the same preprocessed data as described under Section 2 for the station/catchment selection. The raw data are deposited in the institutional long-term repository (Musolff et al. 2020). The idea was to provide data with a higher temporal resolution (monthly) for stations that allow the application of WRTDS. These aggregated data can be used, for example, to estimate trajectories of seasonality and to distinguish trends by seasons (e.g. summer which is more relevant for eutrophication risks). One example is a previous study, Ebeling et al. (2021). We have now explained more clearly the underlying data of the WRTDS models (line 219-226 of the revised manuscript).

Changes in bold: "For **the subset of** stations with high data availability, a Weighted Regression on Time, Discharge and Season (WRTDS; Hirsch et al., 2010) was applied using the R package EGRET (version 3.0.2; Hirsch and De Cicco, 2015). We refer to these stations as 'WRTDS stations' for short. WRTDS represents long-term trends, seasonal components and discharge-related variability of the water quality variables (Hirsch et al., 2010). The criteria **for a WRTDS application were** checked for each station and compound separately **using the preprocessed data as described in Section 2. The criteria** were a time series of at least 20 years **length**, at least 150 samples of water quality, no data gaps larger than 20 % of the total time series length and a complete time series of daily discharge **(see also Section 3.2.2)**."

RC1.9: Line 277 The relationship between these gap-filling and bias-correcting methods and the dataset is unclear. Are data from these methods included in the dataset, or only used in fitting the WRTDS models? If the 'corrected' data are included in the data tables, I think they should be identified as such.

Thank you for the careful reading, we clarified this part now. The gap-filled discharge data are used as input to the WRTDS models and are included in the WRTDS output only, i.e. this also includes the discharge data provided with WRTDS concentration and load estimates. The gap-filling is not included in the provided data of annual observed median discharges. We added this note to section 3.2.2, line 288:

"Note that the gap-filled discharge time series are used for the WRTDS models only. This includes the monthly and annual discharge data provided with the WRTDS data tables (as described in Section 3.1.2)."

RC1.10: Line 307 I remain unsure of the N sinks included in this calculation. Crop harvest is mentioned as an N 'output', and I see no other sinks mentioned. This should be clarified.

Our data set provides the soil surface N budget (N surplus) based on the data of Behrendt et al., (2003) and Häußermann et al. (2020). The latter data sets consider as only N output (sink) withdrawal from harvested crops, which includes also harvest of fodder crops. In the revised manuscript, we use systematically the terms "N inputs" and "N outputs" rather than "N sources" and "N sinks" for clarity and consistency with the terminology used in previous studies (e.g. Behrendt et al., 2003; Häußermann et al., 2020). (e.g. line 313)

RC1.11: Line 345 N deposition on imprevious urban surfaces is not counted as a diffuse N source, but I do not see where is it accounted.

The assumption is that N deposition on impervious urban surfaces is not a diffuse source to the catchment but a point source. We assume it is discharged to the sewer system with rain water and transported to the wastewater treatment plants. Therefore, we do not include this component in our N surplus data, but rather we assume that it is included in the point source N load data from wastewater treatment plant (Sect. 4.3). We recognize that there are alternatives for the fate of the deposition in urban areas, e.g. collection in separate sewer system, or sewer overflow which can occur when the discharge exceeds the capacity of the pipes or treatment plant. However, the partitioning of the N deposition between these different pathways is uncertain. A more exact estimation of the fate of N deposition can be considered in the future, but is not within the scope of this work. In the revised manuscript, we clarify this point (line 354-357).

Changes in bold: "Deposition on urban sealed surfaces was neglected, since we assume this component is collected by the sewer system and **transported to the wastewater treatment plants. We thus assume it** is not a diffuse N source **but part of the point sources (Section 4.3). In contrast,** deposition on urban grassland like public parks was considered."

RC1.12: Line 372 Although they may be beyond the scope of this dataset, I suggest that attributes of the rivers may also provide valuable information. Relevant attributes include riparian or floodplain development (urban or agricultural), geomorphic context (e.g., valley confinement), and the presence of absence of impoundments.

Thank you very much for the suggestions. We agree that these are useful additions, which we might consider for a future extension of the data set. For now, they go beyond the scope of this manuscript.

RC1.13: Line 524 When allowable, I suggest that the inclusion of dis-aggregated (raw) data is worthwile. However, I recognize that the limitations mentioned here may describe much of the source data for this dataset product. I suggest that the authors make and effort to include as much raw data as possible.

Unfortunately, the data limitations exist and raw data cannot be provided until now (see also our response to RC1.7). This is the reason, why we decided to provide the data in aggregated form as ready-to-use metrics together with all the other catchment data.

Dataset Comments

RC1.14: I am able to load, combine, and manipulate the two spatial datasets and all of the .csv data without issue. I appreciate that the catchment polygons overlap, with each polygon representing the complete catchment associated with a station. I also appreciate the clear OBJECTID field, usable to join attributes among the catchments, stations, and tabular data. I also appreciate the description of the various columns in the metadata document, and the consistent units used between fields.

Thank you very much for the checks, we are happy that the data set was easily accessible and understandable.

RC1.15: I may be out of touch with GIS data norms, but I consider the shapefile format to be antiquated and limiting. If the spatial data were instead presented as a geopackage, any limitation on column names (and the number of columns) would be removed, which would aid in the analysis of this comprehensive dataset. A geopackage is also an open, non-proprietary format.

Indeed, there are advantages of the geopackage format, but we think that shapefiles are still widely used and known and the efforts to use them are low from various platforms (R, QGIS etc.). One of the critiques about shapefiles is the multi-file type which should not be problematic as we provided it in a long-term repository and it can be redownloaded at any time.

RC1.16: The naming convention is generally consistent between files, however the concentration tables describe the number of observations with a 'n_' prefix, while the source table describes N concentrations with a 'N_' prefix. These are easily confused. I suggest renaming the columns in the source table to use a '_N' suffix instead (N_total → total_N).

Thank you for this suggestion, we have decided to keep the naming convention as is. We think that the risk to confuse is relatively small as the source data with "N_" prefix are of the double-precision data type and the columns with "n_" prefix are integers, the magnitudes of values differ and all the column meanings of the different files are explained in the metadata. The naming convention as is, is also consistent with the names of the catchment attributes for nutrient sources with the capital letter of the nutrient at the start.

RC1.17: I also suggest renaming the 'attributes' csv file to 'catchment_attributes' to clarify it's affilitation.

We see the point raised here. We have changed the naming as suggested by the reviewer.

RC1.18: If possible, I suggest that the individual monthly concentrations (in addition to the included median concentrations for each month across all years included in the dataset) would add value to this dataset.

Thank you for this interesting idea. Unfortunately, we cannot provide the raw data (as explained above), but we understand that it is valuable to provide a metric for the long-term variability around the monthly median, so that we decided to additionally provide the 25th and the 75th percentiles. The number of the sample size, which is also provided, allows the user to evaluate the robustness of the estimate and define own quality criteria for values to include. We now provide these additional values for each month across the whole time series in the repository and also added this information to the text of the manuscript in corresponding locations. The updated repository can be accessed under http://www.hydroshare.org/resource/0ec5f43e43c349ff818a8d57699c0fe1 (note that DOI provided in the manuscript becomes valid only after publication).

Line 247 in the revised manuscript: "We also include the 25th and 75th percentiles to reflect the long-term variability of a given month."

Line 293 in the revised manuscript: "… we provide long-term monthly median discharge **and the 25th and 75th percentiles** over the whole time series …"

RC1.19: Upon examination, I found one oddity in the c_annual table. I calculated the fraction of total N present as NO3-N, and found some values exceeding 1 (indicating more NO3 than total N). Most of these values were barely greater than 1 and likely due to normal measurement error, yet two had much higher values (one of 2.1 and one of 7.8). I suggest examining these values in the context of the scheme for identifying outliers, and considering a refined approach which flags suspected erroneous records but avoids the challenges associated with the absolute exclusion of outliers. A similar test with P values found evidence of normal measurement errors, but little cause for concern. All other data that I examined appeared reasonable (and interesting!).

Thank you very much for the thorough examination of our data tables. There are several reasons that could cause such deviations: 1.) different number of samples for NO3 and TN, 2.) deviations can result as we are comparing median values instead of single samples, 3.) different sampling times. We therefore revisited the c_annual data table. All examples with median NO3N > median TN have a larger sample size for nitrate compared to total N, i.e. n_NO3 > n_TN for the given year and station.

For the most extreme case mentioned by the reviewer the number of observations n_NO3 is 12 while n_TN is only 8 (OBEJCTID=2, year 2006). The implausible fraction of nitrate in total N thus arises from differences in the frequency of sampling and corresponding differences in the medians. For the single corresponding samples TN is larger than NO3N (see Figure). For the context of outliers please refer to RC1.7. We added a sentence about this to the text (line 205 in revised manuscript).

"Note that the number of samples underlying the median values can differ between the different nutrient species so that the fraction of TN present as NO3-N or TP present as PO4-P may show inconsistencies for single stations (e.g. values above 1)."

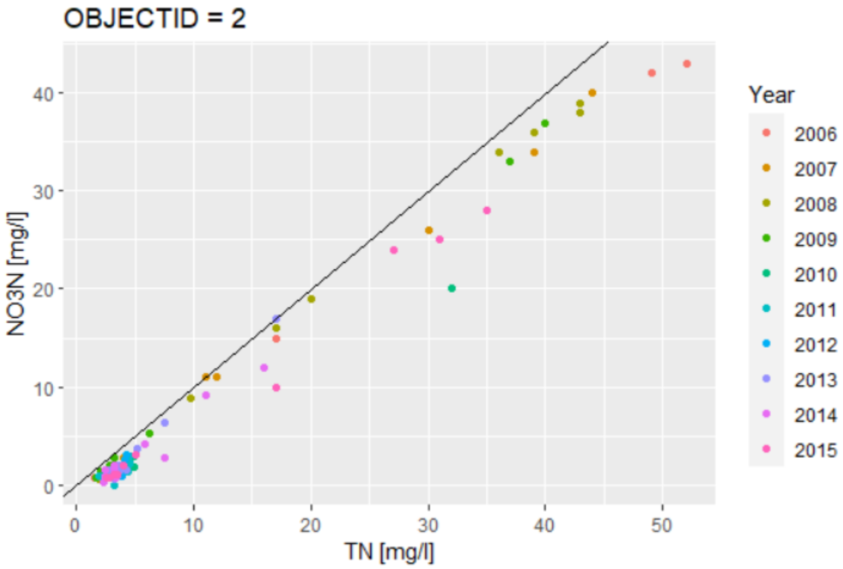

Figure 1: TN and NO3-N concentrations of samples from station with OBJECTID=2.

Reply to comments from Referee 2 of the preprint in ESSD "Water quality, discharge and catchment attributes for large-sample studies in Germany - QUADICA" by Ebeling et al.

RC2.1: 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 highlscientific 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.

We thank Reviewer 2 for the positive feedback on our work and helpful comments. We address the individual comments below with responses in blue. Following suggestions from the first reviewer (RC1.2), we adapted the introduction to more specifically emphasize the advantage and need for such a data set. In the conclusions we also more specifically emphasize the advantage as suggested here. In the abstract we also highlight the large spatial and temporal coverage now.

Line 560 in revised manuscript, changes in bold: "In this study, we provide a comprehensive homogenized data set **with a large spatial and temporal coverage of both water quality and quantity observations along with catchment attributes. Specifically, the data set includes** time series of water quality, co-located discharge, hydroclimatic data and diffuse nitrogen inputs, as well as catchment boundaries and more than 100 catchment attributes **for 1386 German catchments**."

RC2.2: 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?

We agree that this statement was too specific here regarding the variety of data analysis options that exist. We removed the sentence.

RC2.3: 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.

Unfortunately, the information on sampling methods of the water quality parameters is not available to us as this was not provided with the metadata by the federal states. Given the large temporal and spatial coverage of the data and their different sources, differences in the methods over time and between the federal states are possible. However, it is worth noting that these data are also used to report the water quality for the Water Framework Directive at the EU level and that the laboratories commit to analytical quality assurance following the legal regulations for surface waters (OGewV - Oberflächengewässerverordnung) and DIN EN ISO/IEC 17025, e.g. the River Elbe basin management refers to it (see page 130, https://mluk.brandenburg.de/w/WRRL2022-27/Bewirtschaftungsplan/FGG-Elbe-Bewirtschaftungsplan-2022-2027.pdf). With preprocessing, we try to exclude implausible values and address the issue of values below the detection limit, which may depend on the sampling method, but we cannot eliminate other inconsistencies here.

RC2.4: 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.

We agree that extreme values in data sets are useful for certain analyses and that is hard to distinguish between outliers as errors or extreme values. However, given the fact that we cannot provide the raw data (due to licensing issues, see also our response to RC1.7 above) such analysis would unfortunately not be possible in both cases with this data set. We decided to provide robust aggregated values in terms of median values, so that the topic of outliers becomes of lower importance. Therefore, we prefer to keep the approach as presented.

RC2.5: 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?

Indeed, there are different data sources. For some stations the discharge data was provided only for the sampling dates of water quality, however, there is no consistent information available if these are daily averages from a gauge or discharge taken exactly at the time of sampling from a gauge or even from another measuring technique. We assume that in most cases the values also come from gauges, but were not provided as daily time series. Therefore, we cannot provide the information about different possible measurement types. However, we compared the effect of data availability (continuous daily or only on grab sampling dates) on the median annual discharges of the stations with daily/continuous discharge time series (Figure A3 of the revised manuscript). This shows that median values derived from grab sample dates only give quite robust estimates. We added the R^2 and bias values of this comparison to the text and slightly reformulated it. While going through the discharge datasets again, we realized that there was a redundancy in the provided discharge data in the repository and the Figures, which might have been confusing and which we now removed. We now have the discharge calculated from grab sample dates (Q_{grab}) provided consistently and together with the concentration data and the discharge data calculated from daily time series (Q_{daily}). We added both discharge values to Figure 2 and removed Figure 3a and b, while we moved the previous Fig. 2c to the supplement Fig. 3A of the revised manuscript. Similarly, we included the discharge time series in Fig. 1A and added another column for Q_{daily} to Table 2. We hope the different origins of the discharge data are clear now and apologize for any confusion that was created.

Line 276 in the revised manuscript, changes in bold: “For stations with available daily discharge data, both annual median values of the daily data and the data from grab sample days were compared (Fig. 3c). **Our results suggest** that annual median values from grab sample dates **can be considered to be robust estimates of annual median discharge as they have a negligible** bias **(bias=-0.5 %) and low** scatter around the 1:1 line **(R2>0.99)**.”

RC2.6: L271: *provide* is repeated in the same sentence

Thank you, we removed the duplicate.