https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Earth System Science Data
Open Access Discussions

GRQA: Global River Water Quality Archive

Holger Virro[1], Giuseppe Amatulli[2,3], Alexander Kmoch[1], Longzhu Shen[4,5], and Evelyn Uuemaa[1]

[1]Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, Tartu, 51003, Estonia
[2]Yale University, School of the Environment, New Haven, CT, 06511, USA
[3]Yale University, Center for Research Computing, New Haven, CT, 06511, USA
[4]HyperAmp, Barnwell Road, Cambridge CB5 8RQ, UK
[5]Spatial-Ecology, Meaderville House, Wheal Buller, Redruth, TR16 6ST, UK

Correspondence: Holger Virro (holger.virro@ut.ee)

Abstract. A major problem related to global water quality analysis and modelling has been the lack of available good quality and consistent water quality measurement datasets with a global spatial coverage. Current study aims to contribute into improving the global datasets on water quality by aggregating and harmonizing five national, continental and global datasets: CESI, GEMSTAT, GLORICH, WATERBASE and WQP.

The GRQA compilation involved converting observation data from the five sources into a common format and harmonizing the corresponding metadata, flagging outliers, calculating time series characteristics and detecting duplicate observations from sources with a spatial overlap. The final dataset extends the spatial and temporal coverage of previously available water quality data and contains 42 parameters and over 16 million measurements around the globe covering the 1898–2020 time period. Metadata in the form of statistical tables, maps and figures are provided along with observation time series.

The GRQA dataset, supplementary metadata and figures are available for download on the DataCite and OpenAire enabled repository of the University of Tartu, DataDOI, http://dx.doi.org/10.23673/re-273 (Virro et al., 2021).

1 Introduction

Water quality modeling is an integral part of monitoring the health of river ecosystems and studying their interactions with the surrounding environmental conditions. Monitoring and modeling the hydrochemical properties of rivers is essential for understanding and mitigating water quality deterioration due to agricultural and industrial non-point source pollution (Krysanova et al., 1998; Leon et al., 2001; Wu and Chen, 2013). Modeling of different water quality indicators such as nutrients (Caraco and Cole, 1999; He et al., 2011), carbon compounds (Evans et al., 2005; Hope et al., 1994), sediments (Choubin et al., 2018; Ouyang et al., 2018) and oxygen (Radwan et al., 2003; Singh et al., 2009) gives valuable understanding of hydrochemical cycles and enables to estimate the effect of human influence on them.

Traditional approaches to water quality modeling consist of applying bottom-up, physically based models on the catchment level (Wellen et al., 2015). Data for model inputs is usually gathered through *in situ* observations and, more recently, automated sensor networks. Although airborne remote sensing based data acquisition methods have been successfully used to supplement field data for lakes (Chen and Quan, 2011; Toming et al., 2016), applying those methods is only viable in the case of rivers with a large enough surface area (Olmanson et al., 2013). Therefore, improving the river water quality data spatial and temporal coverage with remote sensing is limited. Significant progress has been made in improving the technical capabilities and lowering the installation and maintenance costs of the field sensors, but the spatial and temporal coverage of observation sites remains to be an issue (Pellerin et al., 2016).

In order to improve the spatial coverage of hydrological data, different solutions have been used in predictive hydrological mapping. Until recently, a common approach for predicting hydrological phenomena in ungauged catchments has been the application of already existing process-based models to catchments with similar characteristics (Hrachowitz et al., 2013; Strömqvist et al., 2012; Wood et al., 2011). These physical models usually require extensive calibration along with location-specific knowledge, which limits the wider applicability and spatial upscaling that can be done (Abbaspour et al., 2015; McMillan et al., 2012).

Recently, advances in implementing machine learning (ML) methods in hydrology have given rise to a new, data-driven approach to hydrological modeling (Mount et al., 2016). Comparison of physically based and ML approaches has shown that ML methods can achieve a similar accuracy to the physically based ones and outperform them when describing nonlinear relationships (Chau, 2006; Ouali et al., 2017; Papacharalampous et al., 2019). The recent advent of so-called physics-guided ML, which entails combining process-based models with ML methods is likely to become more applicable in the near future as well (Kratzert et al., 2019; Shen et al., 2018; Marzadri et al., 2021).

Nevertheless, a major problem related to large-scale predictive hydrological modeling has been the lack of available observation data with a good spatiotemporal coverage (Bierkens, 2015). This has affected the reproducibility of previous studies and the potential improvement of existing models (Blöschl et al., 2019; Meals et al., 2010; Stagge et al., 2019). In addition to the observation data itself, insufficient or poor quality metadata has also discouraged researchers to integrate the already available datasets. Here, ambiguities in supplementary metadata such as parameter names, units and methods of measurement has limited the use of open data for large-scale water quality modeling purposes (Archfield et al., 2015; Hutton et al., 2016; Sprague et al., 2017). Therefore, improving both the availability and quality of open water quality data would increase the potential to implement predictive modeling on a global scale. Global ML models have been already successfully used for discharge modeling (Beck et al., 2015; Gudmundsson and Seneviratne, 2015) and recent years have seen the publication of global discharge datasets (Do et al., 2018; Harrigan et al., 2020). The publication of global and continental datasets (Hartmann et al., 2014; Read et al., 2017) could make ML methods applicable for large-scale water quality modeling as well (Shen et al., 2020). However, issues related to a lack of training and validation data due to general data scarcity affects model accuracy and, therefore, limits the further adoption of ML for global water quality predictions (Chen et al., 2020).

We aim to address the aforementioned issues by presenting the novel Global River Water Quality Archive (GRQA) by integrating and harmonizing five different global and regional datasets. The resulting dataset has combined observation data for 42 different forms of some of the most important water quality parameters. Supplementary metadata and statistics are provided with the observation time series to improve the usability of the dataset. We report on developing a harmonized schema and reproducible workflow that can be adapted to integrate and harmonize further data sources. We conclude our study with a call

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Table 1. Source datasets used for compiling GRQA with their total number of observations, parameters and timeframe length in GRQA.

Dataset ID	Dataset name	Extent	Citation	Observations	Timeframe	Parameters
CESI	Water quality in Canadian rivers	Canada		28,877	2002–2018	10
GEMSTAT	Global Freshwater Quality Database	Global	Färber et al. (2018)	1,886,447	1950–2020	30
GLORICH	GLObal RIver Chemistry database	Global	Hartmann et al. (2019)	3,026,488	1942–2011	30
WATERBASE	Waterbase - Water Quality	Europe		275,068	2008–2018	19
WQP	USGS Water Quality Portal	US	Read et al. (2017)	8,689,335	1898–2020	31

for action to extend this dataset and hope that the provided reproducible method of data integration and metadata provenance shall lead as an example.

2 Data

A total of five data sources were used to compile the GRQA with two being global, one regional, and two national level (Table 1). All datasets with the exception of GEMSTAT are publicly available to download online as CSV or Excel file packages. GEMSTAT data can be requested via email. The number of available observation sites was highly dependent on the source with the Water Quality Portal (WQP) maintained by the United States Geological Survey (USGS) having the most sites. Files used during the creation of GRQA are listed in Table 2.

2.1 CESI

The first dataset included in GRQA originated from the Canadian Environmental Sustainability Indicators program (CESI) operated by Environment and Climate Change Canada (ECCC), which is a Canadian governmental department responsible for coordinating environmental policies and programs. CESI consists of water quality measurements collected by federal, provincial and territorial monitoring programs from Canadian rivers from the 2002–2018 time period (Environment and Climate Change Canada, 2020). CESI data is mainly focused on heavy metals, so only eight parameters matched the set that had been previously collected from the the other sources. It is the smallest of the five source datasets with site count ranging from two to 77 per parameter. Mean time series length per site is approximately 13 years and the average number of observations per site is 132.

2.2 GEMSTAT

The Global Freshwater Quality Database GEMStat (Färber et al., 2018) is hosted by the International Centre for Water Resources and Global Change (ICWRGC) and provides inland water quality data within the framework of the GEMS/Water Programme of the United Nations Environment Programme (UNEP). GEMStat contains over 14 million samples from approximately 11,0000 sites in over 80 countries. The data was obtained through a custom request to their data portal (International Centre for Water Resources and Global Change, 2020).

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

Table 2. Source dataset files used for compiling GRQA. WQP sites and observations were downloaded separately for each parameter and file names were assigned during the process.

File name	Size (MB)	Rows	Description	Sheet name	Source
wqi-federal-raw-data-2020-iqe-donnees-brutes-fed.csv	171.5	314,867	Observation data		CESI
data_request.xls	2.4	5,419	Site data	Station_Metadata	GEMSTAT
data_request.xls	2.4	30	Parameter data	Parameter_Metadata	GEMSTAT
data_request.xls	2.4	311	Method data	Methods_Metadata	GEMSTAT
pH.csv	21.9	372,211	Observation data		GEMSTAT
Carbon.csv	19.2	337,928	Observation data		GEMSTAT
Nitrogen.csv	65.1	1,052,823	Observation data		GEMSTAT
Phosphorus.csv	24.3	386,113	Observation data		GEMSTAT
Oxygen_Demand.csv	20.1	331,617	Observation data		GEMSTAT
Solids.csv	11.8	201,628	Observation data		GEMSTAT
Water_Temperature.csv	23.9	370,335	Observation data		GEMSTAT
Oxygen.csv	30.6	488,749	Observation data		GEMSTAT
Sampling_Locations_v1.shp	0.4	15,553	Site point data		GLORICH
sampling_locations.csv	1.6	18,897	Site name data		GLORICH
catchment_properties.csv	10.2	15,514	Catchment data		GLORICH
hydrochemistry.csv	273.3	1,274,102	Observation data		GLORICH
Waterbase_v2019_1_S_WISE6_SpatialObject_DerivedData.csv	15.1	62,288	Site data		WATERBASE
ObservedProperty.csv	0.2	888	Observation data		WATERBASE
Waterbase_v2019_1_T_WISE6_DisaggregatedData.csv	10019.2	39,121,790	Observation data		WATERBASE
WQP_*_sites.csv	2543	9,467,369	Site data		WQP
WQP_*_obs.csv	2749.8	10,088,212	Observation data		WQP

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Approximately 500 water quality parameters were available in the GEMSTAT database, out of which 30 were used when compiling GRQA. Observations cover the period 1950–2020 and mean time series length per parameter is approximately 40 years. Mean time series length per site is nine years. Site count per parameter ranges from less than ten (dissolved and total carbon) to 4,269 (total phosphorus).

2.3 GLORICH

The GLObal RIver CHemistry (GLORICH) database (Hartmann et al., 2014) is a collection of hydrochemical data from more than 1.27 million observations and more than 18,000 sampling locations across the globe. The samples originate from various environmental monitoring programs and scientific literature.

Out of 47 water quality parameters available in the raw data, 30 were chosen to be included in the GRQA. The samples cover the time period of 1942–2011, but the length of the time series is dependent on the parameter. Mean time series length per site is less than a decade for all parameters. The number of available sites per parameter ranges from just four (particulate organic nitrogen) to 8,676 (dissolved inorganic phosphorous). The dataset can be downloaded at Pangaea (Hartmann et al., 2019).

2.4 WATERBASE

Waterbase is the generic name given to the European Environment Agency's (EEA) databases on the status and quality of Europe's rivers, lakes, groundwater bodies and transitional, coastal and marine waters (European Environment Agency, 2020). The database is compiled from data sent by the national European water agencies involved in the Water Framework Directive (WFD).

Over 600 water quality parameters are included in the full dataset out of which 19 parameters were used during building GRQA. Out of all source datasets, WATERBASE had the shortest time series with observations covering only the period 2008–2018. The maximum site count per parameter is 1,976, while there were on average only around 18 observations per site. The mean time series length per site was only 1.4 years.

2.5 WQP

USGS, the U.S. Environmental Protection Agency (EPA) and the National Water Quality Monitoring Council developed the Water Quality Portal (WQP), which is so far the largest standardised water quality database (Read et al., 2017; United States Geological Survey, 2020). Although the portal also includes data from a few other countries (e.g. Mexico, Pacific islands) associated with the National Water Information System (NWIS) network, only a very limited amount of non-US samples were available. For this reason, only US national data was selected to be added to GRQA.

Due to the size of the source dataset, the full set of parameters could not be downloaded at once. Therefore, a scripted download procedure was used to retrieve water quality samples and their corresponding sampling sites separately per parameter. In the case of temperature, the data had to be further divided by state. Unlike other source datasets used in the study, the WQP

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

often had multiple versions of the same parameter available under separate codes, in case the parameter had been measured in different units, using different methods, etc. The final count of parameters used for GRQA was 31.

The longest time series of source datasets is present in the WQP with some dating back to 1898. However, the average time series length per station is just over three years. Like GEMSTAT, WQP is still being updated, so most parameters have their latest observations from 2020. Site count ranges from a single station (dissolved inorganic nitrogen) to 59,000 per parameter (total suspended solids).

3 Methodology

The GRQA compilation workflow was divided into three parts: (1) The pre-processing stage involved converting observation data from the five sources into a common format and harmonizing the corresponding metadata; (2) Pre-processed data were merged by parameter, after which outliers and time series characteristics were detected; (3) Duplicate detection was conducted in the last processing step. The Pandas (McKinney et al., 2010), GeoPandas (Jordahl et al., 2020) and NumPy (Harris et al., 2020) Python libraries were used throughout all data processing stages.

3.1 Source data preprocessing

Parameter selection. The parameters included in GRQA cover the four groups of water quality indicators outlined in the introduction: nutrients, carbon, sediments and oxygen. GLORICH was used as a reference for parameter selection due to being one of the two global source datasets and having the least amount of discrepancies within source data, i.e. each GLORICH parameter had a single matching code, unit, etc.

Parameter harmonization. Preliminary analysis showed that there were ambiguities in the parameter names, codes, units and chemical forms in the different source datasets, which has been identified as a recurring issue when dealing with multi-source water quality data (McMillan et al., 2012; Sprague et al., 2017). For this reason, lookup tables were created for each of the source datasets (**_code_map.csv*) to use as guides in the following processing stages (Table 3). The purpose of the schemas was to match parameter codes and other metadata with the versions used later in the GRQA. For most parameters, this could be done based on the literal names, remarks and descriptions in the metadata. Relevant literature and online resources were consulted for more ambiguous scenarios. One such example was total suspended solids (TSS), which can also be reported as suspended particulate matter (SPM) (Neukermans et al., 2012). Where a reliable decision could not be made (e.g. biological oxygen demand as BOD vs BOD5) the parameters were kept separate.

Unit conversion. Units of measurement were harmonized along with other metadata. All parameters except temperature (°C), pH and dissolved oxygen (%) were converted into mg/l, which was the most prevalent unit in source data. Where units were converted, observation values had to be changed as well. This was done by calculating conversion constants, which were based on both the magnitude of the source unit (e.g. μg/l vs mg/l) and the reported chemical form of the parameter. The latter affected nitrite (NO_2), nitrate (NO_3) and ammonium (NH_4) the most, as these parameters had a variety of forms in the source data that were all converted into corresponding nitrogen versions (NO_2-N, NO_3-N & NH_4-N). In some cases, the chemical form could

Table 3. Summary table of lookup table attributes.

Attribute name	Description	Data type
source_param_code	Parameter code in source dataset	string
source_param_code_meta	Additional code specification used for CESI	string
param_code	Parameter code in GRQA	string
source_param_name	Parameter name in source dataset	string
param_name	Parameter name in GRQA	string
source_param_form	Parameter chemical form in source dataset	string
param_form	Parameter chemical form in GRQA	string
form_ref	Parameter form reference	string
source_unit	Parameter unit in source dataset	string
divisor	Divisor applied to the observation value	float
multiplier	Multiplier applied to the observation value	float
conversion_constant	Unit conversion constant calculated based on divisor and multiplier and applied to the observation value	float
unit	Parameter unit in GRQA	string
source	Source dataset name	string

Table 4. Examples of unit conversion.

Parameter code	Source	Form	Source form	Unit	Source unit	x_1	M_{x_2}	n	M_{x_1}	x_2
TAN	CESI	N	NH3	mg/l	mg/l	0.106	14.007	1	17.031	0.087
NO2N	GEMSTAT	N	NO2	mg/l	mg/l NO2	0.024	14.007	1	46.005	0.007
NO3N	GLORICH	N	NO3	mg/l	µmol/l	210.268	14.007	1000	62.004	0.048
NH4N	WATERBASE	N	NH4	mg/l	mg/l	0.063	14.007	1	18.039	0.049

be identified from the source unit (e.g. mg{N}/L or mg{NO_3}/L), while others were detected by examining parameter names and method descriptions (e.g. "Nitrate, reported as nitrogen"). For other nitrogen (TKN, TN, etc.), all carbon (DOC, TC, etc.) and phosphorus (TP, TIP, etc.) parameters, the chemical were assumed to be either N, C or P even if not reported, because there is only one common element in the molecule (Sprague et al., 2017). GLORICH was the only source dataset, which needed conversion constants for carbon and phosphorus parameters as they had been originally measured in µmol/l. All WQP units matched those intended to be used for GRQA, so no conversion was needed. The formula for conversion constants was

$$x_2 = \frac{x_1 \times M_{x_2}}{n \times M_{x_1}} \qquad (1)$$

where x_1 and x_2 are observation values before and after conversion, M is the corresponding molar mass and n the magnitude difference between source and converted unit. Some examples of unit conversion are given in Table 4.

Site ID duplication. There were some instances of duplicated site IDs in GLORICH (2 site pairs) and WATERBASE (101 pairs) source data, which meant that joining observations with sites would have created duplicate time series as well. Site ID duplicates could indicate that were have been small shifts in the site location or that the site had been closed and reinstated at some point. If the distance between the duplicate pairs was less than a kilometer, only the first instance was retained in the output table. When distance was greater than a kilometer both instances were removed as metadata that could be used to make a decision (e.g. when the site first opened) was not available. Finally, all duplicate pairs were exported as separate files (e.g. *GLORICH_dup_sites*).

Coordinate conversion. CESI and WQP originally had the site coordinates in the North American Datum of 1983 (NAD83). The Pyproj (Snow et al., 2020) Python library was used for converting the North American site coordinates into World Geodetic System 1984 (WGS84) which was the coordinate system chosen for the GRQA.

Observation data filtering. Preliminary cleaning included the removal of observations of negative, missing or low quality values. In this case, low quality refers to measurements that were flagged as either coming from unreliable sources or having any kind of literal quality assessment flag in the source data (e.g. "poor quality"). Additionally, observations marked as below or above detection limit, originating from unreliable sources or otherwise suspect (e.g. unvalidated) were omitted. Three source datasets (GEMSTAT, GLORICH & WATERBASE) had this type of a quality evaluation included in the metadata. Observations from sites marked as "Not for publication" due to national legislation in WATERBASE were also not included in GRQA.

Filtration information. Where possible, supplementary information about whether a sample was filtered or unfiltered was retained as filtration can affect the sample values (Sprague et al., 2017). This information was usually available in a separate metadata column. Both "filtered" and "dissolved" were used depending on the source. GRQA includes the dissolved versions of certain parameters (total nitrogen, total phosphorus and Kjeldahl nitrogen), which originally did not exist as separate parameters in WATERBASE and WQP. In those cases, the filtered/dissolved observations of TN, TP and TKN in the two datasets were treated as the corresponding dissolved forms (TDN, TDP, DKN) in GRQA.

Time and date processing. Observations could have invalid timestamps due to formatting or entry errors, so a validity check was included in the pre-processing scripts. Dates were tested against the presumed source format and observations with incorrectly formatted or implausible dates were removed. The source datasets used different date formats, which were all converted into a common one (%Y-%m-%d). Were possible, observation time was extracted as well. A default value (00:00:00) was used to fill missing information. Time zone information was only possible to extract from the WQP. Other sources lacked time zone information, so it was not possible to determine whether the recorded timestamp was in local or Coordinated Universal Time (UTC) and the time given is up to the user to interpret.

Other metadata. Additional information about methods used or other available observation remarks in the source data were also retained. The metadata depended on the source and was available only sporadically and could not be concatenated in a reasonable way between the datasets, so the information is given in the GRQA for each source separately in the format of *source_meta_sourcecolumnname* (e.g. *GEMSTAT_meta_Analysis Method Code*). Here, the source column names were kept as they appear in raw data, e.g. spaces were not replaced with underscores.

3.2 Outlier treatment, time series availability and continuity

Time series availability and continuity. The analysis of the statistics generated during pre-processing showed that most of the time series extracted from the source datasets are very discontinuous. For example, the mean time series length per site for total phosphorus (TP) in GEMSTAT was 6.5 years and 5 years in GLORICH, while the mean observation count per site was only 55.3 and 52.4, respectively. This means that many sites have observations at a monthly time step at best. Similar findings have been previously reported about WQP time series (Read et al., 2017; Shen et al., 2020).

In order to illustrate the suspected temporal fragmentation in observation data, monthly availability and monthly continuity statistics appropriated from the strategy used by Crochemore et al. (2019) were calculated for each site in each of the merged parameter time series. Both characteristics can give insight to the granularity of the time series and can affect the applicability of different modeling methods. Monthly availability of observation data was defined as the ratio between number of months with at least one observation and the total number of months a particular site had any observations. A ratio of 1.0 would mean that there was at least one observation in every month of the time series. Monthly continuity was calculated as the ratio between the longest period of consecutive months with any measurements and the length of time series in months. Here, a ratio of 1.0 would mean that there were no months without observations and the time series is continuous on a monthly level. The resulting characteristics were added as columns in the output files.

Outlier flagging. Water quality modeling often involves dealing with numerous outliers and uncertainties in observation data, particularly when integrating time series from multiple sources (McMillan et al., 2012; Sprague et al., 2017). Due to the differences in environmental conditions and water regimes, the potential range of observation values can vary a lot between catchments. Although extreme outliers caused by faulty equipment or data entry errors can sometimes be detectable by examining distribution plots, it is often difficult to decide whether an outlier is an error or not. For example, sudden spikes in observation time series can be caused by events such as agricultural spills, which can have long-lasting effects on water quality and, therefore, should not be removed from data. However, flagging outliers can still help researchers troubleshoot potential issues at the modeling stage.

For this reason, no observations were omitted from the time series and two flags associated with outliers were added to the output tables instead. First flag (*obs_iqr_outlier*) shows whether an observation was deemed to be an outlier by the interquartile range (IQR) test. IQR is defined as the difference between the third ($Q3$) and first ($Q1$) quartile. All values greater than $Q3 + 1.5 \times IQR$ or less than $Q1 - 1.5 \times IQR$ are considered outliers. The second flag (*obs_percentile*) was an indicator (0.0–1.0) showing which percentile a particular observation belongs to. Histograms along with box and whisker plots were used to visually show the range and distribution of the parameter observations. The plots were produced for every parameter and are included in the GRQA data repository.

3.3 Duplicate observation detection

The global datasets (GEMSTAT and GLORICH) used in this study had at least partial spatial overlap with the other three sources, which means that merging could have created duplicate sites in the GRQA. Contrary to site ID duplicates within the

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

same dataset discussed in section 3.1, site duplicates from different sources would likely also have different IDs. Therefore, rather than comparing ID information, the duplicates had to be identified by spatial proximity and time series similarity. Similar to procedures described in section 3.2, duplicate detection was done separately for each parameter.

First stage of duplicate detection was clustering sites based on their geographic location. The DBSCAN (density-based spatial clustering of applications with noise) algorithm (Xu et al., 1998) from the Scikit-learn Python library (Pedregosa et al., 2011) was used to create clusters of sites within a one kilometer radius of each other, which is the approximate accuracy of around two decimal points in latitude/longitude degrees. A major advantage of DBSCAN compared to similar density-based clustering methods is that the algorithm can be run without determining a priori the number of output clusters (Birant and Kut, 2007). In addition, DBSCAN has shown to be more applicable than others when dealing with large-scale datasets (Khan et al., 2014; Parimala et al., 2011).

Although there are time series similarity detection methods that can be applied to irregular time series and handle some degree of discontinuity, the focus of those methods is on misalignment of the time of observations rather than differences in the pattern of time series gaps (Berndt and Clifford, 1994). Therefore, it is likely that GRQA time series are too fragmented for these advanced methods to yield reliable results. A conservative approach based on root-mean-square error (RMSE) was chosen here instead. Output site clusters were converted into unique site pairs, so that all sites within a cluster could be compared to one another (e.g. a cluster of four would yield six unique ID pairs). Site ID pairs were then used to extract corresponding time series from observation data. Only observations made on matching dates were used for calculating the RMSE and only pairs where RMSE was equal to zero were considered as potential duplicates. Finally, the duplicates were exported into separate CSV files (e.g. *TP_dup_obs.csv*) along with relevant metadata to help the user decide whether the sites can be considered duplicate (Table 5). A high number of matching dates with the same observation value (column *date_match_count*) would indicate a higher likelihood of duplication.

4 Results

GRQA data model and descriptive overview. The GRQA dataset consists of observation time series for 42 different water quality parameters provided in tabular form as CSV files. Each of the observation files is accompanied by corresponding metadata files (tables and images) describing the spatial and temporal characteristics of the time series.

GRQA is made up of the following files:

– Water quality observation time series files (named *paramcode_GRQA.csv*)

– Harmonization schemas used in the preprocessing stage (*source_code_map.csv*) for each source dataset

– Summary statistics of observation values by parameter for each source dataset before (*paramcode_source_raw_stats.csv*) and after (*paramcode_source_processed_stats.csv*) processing

– Histograms (*paramcode_GRQA_hist.png*) and box plots (*paramcode_GRQA_box.png*) showing the distribution of observation values by source dataset

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Table 5. Summary table of duplicate observation file attributes.

Attribute name	Description	Data type
obs_id_1	Observation ID of first site	string
lat_wgs84_1	Latitude of first site	float
lon_wgs84_1	Longitude of first site	float
site_id_1	First site ID	string
site_name_1	First site name	string
obs_value_1	First site observation value	float
source_1	First site source	string
site_ts_availability_1	First site availability	float
site_ts_continuity_1	First site continuity	float
obs_date	Observation date	string
obs_id_2	Observation ID of second site	string
lat_wgs84_2	Latitude of second site	float
lon_wgs84_2	Longitude of second site	float
site_id_2	Second site ID	string
site_name_2	Second site name	string
obs_value_2	Second site observation value	float
source_2	Second site source	string
site_ts_availability_2	Second site availability	float
site_ts_continuity_2	Second site continuity	float
date_match_count	Number of matching dates with the same observation value	int
param_code	Parameter code	string

– Maps showing the spatial distribution of the observations by source (*paramcode_GRQA_spatial_dist.png*)

– Maps showing the median observation values of sites (*paramcode_GRQA_median.png*)

– Maps showing the monthly availability (*paramcode_GRQA_availability.png*) and continuity (*paramcode_GRQA_continuity.png*) of the observations

– Where relevant, duplicate site ID files (*source_dup_sites.csv*)

– Where relevant, duplicate observation files (*source_dup_obs.csv*)

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

Table 6. Summary table of output water quality observation file attributes.

Attribute name	Description	Data type
obs_id	Unique observation ID generated by hashing	string
lat_wgs84	Observation site latitude in WGS84	float
lon_wgs84	Observation site longitude in WGS84	float
obs_date	Observation date in the %Y-%m-%d format	string
obs_time	Observation time in the %H:%M:%S format	string
obs_time_zone	Observation time zone code	string
site_id	Observation site ID	string
site_name	Observation site name	string
site_country	Observation site country	string
upstream_basin_area	Site upstream basin area	string
upstream_basin_area_unit	Site upstream basin area unit	string
drainage_region_name	Drainage region where site is located in	string
param_code	Parameter code in GRQA	string
source_param_code	Parameter code in source dataset	string
param_name	Parameter name in GRQA	string
source_param_name	Parameter name in source dataset	string
obs_value	Observation value in GRQA	float
source_obs_value	Observation value in source dataset	float
param_form	Parameter chemical form in GRQA	string
source_param_form	Parameter chemical form in source dataset	string
unit	Parameter unit in GRQA	string
source_unit	Parameter unit in source dataset	string
filtration	Sample filtration information	string
source	Source dataset name	string
obs_percentile	Percentile of the observation value	float
obs_iqr_outlier	Flag to mark whether observation value is an outlier according to the interquartile range test	string
site_ts_availability	Monthly availability of the time series per site	float
site_ts_continuity	Monthly continuity of the time series per site	float
*_meta_*	Other observation metadata with a reference to the corresponding source column (e.g., GEMSTAT_meta_Method Description)	string
...	...	

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

Table 7. GRQA water quality parameter statistics.

Parameter code	Parameter name	Sites	Observations	Median value	Unit	Start year	End year	Outlier %
BOD	Biochemical Oxygen Demand	2,924	131,026	3.877	mg/l	1974	2019	13.5
BOD5	Biochemical Oxygen Demand (BOD5)	13,140	272,857	5.786	mg/l	1905	2020	8.5
BOD7	Biochemical Oxygen Demand (BOD7)	386	5,263	2.2	mg/l	2013	2018	5.9
COD	Chemical Oxygen Demand	2,763	109,145	26.803	mg/l	1974	2019	11.2
CODCr	Chemical Oxygen Demand (Cr)	625	6,919	25.8	mg/l	2013	2018	4.2
CODMn	Chemical Oxygen Demand (Mn)	227	2,020	3.9	mg/l	2013	2018	4.6
DC	Total Dissolved Carbon	7	9	4.8	mg/l	2000	2001	0
DIC	Dissolved Inorganic Carbon	960	29,691	11.838	mg/l	1968	2020	3.7
DIN	Dissolved Inorganic Nitrogen	119	7,808	4.2	mg/l	1998	2019	2.4
DIP	Dissolved Inorganic Phosphorus	8,873	567,530	0.046	mg/l	1942	2017	13.5
DKN	Dissolved Kjeldahl Nitrogen	2,366	71,882	0.385	mg/l	1973	2020	7.2
DO	Dissolved Oxygen	48,067	1,484,174	8.851	mg/l	1898	2020	2
DOC	Dissolved Organic Carbon	14,769	404,752	2.896	mg/l	1968	2020	6.9
DON	Dissolved Organic Nitrogen	10,810	154,098	0.384	mg/l	1951	2020	7.3
DOP	Dissolved Organic Phosphorus	142	899	0.01	mg/l	1971	2003	8.7
DOSAT	Dissolved Oxygen Saturation	34,911	949,457	92.302	%	1898	2020	8.3
NH4N	Ammonium Nitrogen	10,213	584,820	0.016	mg/l	1942	2018	15.5
NO2N	Nitrite Nitrogen	29,417	623,594	0.012	mg/l	1900	2020	12.2
NO3N	Nitrate Nitrogen	45,251	1,206,290	0.48	mg/l	1900	2020	10.9
PC	Particulate Carbon	2,898	51,049	0.908	mg/l	1995	2020	11
pH	pH	27,544	1,372,510	6.886	pH	1900	2020	14.1
PIC	Particulate Inorganic Carbon	693	6,285	0.12	mg/l	1974	2020	14.1
PN	Particulate Nitrogen	2,995	54,534	0.133	mg/l	1981	2020	9.4
POC	Particulate Organic Carbon	22,846	609,144	1.652	mg/l	1900	2020	9.8
PON	Particulate Organic Nitrogen	28	1,053	0.12	mg/l	1989	2019	13.1
POP	Particulate Organic Phosphorus	12	13	0.02	mg/l	1999	2000	7.7
TAN	Total Ammonia Nitrogen	27,980	717,419	0.065	mg/l	1900	2020	13.3
TC	Total Carbon	1,181	12,338	27	mg/l	1968	2007	3.3

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Table 7. Continued.

Parameter code	Parameter name	Sites	Observations	Median value	Unit	Start year	End year	Outlier %
TDN	Total Dissolved Nitrogen	967	62,737	0.312	mg/l	1972	2020	11.2
TDP	Total Dissolved Phosphorus	2,946	144,307	0.04	mg/l	1965	2020	11.1
TEMP	Water Temperature	26,827	1,113,048	18.932	Deg C	1912	2020	9.3
TIC	Total Inorganic Carbon	1,963	22,039	12.4	mg/l	1968	2019	3.7
TIN	Total Inorganic Nitrogen	78	11,889	4.314	mg/l	1992	2020	0.7
TIP	Total Inorganic Phosphorus	1,276	36,749	0.025	mg/l	1971	2017	12.4
TKN	Total Kjeldahl Nitrogen	9,024	407,365	0.693	mg/l	1962	2020	8.4
TN	Total Nitrogen	18,427	552,918	1.369	mg/l	1958	2020	11.7
TOC	Total Organic Carbon	18,029	417,672	4.637	mg/l	1958	2020	7.2
TON	Total Organic Nitrogen	22,796	580,450	0.634	mg/l	1900	2020	8.8
TOP	Total Organic Phosphorus	294	1,811	0.03	mg/l	1971	2020	11.9
TP	Total Phosphorus	44,741	1,890,491	0.11	mg/l	1900	2020	11.5
TPP	Total Particulate Phosphorus	76	4,853	0.032	mg/l	1978	2019	9.8
TSS	Total Suspended Solids	68,373	1,920,104	10.207	mg/l	1898	2020	18.9

The structure of GRQA observation files is given in Table 6. In addition to the attributes outlined in section 3, the extracted metadata also includes information about the upstream basin and drainage region of the observation site. It has to be noted that the availability of this information was dependent on both the source and the observation site itself and is therefore available only sporadically in GRQA as well. Parameter codes, names, forms and observation values are given as they appeared in source data alongside their harmonized and processed GRQA versions, so that end users could assess the validity of conversion and make corrections if needed.

Statistical overview of the parameters included in GRQA is shown in Table **??**. The number of sites per parameter ranges from only 15 (POP) up to 90,792 (pH). Parameters having more sites generally also have more observations. Parameters with a small number of sites and observations were usually present in only one or two source datasets. For example, dissolved organic phosphorus (DOP) only existed in WQP. Different versions of biochemical and chemical oxygen demand that could not be harmonized based on source metadata were kept separate, although the median value for BOD and BOD5 ended up being equal.

Spatial distribution of water quality observation sites depended on the parameter and is illustrated in Fig. 1 using dissolved oxygen (DO), dissolved organic carbon (DOC), TP and TSS. These parameters were the largest in terms of number of sites and observations in their corresponding groups (oxygen, carbon, nutrients and sediments). They are also used in the following figures. Some observations that could be made when examining site maps were the following:

- Europe and North America are the best represented in the case of all parameters

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

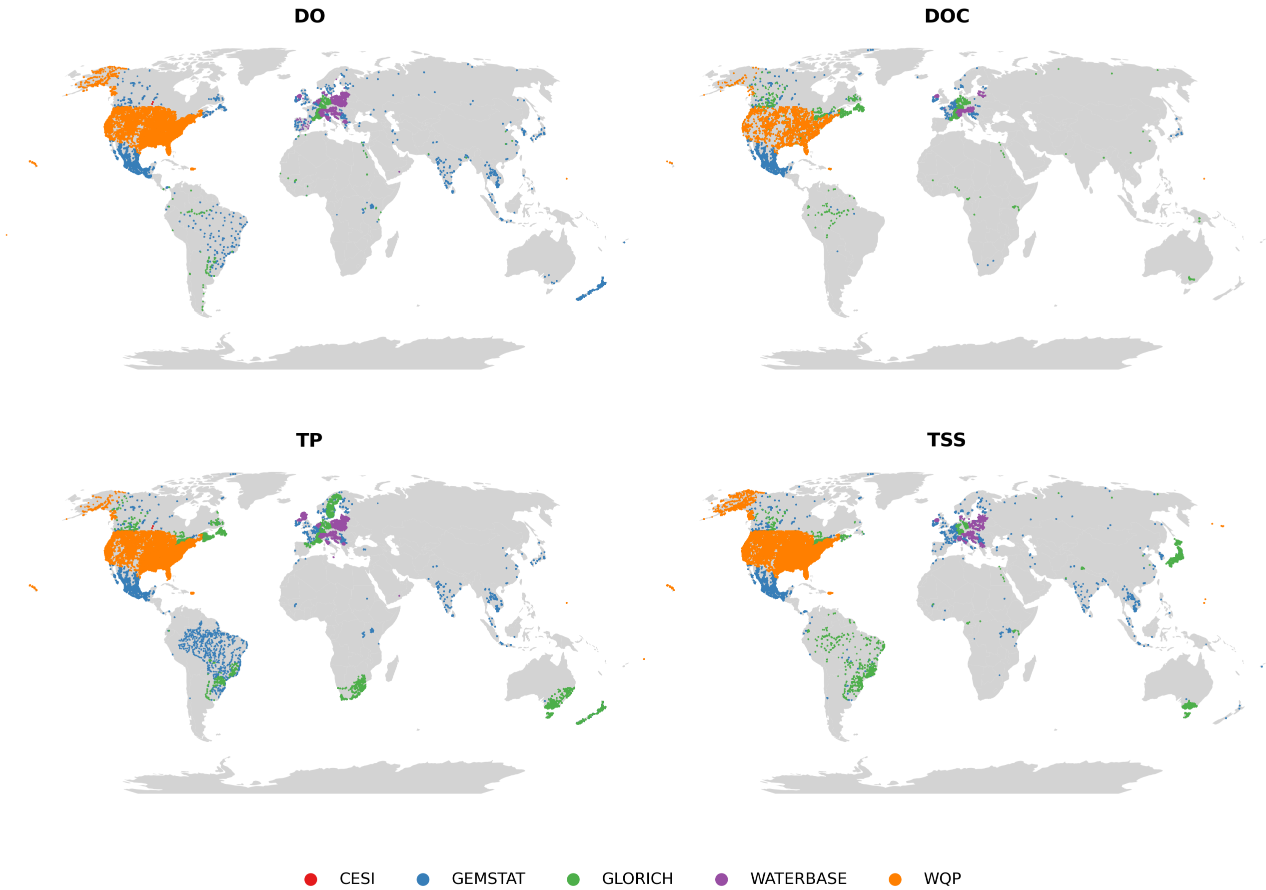

Figure 1. Distribution of observation sites for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS).

- Coverage is also good in Australia, New Zealand, parts of East Asia and Brazil in the case of some of the key parameters (e.g. TP, TN)
- Rest of the world (Africa, most of Asia) only has sporadic coverage

The temporal distribution of the four parameters is given in Fig. 2. Similar to the spatial distribution, temporal coverage of observations depended on both source data and parameter with WQP having the longest and WATERBASE the shortest time series. Most of the data from GEMSTAT are from the past decade, while GLORICH has a more even observation distribution throughout the time series.

Statistical characteristics of GRQA observation time series. As mentioned in the previous section, each of the observation files was accompanied by a set of images and tables giving insight into the characteristics of the observation time series. The structure of tabular summary statistics is shown in Table 8. These files contain some basic statistics (standard deviation, etc) about observation values per parameter and source. In addition, information about the temporal characteristics of time series

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

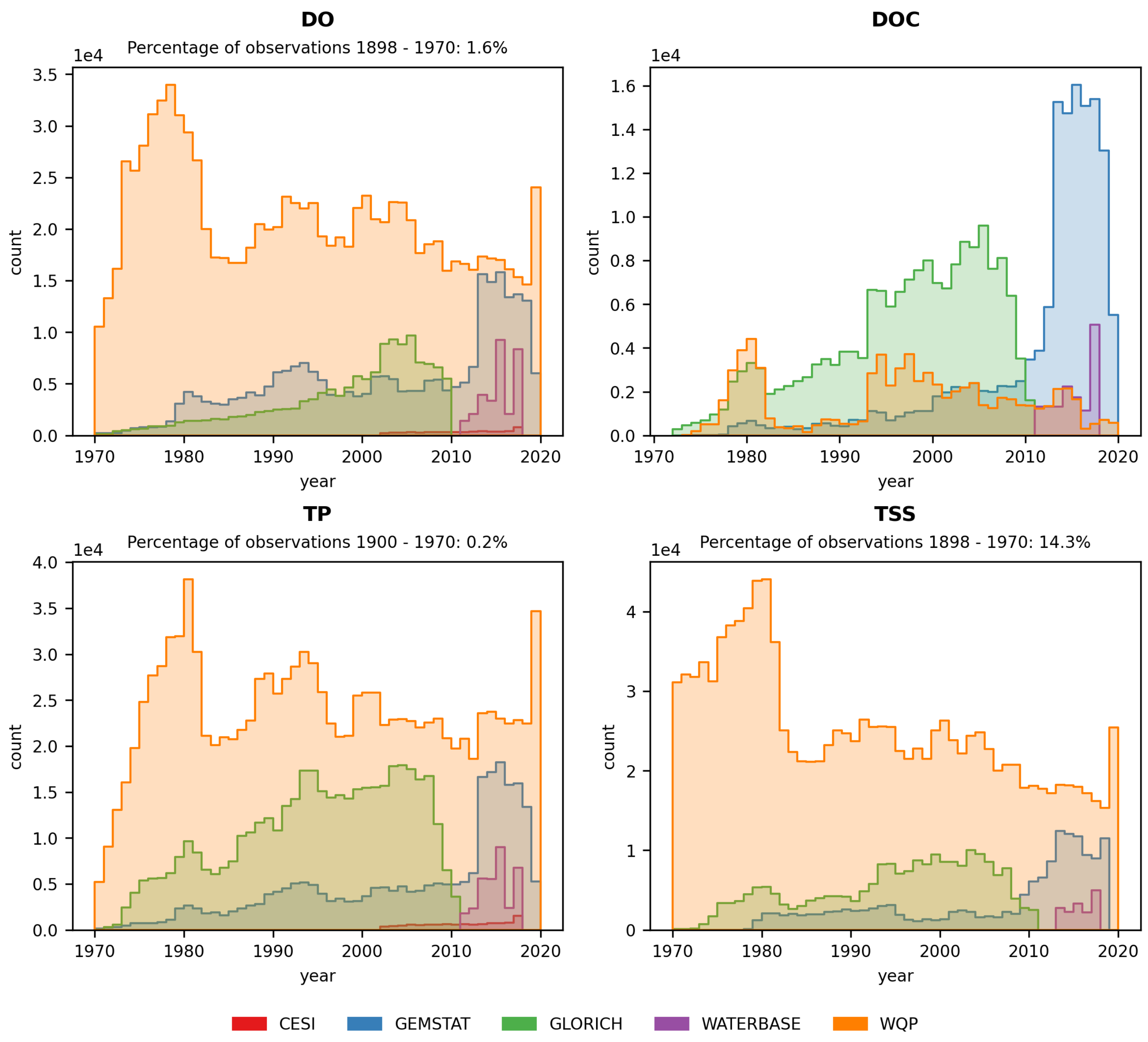

Figure 2. Temporal distribution of observations for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS) for the period 1970–2020.

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

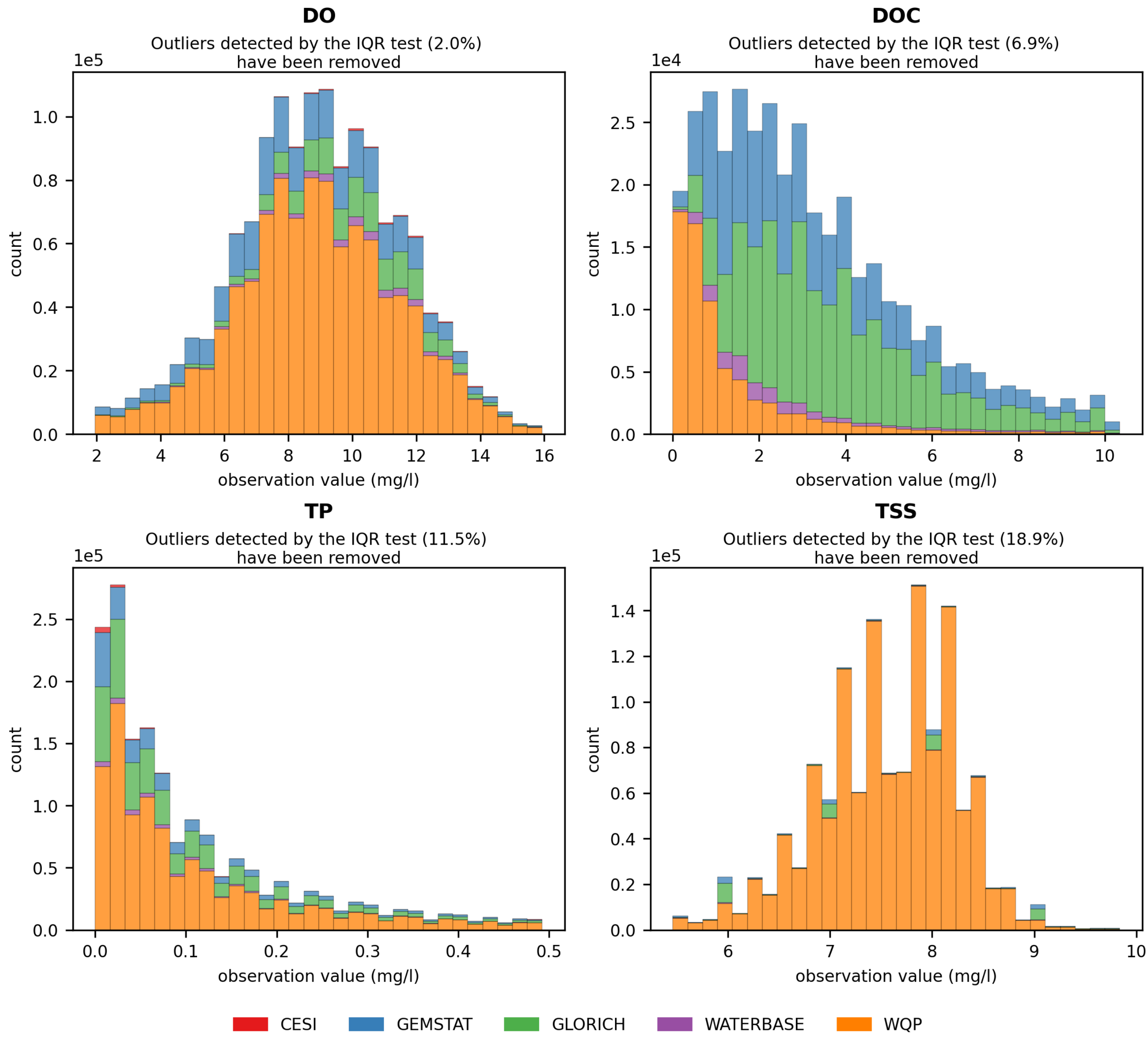

Figure 3. Distribution of observation values for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS). Outliers determined by the IQR test are not shown on the plot.

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Table 8. Summary table of observation time series statistics file attributes.

Attribute name	Description	Data type
source_param_code	Parameter code in source dataset	string
param_code	Parameter code in GRQA	string
param_name	Parameter name in source dataset	string
source_param_form	Parameter form in source dataset	string
param_form	Parameter form in GRQA	string
source_unit	Parameter unit in source dataset	string
unit	Parameter unit in GRQA	string
count	Total number of observations	int
min	Minimum observation value	float
max	Maximum observation value	float
mean	Mean observation value	float
median	Median observation value	float
std	Standard deviation of observation values	float
min_year	Time series start	int
max_year	Time series end	int
ts_length	Total time series length per parameter	float
site_count	Total number of sites per parameter	int
mean_obs_count_per_site	Mean observation count per site	float
mean_ts_length_per_site	Mean time series length in years per site	float

(mean length per site, etc) is given as well as this can be important when assessing the suitability of the data for modeling purposes.

The applicability of water quality modeling is greatly affected by the distribution of observation values as a majority of modeling methods require a near normal distribution. The skewness caused by extreme outliers is a common problem in hydrological modeling and the data often needs to be transformed and normalized in order to be usable (Helsel, 1987; Hirsch et al., 1982; Parmar and Bhardwaj, 2014). Similar behavior was also examined in GRQA, where values of most parameters showed a strong positive skew. This can be seen in histograms (Fig. 3) and box plots (Fig. 4). For illustrative purposes, values determined as outliers by the IQR test have been omitted from the figures. In the case of parameters such as TP and TSS, the skewness remains even after outlier omission. This is confirmed by the violin plots, where the total range of the values greatly exceeds the median.

Availability (Fig. 5) and continuity (Fig. 6) plots were used to examine the temporal fragmentation of the time series. In general, observations from national sources (CESI and WQP) exhibited slightly higher availability and continuity than others, likely caused by more consistent data acquisition frameworks. No clear spatial pattern emerged from the analysis meaning that differences in both indicators exist at the site level even within the same country. Due to how the metrics were calculated,

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

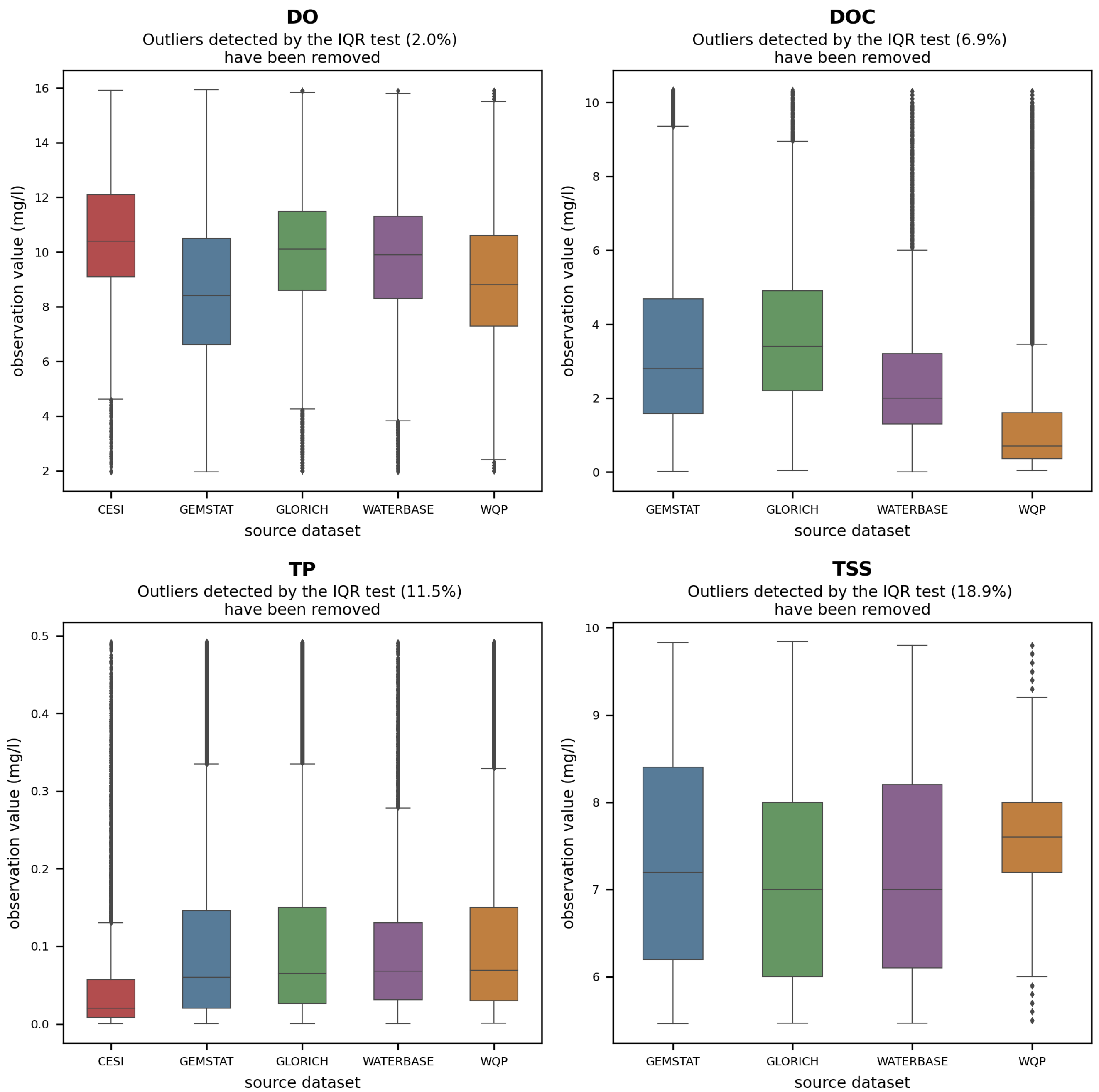

Figure 4. Box plot of observation values for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS). Outliers determined by the IQR test are not shown on the plot.

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

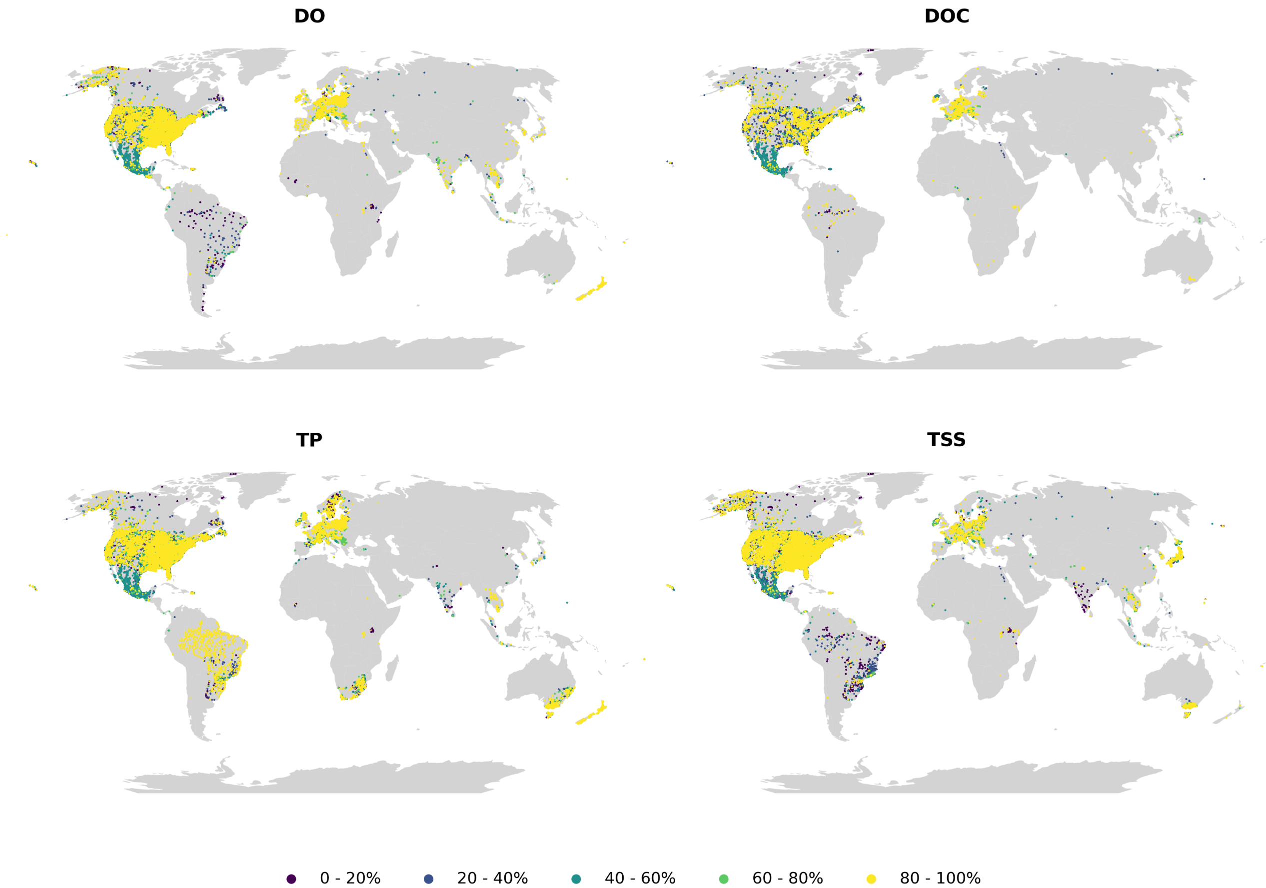

Figure 5. Monthly availability for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS).

shorter time series outperformed longer ones. An example of this is TP in Brazil, where the examined high continuity correlated with very short mean time series length (less than a year). Parameters with very fragmented time series (e.g. TSS) had only a limited number of sites where observations had been collected consistently throughout the whole time frame.

The GRQA also includes plots of median observation values, which were calculated over the whole time series for each site. Seasonal fluctuations cannot be identified on this aggregation level, so the maps are meant to be only indicative. Nevertheless, certain spatial patterns can be observed (Fig. 7). DOC concentrations are lower in higher altitudes (Alps, Rocky Mts and Appalachian Mts), which has been also observed before by Toming et al. (2020) and is possibly related organic soil horizons being thinner on steeper slopes (Rasmussen et al., 1989) and to a smaller proportion of wet soils compared to lowlands (D'Arcy and Carignan, 1997). The United States corn belt stands out with high TP concentrations, which are likely caused by agricultural pollution (**?**). TP concentrations are also high in Central Europe due to combined pressures of agricultural production and urban point sources (Grizzetti et al., 2017; Mekonnen and Hoekstra, 2018).

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

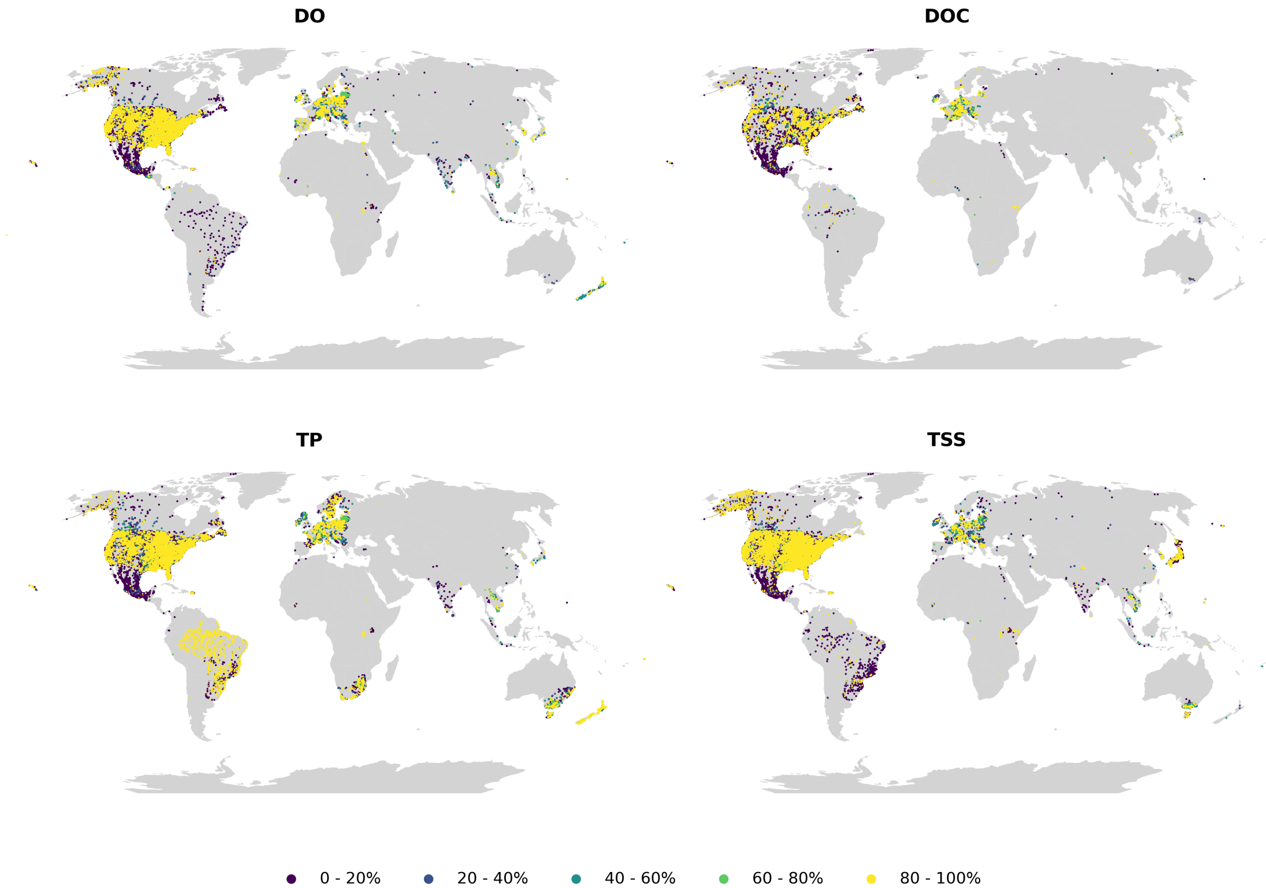

Figure 6. Monthly continuity for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS).

5 Discussion

5.1 Limitations and considerations regarding the use of GRQA

Taking into account aforementioned issues encountered during the compilation of GRQA, certain limitations and potential remaining errors have to be considered when using the dataset for water quality modeling.

Potential errors in unit conversion. As described in section 3, several assumptions had to be made when creating harmonization schemas about the chemical form of certain nitrogen parameters (NO_2, NO_3 and NH_4). Since the conversion constants were calculated based on the molar mass of the chemical form, using the wrong form would affect the resulting value. These potential conversion errors are more likely in observations originating from GLORICH, as it lacked parameter form information and also had measurements only in μg/l. For this reason, the source observation values along with source units were retained and the users can retrace the conversion steps using the harmonization schemas.

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

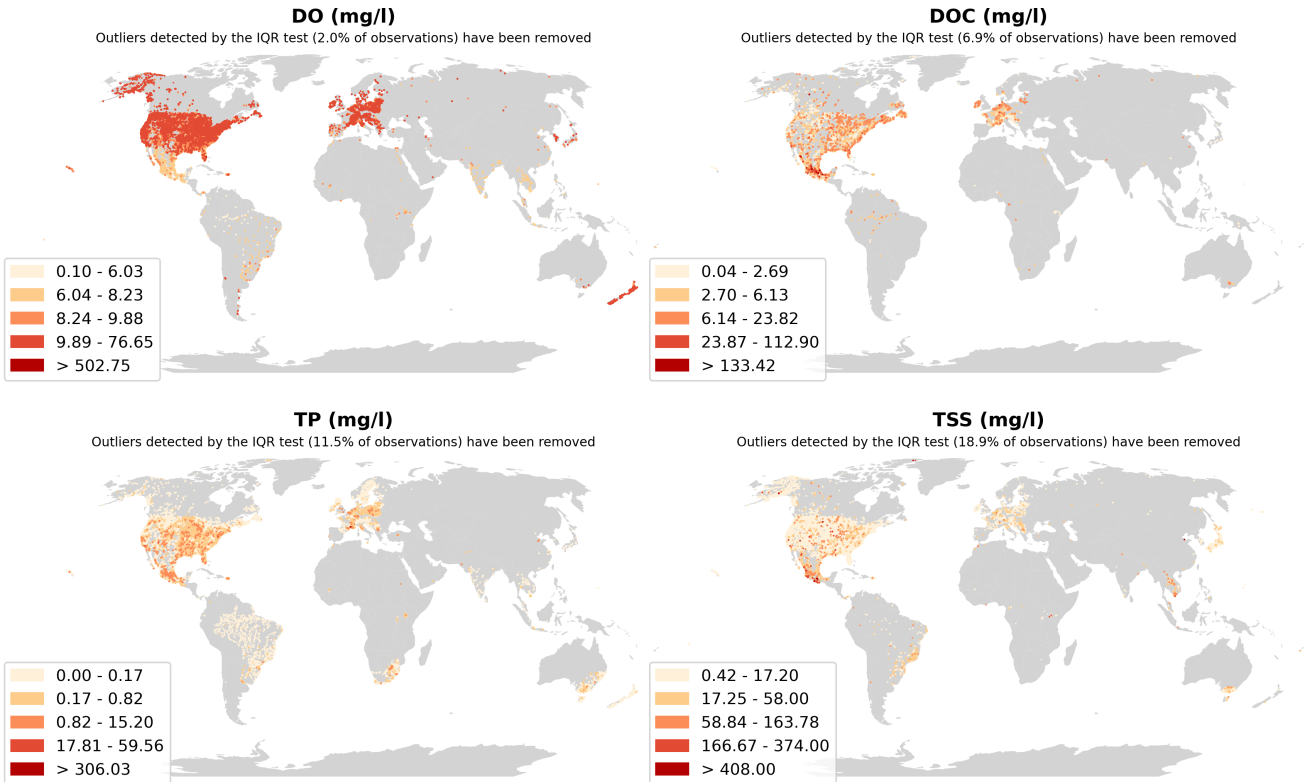

Figure 7. Spatial distribution of yearly median observation values for dissolved oxygen (DO), dissolved organic carbon (DOC), total phosphorus (TP) and total suspended solids (TSS). Outliers determined by the IQR test are not shown on the plot.

Skewness of observation values. The outlier treatment strategy used for GRQA involved only flagging the values based on the IQR test, which means that the skewness illustrated in section 4 still remains. Although the described strong positive skew existed also in source data, potential unit conversion errors could have exaggerated it. As shown by histograms, omitting flagged outliers is not enough to eliminate the skewness in some cases (TP and TSS), so additional processing could be needed to transform the data into a normal shape. Power transformation methods like the Box-Cox transformation (Box and Cox, 1964) could be used to further minimize skewness.

5.2 Suggestions for improving multi-source water quality data compilation

Metadata quality. When merging datasets from different sources, most of the complications stemmed from inadequate metadata of water quality observations, such as ambiguous parameter names and codes, and missing details on the chemical forms of parameters. This information would be integral for harmonizing units and observation values. The terms used for indicating the filtration status of samples are often dependent on the interpretation of the authors (total vs unfiltered, dissolved vs filtered), which can affect results when merging (McMillan et al., 2012; Sprague et al., 2017). Annotation of suspect or incomplete

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

data is another aspect of good quality metadata (Gudivada et al., 2017). Internal quality control measures such as the ones in GEMSTAT and WATERBASE would help the end user in the data cleaning stage and eliminate some of the outliers.

The following aspects should be considered to make multi-source data harmonization more efficient in the future:

- Parameter forms should be reported with the units
- The filtration status of the samples should be reported and the terms filtered/unfiltered should be preferred as opposed to the more ambiguous dissolved/total
- Machine-readable quality flags as found in GEMSTAT (columns *Value Flags* and *Data Quality*) or WATERBASE (columns *resultObservationStatus*, *metadata_statusCode* and *metadata_observationStatus*) should be added
- Whether observations are daily or monthly at the source level should be clearly defined
- Area units (m^2, km^2, etc) should be included, when the upstream catchment area of the site is reported
- Other information about potential errors in the data (potential duplicates, typographical errors, etc)
- When certain assumptions or decisions are made when harmonizing data from different sources, they should be reported when the data is published

Spatial and temporal discontinuity. Although spatial coverage of water quality observations in GRQA exceeds that of the existing global datasets (GEMSTAT and GLORICH), large areas of Africa and Asia are empty. A major reason might be a lack of knowledge and funding to update and extend site networks, particularly in hard to reach areas. In addition, not all governments adhere to an open data policy. For this reason, further adoption of ML methods for water quality mapping could help fill the gaps in global coverage as aforementioned problems are far less likely to improve in the near future.

The availability and continuity analysis showed that the GRQA time series are fragmented and significant gaps remain in the data, which will negatively affect large-scale modeling performance. These gaps could be caused by both issues with sensor maintenance or technical limitations under certain conditions (weather, etc) and inconsistencies in the data acquisition practices on the local level. Recently, ML based solutions for time series augmentation have been used to fill in gaps in historical monitoring data (Gao et al., 2018; Ren et al., 2019). However, this kind of gap filling still requires enough good quality training data in the existing time series fragments to be effective.

Another option for improving continuity is using data from one time series to fill in gaps in another. For example, turbidity has been successfully translated into TP and TSS content (Castrillo and García, 2020; Jones et al., 2011). As turbidity data can be acquired at a higher frequency than TP and TSS, the use of such surrogate parameters can be helpful in data scarce regions for certain parameters.

General remarks. An important part in improving the spatiotemporal coverage of water quality is raising awareness about the existing datasets (e.g. GEMSTAT), so that new institutions could join the contributor network and submit their own site data. Continued growth of international collaboration will be vital in improving open global water quality data (Blöschl et al.,

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Open Access Earth System Science Data Discussions

2019; Tang et al., 2019). Most of the data collected locally is intended only for regional or national use. Thus, the data is not compatible with those from other countries due to lack of common metadata management practices with problems discussed above being a major bottleneck (Hutton et al., 2016; Sprague et al., 2017; Stagge et al., 2019). Providing those institutions with an example workflow when designing water quality data pipelines, such as the schema recently proposed by Plana et al. (2019), would help them develop their own data management strategy. The workflow used to compile GRQA along with the issues raised in this study will hopefully also help to draw attention to this topic.

6 Conclusions

The GRQA dataset was created with the intention to improve the spatiotemporal coverage of previously available open water quality data and provide an example workflow for multi-source data compilation that can be accustomed for other data sources as well. The current version of GRQA is mainly focused on different forms of the main nutrients (N and P) and carbon compounds, although GEMSTAT, WATERBASE and WQP also had many other types of parameters that are used as water quality indicators (heavy metals, pesticides, etc). Other researchers are able to make additions and customize the dataset to their needs for parameter-specific studies using the scripts published with GRQA.

Updates and additions by the hydrological community are encouraged to further develop GRQA. The dataset is expected to have yearly updates after publishing, so that updates in source data can be taken into account. As it stands, GRQA is a set of well structured CSV files rather than a queryable database. Converting the files into a database would greatly improve data management and make extending GRQA easier in the future. We also consider the addition of an online dashboard for data visualization and download. A versioning system along with a metadata validation strategy similar to Welty et al. (2020) could be implemented to ensure metadata quality.

Future work could also include the development of a dataset for catchment characteristics in order to better study how water quality in rivers and streams is affected by land use changes in their catchments. The CAMELS dataset (Addor et al., 2017) and its regional implementations (Chagas et al., 2020; Coxon et al., 2020) can be used as an example. In addition, interactions between water quality and streamflow can be further studies by linking water quality observations to streamflow data from the Global Streamflow Indices and Metadata Archive (GSIM) (Do et al., 2018).

Code and data availability. The GRQA dataset, supplementary metadata and figures are available for download on the DataCite and OpenAire enabled repository of the University of Tartu, DataDOI, http://dx.doi.org/10.23673/re-273 (Virro et al., 2021).

The data processing scripts used for the compilation of GRQA are available on the University of Tartu Landscape Geoinformatics Lab GitHub page (https://github.com/LandscapeGeoinformatics/GRQA_src).

Author contributions. Holger Virro conceived the manuscript, conducted the data processing and scripting. All authors contributed to the development of the workflow and writing the manuscript.

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

Competing interests. The authors declare no competing interests.

Acknowledgements. This research was funded by Mobilitas+ programme grant no. MOBERC34, Marie Skłodowska-Curie Actions individual fellowship under the Horizon 2020 Programme grant agreement number 795625, grant PRG352 from the Estonian Research Council, NUTIKAS programme and the Dora Plus PhD student mobility scholarship number 36.9-6.1/1124 of the Archimedes foundation and European Regional Development Fund (EcolChange Centre of Excellence). Holger Virro is also thankful for technical support from the Yale Center for Research Computing support and the High Performance Computing Center of University of Tartu.

References

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., and Kløve, B.: A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model, Journal of Hydrology, 524, 733–752, https://doi.org/10.1016/j.jhydrol.2015.03.027, http://www.sciencedirect.com/science/article/pii/S0022169415001985, 2015.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences (HESS), 21, 5293–5313, 2017.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., Seibert, J., Hakala, K., Bock, A., Wagener, T., et al.: Accelerating advances in continental domain hydrologic modeling, Water Resources Research, 51, 10 078–10 091, 2015.

Beck, H. E., De Roo, A., and van Dijk, A. I.: Global maps of streamflow characteristics based on observations from several thousand catchments, Journal of Hydrometeorology, 16, 1478–1501, 2015.

Berndt, D. J. and Clifford, J.: Using dynamic time warping to find patterns in time series., in: KDD workshop, vol. 10, pp. 359–370, Seattle, WA, USA:, 1994.

Bierkens, M. F.: Global hydrology 2015: State, trends, and directions, Water Resources Research, 51, 4923–4947, 2015.

Birant, D. and Kut, A.: ST-DBSCAN: An algorithm for clustering spatial–temporal data, Data & knowledge engineering, 60, 208–221, 2007.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H., Sivapalan, M., et al.: Twenty-three unsolved problems in hydrology (UPH)–a community perspective, Hydrological Sciences Journal, 64, 1141–1158, 2019.

Box, G. E. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211–243, 1964.

Caraco, N. F. and Cole, J. J.: Human impact on nitrate export: an analysis using major world rivers, Ambio, 28, 167–170, 1999.

Castrillo, M. and García, Á. L.: Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods, Water Research, 172, 115 490, 2020.

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth System Science Data, 12, 2075–2096, 2020.

Chau, K.-w.: A review on integration of artificial intelligence into water quality modelling, Marine pollution bulletin, 52, 726–733, 2006.

Chen, J. and Quan, W.: Using Landsat/TM imagery to estimate nitrogen and phosphorus concentration in Taihu Lake, China, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5, 273–280, 2011.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., et al.: Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Research, 171, 115 454, 2020.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., and Kløve, B.: River suspended sediment modelling using the CART model: a comparative study of machine learning techniques, Science of the Total Environment, 615, 272–281, 2018.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis, M., Robinson, E. L., et al.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth System Science Data, 12, 2459–2483, 2020.

Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A., and Arheimer, B.: Lessons learnt from checking the quality of openly accessible river flow data worldwide, Hydrological Sciences Journal, 0, 1–13, https://doi.org/10.1080/02626667.2019.1659509, https://doi.org/10.1080/02626667.2019.1659509, 2019.

D'Arcy, P. and Carignan, R.: Influence of catchment topography on water chemistry in southeastern Quebec Shield lakes, Canadian Journal of Fisheries and Aquatic Sciences, 54, 2215–2227, 1997.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM)-Part 1: The production of a daily streamflow archive and metadata, Earth System Science Data, 10, 765–785, 2018.

Environment and Climate Change Canada: Water quality in Canadian rivers, https://open.canada.ca/data/en/dataset/55cc50dc-feb3-46d1-b40f-09254f3c00c5, accessed on November 16, 2020.

European Environment Agency: Waterbase - Water Quality ICM, https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-icm, accessed on November 16, 2020.

Evans, C., Monteith, D., and Cooper, D.: Long-term increases in surface water dissolved organic carbon: observations, possible causes and environmental impacts, Environmental pollution, 137, 55–71, 2005.

Färber, C., Lisniak, D., Saile, P., Kleber, S.-H., Ehl, M., Dietrich, S., Fader, M., and Demuth, S.: Water quality at the global scale: GEMStat database and information system, EGUGA, p. 15984, 2018.

Gao, Y., Merz, C., Lischeid, G., and Schneider, M.: A review on missing hydrological data processing, Environmental earth sciences, 77, 47, 2018.

Grizzetti, B., Pistocchi, A., Liquete, C., Udias, A., Bouraoui, F., and Van De Bund, W.: Human pressures and ecological status of European rivers, Scientific reports, 7, 1–11, 2017.

Gudivada, V., Apon, A., and Ding, J.: Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations, International Journal on Advances in Software, 10, 1–20, 2017.

Gudmundsson, L. and Seneviratne, S. I.: Towards observation-based gridded runoff estimates for Europe, Hydrology and Earth System Sciences, 19, 2859–2879, 2015.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979-present, Hydrol. Soil Sci. Hydrol, 2020.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al.: Array programming with NumPy, Nature, 585, 357–362, 2020.

Hartmann, J., Lauerwald, R., and Moosdorf, N.: A Brief Overview of the GLObal RIver Chemistry Database, GLORICH, Procedia Earth and Planetary Science, 10, 23–27, https://doi.org/10.1016/J.PROEPS.2014.08.005, https://www.sciencedirect.com/science/article/pii/S1878522014000678, 2014.

Hartmann, J., Lauerwald, R., and Moosdorf, N.: GLORICH-Global river chemistry database, PANGAEA https://doi.org/10.1594/PANGAEA, 902360, 2019.

He, B., Kanae, S., Oki, T., Hirabayashi, Y., Yamashiki, Y., and Takara, K.: Assessment of global nitrogen pollution in rivers using an integrated biogeochemical modeling framework, Water research, 45, 2573–2586, 2011.

Helsel, D. R.: Advantages of nonparametric procedures for analysis of water quality data, Hydrological Sciences Journal, 32, 179–190, 1987.

Hirsch, R. M., Slack, J. R., and Smith, R. A.: Techniques of trend analysis for monthly water quality data, Water resources research, 18, 107–121, 1982.

Hope, D., Billett, M., and Cresser, M.: A review of the export of carbon in river water: fluxes and processes, Environmental pollution, 84, 301–324, 1994.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., et al.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, 2013.

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., and Arheimer, B.: Most computational hydrology is not reproducible, so is it really science?, Water Resources Research, 52, 7548–7555, 2016.

International Centre for Water Resources and Global Change: Global Water Quality Database GEMStat, https://gemstat.org/data/data-portal/, accessed on November 16, 2020.

Jones, A. S., Stevens, D. K., Horsburgh, J. S., and Mesner, N. O.: Surrogate Measures for Providing High Frequency Estimates of Total Suspended Solids and Total Phosphorus Concentrations 1, JAWRA Journal of the American Water Resources Association, 47, 239–253, 2011.

Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F.: geopandas/geopandas: v0.8.1, https://doi.org/10.5281/zenodo.3946761, https://doi.org/10.5281/zenodo.3946761, 2020.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S.: DBSCAN: Past, present and future, in: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014), pp. 232–238, IEEE, 2014.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11 344–11 354, 2019.

Krysanova, V., Müller-Wohlfeil, D.-I., and Becker, A.: Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds, Ecological modelling, 106, 261–289, 1998.

Leon, L., Soulis, E., Kouwen, N., and Farquhar, G.: Nonpoint source pollution: a distributed water quality modeling approach, Water Research, 35, 997–1007, 2001.

Marzadri, A., Amatulli, G., Tonina, D., Bellin, A., Shen, L. Q., Allen, G. H., and Raymond, P. A.: Global riverine nitrous oxide emissions: the role of small streams and large rivers, Science of The Total Environment, p. 145148, 2021.

McKinney, W. et al.: Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56, Austin, TX, 2010.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrological Processes, 26, 4078–4111, 2012.

Meals, D. W., Dressing, S. A., and Davenport, T. E.: Lag time in water quality response to best management practices: A review, Journal of environmental quality, 39, 85–96, 2010.

Mekonnen, M. M. and Hoekstra, A. Y.: Global anthropogenic phosphorus loads to freshwater and associated grey water footprints and water pollution levels: A high-resolution global study, Water resources research, 54, 345–358, 2018.

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., and Abrahart, R.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, Hydrological Sciences Journal, 61, 1192–1208, 2016.

Neukermans, G., Ruddick, K., Loisel, H., and Roose, P.: Optimization and quality control of suspended particulate matter concentration measurement using turbidity measurements, Limnology and Oceanography: Methods, 10, 1011–1023, https://doi.org/10.4319/lom.2012.10.1011, https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.4319/lom.2012.10.1011, 2012.

Olmanson, L. G., Brezonik, P. L., and Bauer, M. E.: Airborne hyperspectral remote sensing to assess spatial distribution of water quality characteristics in large rivers: The Mississippi River and its tributaries in Minnesota, Remote Sensing of Environment, 130, 254–265, 2013.

Ouali, D., Chebana, F., and Ouarda, T. B.: Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites, Journal of Advances in Modeling Earth Systems, 9, 1292–1306, 2017.

Ouyang, W., Yang, W., Tysklind, M., Xu, Y., Lin, C., Gao, X., and Hao, Z.: Using river sediments to analyze the driving force difference for non-point source pollution dynamics between two scales of watersheds, Water research, 139, 311–320, 2018.

Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D.: Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes, Stochastic Environmental Research and Risk Assessment, 33, 481–514, 2019.

Parimala, M., Lopez, D., and Senthilkumar, N.: A survey on density based clustering algorithms for mining large spatial databases, International Journal of Advanced Science and Technology, 31, 59–66, 2011.

Parmar, K. S. and Bhardwaj, R.: Water quality management using statistical analysis and time-series prediction model, Applied Water Science, 4, 425–434, 2014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825–2830, 2011.

Pellerin, B. A., Stauffer, B. A., Young, D. A., Sullivan, D. J., Bricker, S. B., Walbridge, M. R., Clyde Jr, G. A., and Shaw, D. M.: Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection, JAWRA Journal of the American Water Resources Association, 52, 993–1008, 2016.

Plana, Q., Alferes, J., Fuks, K., Kraft, T., Maruéjouls, T., Torfs, E., and Vanrolleghem, P. A.: Towards a water quality database for raw and validated data with emphasis on structured metadata, Water Quality Research Journal, 54, 1–9, 2019.

Radwan, M., Willems, P., El-Sadek, A., and Berlamont, J.: Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model, International Journal of River Basin Management, 1, 97–103, 2003.

Rasmussen, J. B., Godbout, L., and Schallenberg, M.: The humic content of lake water and its relationship to watershed and lake morphometry, Limnology and Oceanography, 34, 1336–1343, 1989.

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., Kreft, J., Read, J. S., and Winslow, L. A.: Water quality data for national-scale aquatic research: The Water Quality Portal, Water Resources Research, 53, 1735–1745, 2017.

Ren, H., Cromwell, E., Kravitz, B., and Chen, X.: Using deep learning to fill spatio-temporal data gaps in hydrological monitoring networks, Hydrology and Earth System Sciences Discussions, pp. 1–20, 2019.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., et al.: HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, Hydrology and Earth System Sciences (Online), 22, 2018.

Shen, L. Q., Amatulli, G., Sethi, T., Raymond, P., and Domisch, S.: Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework, Scientific data, 7, 1–11, 2020.

Singh, K. P., Basant, A., Malik, A., and Jain, G.: Artificial neural network modeling of the river water quality—a case study, Ecological Modelling, 220, 888–895, 2009.

Snow, A. D., Whitaker, J., Cochran, M., den Bossche, J. V., Mayo, C., de Kloe, J., Karney, C., Ouzounoudis, G., Dearing, J., Lostis, G., Heitor, Filipe, May, R., Itkin, M., Couwenberg, B., Berardinelli, G., Badger, T. G., Eubank, N., Dunphy, M., Brett, M., Raspaud, M., da Costa,

https://doi.org/10.5194/essd-2021-51
Preprint. Discussion started: 2 March 2021

M. A., Evers, K., Ranalli, J., de Maeyer, J., Popov, E., Gohlke, C., Willoughby, C., Barker, C., and Wiedemann, B. M.: pyproj4/pyproj: 2.6.1 Release, https://doi.org/10.5281/zenodo.3783866, https://doi.org/10.5281/zenodo.3783866, 2020.

Sprague, L. A., Oelsner, G. P., and Argue, D. M.: Challenges with secondary use of multi-source water-quality data in the United States, Water research, 110, 252–261, 2017.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., and James, R.: Assessing data availability and research reproducibility in hydrology and water resources, Scientific data, 6, 190 030, 2019.

Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C., and Lindström, G.: Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale, Hydrological Sciences Journal, 57, 229–247, 2012.

Tang, T., Strokal, M., van Vliet, M. T., Seuntjens, P., Burek, P., Kroeze, C., Langan, S., and Wada, Y.: Bridging global, basin and local-scale water quality modeling towards enhancing water quality management worldwide, Current opinion in environmental sustainability, 36, 39–48, 2019.

Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., and Nõges, T.: First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery, Remote Sensing, 8, 640, 2016.

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., and Tranvik, L. J.: Predicting lake dissolved organic carbon at a global scale, Scientific reports, 10, 1–8, 2020.

United States Geological Survey: Water Quality Portal, https://www.waterqualitydata.us/portal/, accessed on November 16, 2020.

Virro, H., Amatulli, G., Kmoch, A., Shen, L., and Uuemaa, E.: GRQA: Global River Water Quality Archive, https://datadoi.ee/handle/33/331, 2021.

Wellen, C., Kamran-Disfani, A.-R., and Arhonditsis, G. B.: Evaluation of the current state of distributed watershed nutrient water quality modeling, Environmental science & technology, 49, 3278–3290, 2015.

Welty, E., Zemp, M., Navarro, F., Huss, M., Fürst, J. J., Gärtner-Roer, I., Landmann, J., Machguth, H., Naegeli, K., Andreassen, L. M., et al.: Worldwide version-controlled database of glacier thickness observations, Earth System Science Data, 12, 3039–3055, 2020.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., et al.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, 2011.

Wu, Y. and Chen, J.: Investigating the effects of point source and nonpoint source pollution on the water quality of the East River (Dongjiang) in South China, Ecological Indicators, 32, 294–304, 2013.

Xu, X., Ester, M., Kriegel, H.-P., and Sander, J.: A distribution-based clustering algorithm for mining in large spatial databases, in: Proceedings 14th International Conference on Data Engineering, pp. 324–331, IEEE, 1998.