Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-6-RC1, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on essd-2022-6

Anonymous Referee #1

Referee comment on "QUADICA: water QUAlity, DIscharge and Catchment Attributes for large-sample studies in Germany" by Pia Ebeling et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-6-RC1, 2022

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:

Manuscript Comments

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.

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..."*

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).

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.

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.

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.

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.

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.

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.

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.

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.

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.

Dataset Comments

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.

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.

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).

I also suggest renaming the ‘attributes’ csv file to ‘catchment_attributes’ to clarify it’s affilitation.

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.

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!).