https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Open Access Earth System Science Data Discussions

Water quality dataset in China

Jingyu Lin[1,2], Peng Wang[3], Jinzhu Wang[4], Youping Zhou[5], Xudong Zhou[6], Pan Yang[1,2], Hao Zhang[7], Yanpeng Cai[1,2*], Zhifeng Yang[1,2]

[1]Guangdong Provincial Key Laboratory of Water Quality Improvement and Ecological Restoration for Watersheds, School of Ecology, Environment and Resources, Guangdong University of Technology, Guangzhou, 510006, China
[2]Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, 511458, China
[3]Coastal and Ocean Management Institute, College of the Environment and Ecology, Xiamen University, Xiamen 361102, China
[4]School of Life and Environmental Sciences, Deakin University, Burwood, Vic, 3125, Australia
[5]Department of Ocean Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen 518055, China
[6]Global Hydrological Prediction Center, Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
[7]CAS Key Laboratory of Tropical Marine Bio-Resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China

Correspondence to: Yanpeng Cai (yanpeng.cai@gdut.edu.cn)

Abstract. Water data is a crucial asset for sustainable water resource management. However, the availability of China's water datasets lags far behind modern expectations for open geoscientific data. This dataset is a part of the China Water Data Archive (CWDA), an upcoming national collection of water-related data covering all aspects of water data for boosting data sharing in China. The CWDA aims at providing free, clean, non-sensitive, coherent, and reliable water data within China for global researchers to support the national and global water resources management and the United Nations-Water Integrated Monitoring Initiative for Sustainable Development Goals 6 and 14. In this paper, we used Python and R language to collect, tidy, reorganize, and curate the publicly available inland and coastal/ocean surface water quality data in China, following a series of data quality dimensions (integrity, completeness, consistency, and accuracy). As the most comprehensive, publicly available, handy, and clean water quality dataset in China so far, it included water quality data for daily, weekly, and monthly in the period of 1980-2022, with 17 indicators for over 330,000 observations at 2384 sites from inland to coastal/ocean areas. This dataset will greatly support works relevant to the assessment, modelling, and projection of water quality, ocean biomass, and biodiversity in China.

1 Introduction

The implication of the 2030 Agenda for Sustainable Development requires high-quality of monitoring data to measure progress and inform policymaking (Allen et al., 2021). As the core of sustainable development (UNESCO-WWAP, 2019), water is related to many targets of the SDGs, particularly the SDG 6 (i.e., to ensure availability and sustainable management of water and sanitation for all) and the SDG 14 (i.e., to conserve and sustainably use the oceans, seas, and marine resources). With the campaign of ecological civilization and a series of marine policies (i.e., Maritime Power and Strategy, Chen et al., (2019)), China aims at maintaining water resources while improving

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

resources management. To achieve the United Nation's SDGs and President Xi's version of Chinese Dream, it is important to compile water data from inland to coastal/ocean areas (Dai et al., 2022). Amongst the water quality data is a key aspect used to identify the pollutions in the Source-to-Sea (S2S) aquatic continuum for sustaining water resources and sanitation services (Regnier et al., 2022).

The recognition of the significance of the aquatic system to food, society, and security has accelerated local and national water datasets for example, dataset from United State (Read et al., 2017), Germany (Ebeling et al., 2022), Arctic watersheds (Shogren et al., 2022), United Kingdom (Bowes et al., 2018), and global water quality databases such as the Global River Water Quality Archive (GRQA; Virro et al., 2021). However, Chinese water quality data is sparsely represented in this handy worldwide dataset and lack of data from coastal/ocean areas. Few water quality data are accessible via China's national data centers (e.g., China National Environmental Monitoring Center and National Tibetan Plateau Data Centre; Lin et al., 2023). In addition, these data are stored in a user-unfriendly format that are difficult to edit and compute. For example, monthly water quality data spanning from 2006 to 2022 were presented as reports with figures derived from statistical analysis, instead of providing more reliable monitoring data. With observed data, for instance, the weekly inland surface water quality data from 2006 to 2018, were reported as a list in WORD file format only in Chinese. By now, there is no other publicly national water quality datasets covering in China.

Besides, these raw data suffered from data quality issues (i.e., inconsistency problem, Cai and Zhu (2015)), which need lots of additional efforts to make it creditable, editable, and reusable. For example, one station may have different metadata information (e.g., the longitude and latitude information for coastal/ocean water quality datasets, http://ep.nmemc.org.cn:8888/Water/). Additionally, methods and equipment used to sample and analyse are always unclear, which need further clarification (https://chinadialogue.net/en/pollution/8922-clear-as-mud-how-poor-data-is-thwarting-china-s-water-clean-up/). Finally, there are disparities in water data descriptions for the missing data (e.g., if the missing of data is caused by design, or simply due to extremely low concentration, etc.; http://www.cnemc.cn/). There is a great need to reorganize, curate, and manage water quality data to support researchers and decision-makers in China.

Despite there are some studies used national scale water quality data for water quality assessment and modelling in China (Ma et al., 2020a; Ma et al., 2020b; Huang et al., 2021; Zhang et al., 2022), the datasets underlying these studies are not publicly available due to licensing restrictions and/or government-sanctions (Lin et al., 2023). The significance of publicly available datasets is highlighted by the fact that using publicly unavailable water quality data in China to conduct water quality study may lead to the retraction of published work. For example, a study that used water data from China's Ministry of Environmental Protection to analyse the country's water quality from Science Advances was later discovered that the data was obtained without proper licensing, and the authors were forced to retract it. In 2022, a study published in the Journal of Cleaner Production used water quality data from Yangtze River was retracted as the authors did not obtain permission to publish it. It is important for researchers to obtain permission from the appropriate authorities before using water quality data to avoid potential legal and ethical issues and ensure the accuracy and reliability of their findings.

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Remote sensing techniques have been increasingly used for monitoring water quality due to their ability to provide synoptic and frequent coverage of large areas (Altenau et al., 2021). With the advancement of remote sensing technologies, researchers can monitor water quality parameters such as chlorophyll-a, turbidity, and total suspended matter from space-based sensors. For instance, Landsat data was used to estimate chlorophyll-a and total suspended matter concentrations using a semi-analytical algorithm (Yin et al., 2023). Virdis et al. (2022) used Sentinel-2 data to estimate total suspended matter and turbidity. Machine learning algorithms have also been applied to remote sensing data to improve the accuracy of water quality assessment (Cao et al., 2020). Despite these efforts in getting water quality data by using remote sensing techniques (Sagan et al., 2020), it still requires a large amount of on-site monitoring data to provide further validation.

Therefore, continuous, long-time series, standardized, well-organized, and consistent water quality datasets from inland to coastal/ocean areas are valuable assets to study the status of water quality from rivers to ocean, model different aspects of water quality indicators, and predict the impacts of emerging water pollution (i.e., coastal eutrophication and ocean harmful algal blooms due to additional nitrogen input from land and releases of radionuclides from inland unexpected nuclear power plant accidents). Such dataset is also valuable to the effective management of water resources to support the United Nation Water Action Decade (2018-2028) and Ocean Decade (2021-2030; Folke et al., 2021).

The China Water Data Archive (CWDA) is initiated to meet the huge demand for Chinese water data, to boost national water data sharing, and to advance global water-related research and applications. Maintained by the water community in China, the activities of CWDA intend to collect non-sensitive and publicly available water data, applying consistency to the formatting and curation, establishing a standardized set of metadata for different water aspects (e.g., water quantity, water quality, water demand & use, water pollutants loadings, water infrastructure & utilities, water policies, etc.), and making water data clean. Data available in this paper is a part of CWDA, which focuses on daily, weekly, and monthly surface water quality aspects from inland to coastal areas in China. The CWDA aims at supporting the establishment of China's national water data infrastructure in the future.

2 Data and methods

2.1 Data

This Chinese surface water quality dataset was compiled from a total of three data sources that are publicly available to obtain online. The files used for the creation of water quality dataset of CWDA were listed in Table 1.

Table 1. Source datasets for compiling China water quality dataset.

Name	Data Sources	Timestep	Observations	Timeframe	Number of the parameters	Number of the sites
Global daily water quality data	Global River Water Quality Archive (GRQA)	Daily	> 17,000,000	1898-2020	42	93,057

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

National weekly water quality data	China National Environmental Monitoring Centre (CNEMC)	7-day moving average	225,336	2007-2018	4	150
National monthly water quality data	National Marine Environmental Monitoring Center (NMEMC)	3-month moving average	116,296	2017-2022	6	1,991

2.1.1 GRQA

As the most comprehensive water quality dataset, GRQA has incorporated inland water quality data from five existing sources, including the Canadian Environmental Sustainability Indicators program, Global Freshwater Quality Database, GLObal RIver Chemistry database, European Environment Agency, and USGS Water Quality Portal for selected 42 water quality parameters (e.g., nutrients, carbon, oxygen, and sediments) (Read et al., 2017; Virro et al., 2021) with globally 93,057 sites in total spanning from 1898-2020 (Table 1).

2.1.2 CNEMC

As the most advanced and complete environmental data center, the China National Environmental Monitoring Centre (CNEMC) is an online information system managed by the agency of the China Ministry of Ecology and Environment (MEE). The CNEMC was established in 1979 to monitor all environmental aspects (e.g., quality of air, water, soil), to provide publicly online data, to assess environmental impacts, and to report on water environment for local and national governments. Water quality data available from this center includes yearly water quality reports spanning from 2006-2022 (**http://www.cnemc.cn/jcbg/qgdbsszyb/index_6.shtml**), 7-day moving average inland water quality data stored into individual WORD file or PDF file named by year with week number spanning from the year of 2007 to 2018 (Table 1), and real-time water quality data with a frequency of 4 hours (**https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html**) with data licensing and sharing restrictions. In this paper, we provide the digital 7-day moving water quality data which is publicly available.

2.1.3 NMEMC

Maintained by the China's MEE since 2018, the National Marine Environmental Monitoring Center (NMEMC) is an agency of a history of 60 years that specialized in marine ecological and environmental monitoring and protection. Monthly average coastal/ocean water quality data is accessible via **http://ep.nmemc.org.cn:8888/Water/** that were recorded from the year 2017 to 2022 and keep updated until now. Meanwhile 7-day moving average water quality reports of some important beaches along the coastal areas of China from 2019-2022 are available via **http://www.nmemc.org.cn/hjzl/hsycszzb/index.shtml**, as well as annual average ocean ecological environment bulletins on **http://www.nmemc.org.cn/hjzl/sthjgb/**. Observation data are only available for 3-month moving average coastal/ocean water quality data.

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

2.2 Methods

2.2.1 Data capturing and cleaning

We extracted those sites located in China based on the geopolitical map after importing all coordinate data of the GRQA dataset into ArcGIS10.8. Afterwards, metadata information of countries/regions from GRQA were tidied and renamed for consistency. For instance, regions identified as "HK", "Macao", and "Taiwan" were revised as "China". Therefore, we obtained daily water quality data in China from GRQA, which consists of 244 stations for 15 key water quality indicators. Indicators of this water quality data included Biochemical Oxygen Demand (BOD), Dissolve Oxygen (DO), Chemical Oxygen Demand (COD), Dissolved Inorganic Phosphorus (DIP), Dissolved Organic Carbon (DOC), Dissolved Oxygen Saturation (DOSAT), Ammonia Nitrogen (NH4N), Nitrite Nitrogen (NO2N), Nitrate Nitrogen (NO3N), Potential of Hydrogen (pH), Total Dissolved Phosphorus (TDP), Water Temperature (TEMP), Total Organic Carbon (TOC), Total Phosphorus (TP), and Total Suspended Solids (TSS).

Weekly average inland water quality data is tidied up from the reports collection deriving from **http://www.cnemc.cn/sssj/szzdjczb/index.shtml**. To obtain all these files automatically, we inspected the elements of the webpage to locate the key nodes where *href* attribute specifies the URL of the page the link goes for each report. Subsequently, a series of packages (i.e., *rvest*, *RSelenium*, *XML*, *purrr*) in R language were used to request remote URL and scrape the hyperlinks. A collection of hyperlinks was listed to download the original reports using *downloader* package. There are up to 500 reports with WORD file extension (i.e., in the format of DOC, DOCX, and PDF). Each filename was named by using a combination of the year and the week number. The start date and end dates for that specific week were estimated using R according to the international standard ISO 8601 that Monday is considered the first day of the week. These files were then converted into editable CSV files individually. All the CSV files were appended into a single worksheet file. Four additional columns were added to indicate the specific year, week number, estimated start date, and estimated end date. Indicators of this dataset included DO, COD, NH4N, and pH.

We have collected the 3-month moving average coastal/ocean water quality data from the NMEMC manually for each year. All data were stored as CSV files. Indicators of the coastal/ocean water quality data included COD, Dissolved Inorganic Nitrogen (DIN), DO, Dissolved Inorganic Phosphorus (DIP), pH, and Total Petroleum Hydrocarbons (TPH).

2.2.2 Metadata information processing

Metadata information of longitude and latitude is the fundamental information for identifying the location of a monitoring site. In this study, we aim at providing accurate location information for all available water quality data in China. The information of longitude and latitude was also used to export spatial point data and was overlapped with other maps to obtain other metadata information.

For daily inland water quality data, the longitude and latitude information were given by the GRQA dataset. Site location for weekly inland water quality data was coded as plain text of the administrative address, lacking

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

geographic coordinates (i.e., longitude, latitude). We first used geocoding API methods to find the address for a given place, which will convert the address into a geographic entity. Afterwards, we validate each of them by overlapping with the layers of watersheds and rivers according to the official maps obtained from the National Geomatics Center of China (http://www.ngcc.cn/ngcc/html/1/391/392/16114.html). All sites are confirmed to be located at the outlet of a river reach.

General information of metadata for 3-month moving average coastal/ocean water quality data is findable via the NMEMC. However, there are some information inconsistencies of a single site. For example, the station with code number FJD10003 was recorded with 120.57 E and 26.84 N in the year 2021 but with 120.58 E and 26.84 N in 2022. In addition, some stations with the same longitude and latitude may have different code names. Therefore, we first grouped them by code names and computed the average value of the longitude and latitude of that station to replace the initial value. Subsequently, we removed the column of the code name to avoid the same stations. Finally, we dropped the duplicated rows to get the unique stations.

All transferred longitude and latitude information was merged into a single table and was imported into ArcGIS as point shapefile. After overlapping with the city-level administrative map and watersheds delineation map that obtained from the National Geomatics Center of China, we derived other discrete information such as city, sub-watersheds, etc. The code of province and city are referred to the China Area Code and Zip code of Version 2021.

2.2.3 Technical Validation

Firstly, duplicated and irrelevant rows were removed from the inland and coastal/ocean water quality datasets. Afterwards, some observations for different indicators were messed into a single column when converting the PDF file to editable file for weekly inland water quality data. Those columns were selected to be divided and tidied up into several columns via regular expression automatically and validation manually. In addition, missing (e.g., noted as '-') and empty data were replaced with *NA*. Observations noted "*未检出*" from coastal/ocean water quality data were marked with "*No Detected*" for further clarity. Location of the station named "*河南信阳徐桥*" that collected weekly inland water quality data can't be recognized and identified based on the given information. Therefore, this station was removed from the stations' list. We provided water quality dataset including *NA* value and excluding *NA* value for different data users.

3 Data Records

3.1 General information of metadata

All data were constructed in the form of CSV, while site information is provided with point shapefile (.shp) map (available for download at **https://figshare.com/s/4f4af7fa7b8457467ea7**). Referring to the inventory information of USGS Water Quality Portal, descriptions of the location metadata for the water quality dataset were explained in Table 2. There are four types of monitoring locations, including rivers, lakes, and reservoirs in inland areas and coastal/ocean areas (Table 3).

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Table 2. Location metadata information for water quality data

Sheet file name	Field name	General introduction	Descriptions
Metadata_all	ID	/	Identifier for each indicator and each site
	WaterDataType	Water data type within a broader aspect	"W2" stands for water quality data
	MonitoringLocationIdentifier	Identifier for monitoring location	Identifiers for the stations
	MonitoringLocationDescriptionText	Given by the data source	
	MonitoringLocationName	Given by the data source	Name of the station
	MonitoringLocationType	Indicate the type of monitoring site	River, Lake, Reservoir, Ocean
	MonitoringLocationTypeCode	Using code to indicate the type	River(R), Lake(L), Reservoir(V), Ocean(C)
	MonitoringLocationTypeName	Specific the name of that monitoring site	In which rivers, which lakes
	Source_MonitoringLocationCode	Location code from the original datasets	
	LongitudeMeasure		
	LatitudeMeasure		
	ProvinceName	The acronym of a specific province	
	ProvinceCode	China area code and zip code	
	CityCode	China area code and zip code	
	IndicatorsName		
	IndicatorsUnit		
	ResolutionCode	Using numbers to identify the spatial resolution	A larger value, a higher resolution
	ResolutionName	spatial resolution	
	CountryName		
	StartDate		
	EndDate		
	SourceProvider	Data source	
	SourceProviderID	To separate the type of data source	Classified as authoritative and non-authoritative

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Table 3. Stats for different types of the monitoring sites.

MonitoringLocationType	Site_Number	Number of the indicators	IndicatorName	Number of the sites with specific indicators
Coastal/Ocean	1991	6	COD	1991
			DIN	1991
			DIP	1991
			DO	1991
			pH	1991
			TPH	1991
River	365	15	BOD	10
			COD	132
			DIP	3
			DO	135
			DOC	5
			DOSAT	24
			NH4N	123
			NO2N	13
			NO3N	119
			pH	251
			TDP	3
			TEMP	92
			TOC	1
			TP	10
			TSS	12
Lake	22	4	COD	22
			DO	22
			NH4N	22
			pH	22
Reservoir	5	4	COD	5
			DO	5
			NH4N	5
			pH	5

3.2 Data performance

Water data performance is a necessary step to determine the shortcomings, errors, and issues of research results, and ensure robust study for water data users (Koelmans et al., 2019). A large amount of missing data and outliers will

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

generate bias and uncertainty for the results despite conducting imputation (Tiyasha et al., 2020). We therefore used the *Dlookr* package, one of the most powerful tools for data inspections and data quality assessment (Mariño et al., 2022) in our work to diagnose the quality of the dataset via visualizing the distribution of missing data and outliers distribution. Since the missing data from the GRQA dataset were removed, we only analyzed the missing data from the other two (sub)datasets.

For weekly average inland water quality data, data for all indicators are in good condition with an average of 2.3% for the frequency of missing values (Figure 1a). The proportion of missing data experienced a slight increase in the years 2009 and 2016 compared to other monitoring years (Figure 1b). The data quality of 3-month moving average coastal/ocean water quality data is better than the weekly average inland water quality data (Figure 2). Indicators of COD, DIN, DIP, COD, pH, and TPH are in good condition (Figure 2a). Data quality of indicators of DO deteriorate seriously from the year 2017 to 2022 (Figure 2b). In addition, ~12.6% of observations from TPH were marked as 'No Detected', amongst mostly are in the year 2020-2023. Data users should be cautious when using the indicator of TPH and treating 'No Detected' value as missing data, which will cause bias for the results. Boxplots for each indicators illustrate that BOD, COD, DIN, DIP, DO, NH4, TOC, TP, and TPH display less variability with more outliers (Figure 3).

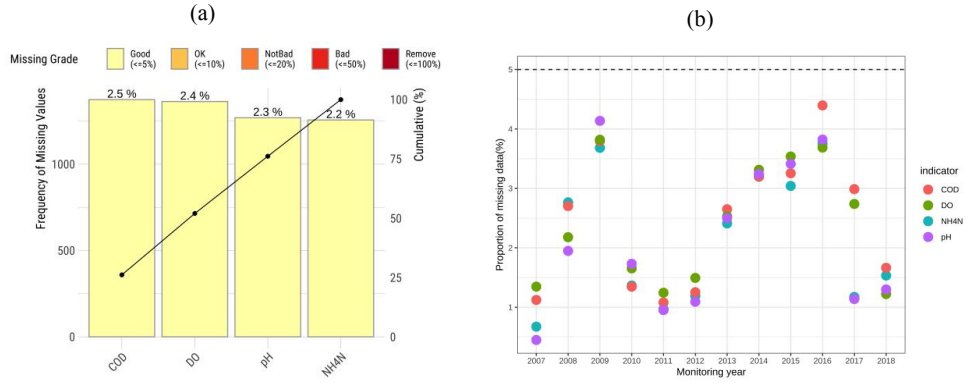
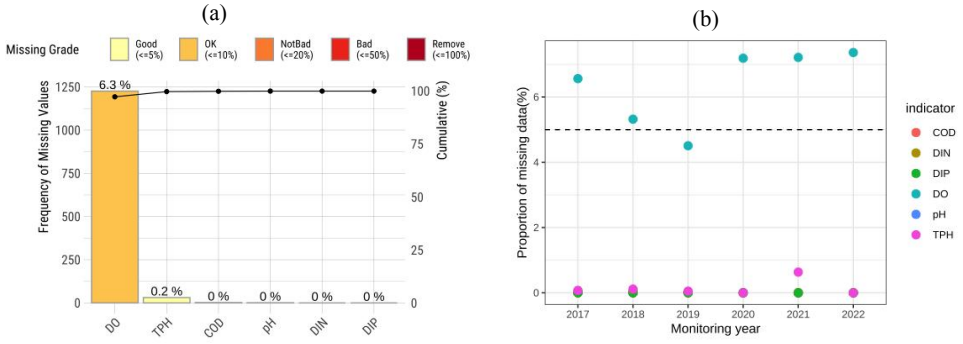
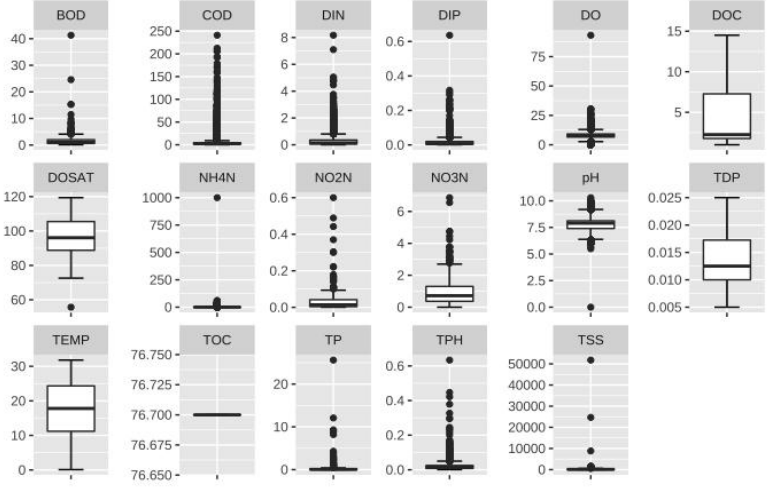


Figure 1. Data quality diagnosis for weekly average inland water quality data including visualizing the relationship between variables with missing values in panel (a) and proportion of missing values between years in panel (b).

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023



225 **Figure 2. Data quality diagnosis for ocean water quality data for analysing the relationships between variables including missing values in panel (a) and the proportion of missing values between years in panel (b). Note that value marked as 'No detected' were numbered as 0.**



230 **Figure 3. Outliers plot determined by the interquartile range (IQR) test.**

3.3 Spatial distribution of monitoring sites

This dataset provides a high intensity of water quality records along the coastal/ocean areas (Figure 4). A majority of inland water quality monitoring sites are located on the River Class 'I', 'II', and 'III' according to the Chinese river grade classification (Figure 4).

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Figure 4. Spatial distribution of monitoring sites with drainages in China.

4 Applications

Given the amount of metadata information included in our inventory and the observations, this database will be particularly useful and important for researchers and decision-makers in the fields of hydrology, environmental management, and oceanography. For example, the indicator of NH4N can be used by hydrologists to calibrate water quality models and generate projections within China. The inland and coastal/ocean surface water quality data can be connected to display the dynamic of water quality from land to ocean, thereby routing the import, transport, and export of pollutants. The high intensity of coastal/ocean water quality data can be used to indicate coastal/ocean water environment for food web (i.e., living conditions of plankton). For instance, Phytoplankton and zooplankton communities are sensitive to changes in water quality. Plankton respond to low DO levels, high nutrient levels (i.e., DIN), toxic contaminants (i.e., TPH). Therefore, such continuous coastal/ocean water quality dataset is helpful for characterizing the patterns of spatial-temporal distributions of plankton, assessing the status and trends of biodiversity, and predicting the population succession in the changing ocean world.

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

5 Data availability

All data records can be found via the temporary link **https://figshare.com/s/4f4af7fa7b8457467ea7** and **http://doi.org/10.6084/m9.figshare.22584742** (Lin et al., 2023).

6 Conclusion

This paper provides a clean, editable, and sharable national water quality dataset across inland and coastal/ocean areas in China, compiling three publicly available (sub)datasets from the public and government. It included water quality data for daily, weekly, and monthly in the period of 1980-2022, with 17 indicators for over 330,000 observations at 2384 sites. Data quality for most indicators are in conditions except data of DO from coastal/ocean areas. Since a large proportion of observations from TPH were marked as 'No Detected', data users should be cautious when using the indicator of TPH from coastal/ocean areas. As a part of the China Water Data Archive (CWDA), this paper also proposes the metadata framework for the upcoming national datasets. This database will be particularly useful and important for researchers and decision-makers in the fields of hydrology, environmental management, and oceanography for advancing the assessment, modeling, and projection of water quality, ocean biomass, and biodiversity in China.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number 52200213), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (grant number 2019ZT08L213), and the National Natural Science Foundation of China (grant numbers 52239005).

Author contributions

Yanpeng Cai, Zhifeng Yang were involved in planning and supervised the work. Jingyu Lin, Peng Wang, and Jinzhu Wang designed the code. Jingyu Lin carried out the data processing with contributions from Peng Wang and Jinzhu Wang. Jingyu Lin mapped the monitoring sites and developed the outlier detection strategy. Youping Zhou helped improve the grammar and flow of the manuscript. Jingyu Lin prepared the manuscript and Jinzhu Wang, Youping Zhou, Xudong Zhou, Pan Yang, and Hao Zhang provided critical feedback and helped shape the research, analysis, and manuscript.

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

References

Allen, C., Smith, M., Rabiee, M., & Dahmm, H. (2021). A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals. Sustainability Science, 16(5), 1701–1716.

Altenau, E. H. *et al.* (2021). The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A Global River Network for Satellite Data Products. Water Resources Research, 57(7), e2021WR030054.

Bowes, M. J., Armstrong, L. K., Harman, S. A., Wickham, H. D., Nicholls, D. J. E., Scarlett, P. M., et al. (2018). Weekly water quality monitoring data for the River Thames (UK) and its major tributaries (2009–2013): the Thames Initiative research platform. Earth System Science Data, 10(3), 1637–1653.

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., & Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. Remote Sensing of Environment, 248, 111974.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(0), 2.

Chen, J. et al. (2019). Alternative Maritime Power application as a green port strategy: Barriers in China. Journal of Cleaner Production, 213, 825–837.

Dai, M., Su, J., Zhao, Y., Hofmann, E. E., Cao, Z., Cai, W.J., et al. (2022). Carbon Fluxes in the Coastal Ocean: Synthesis, Boundary Processes, and Future Trends. Annual Review of Earth and Planetary Sciences, 50(1), 593–626.

Ebeling, P., Kumar, R., Lutz, S. R., Nguyen, T., Sarrazin, F., Weber, M., et al. (2022). QUADICA: water QUAlity, DIscharge and Catchment Attributes for large-sample studies in Germany. Earth System Science Data, 14(8), 3715–3741.

Folke, C., Polasky, S., Rockström, J., Galaz, V., Westley, F., Lamont, M., et al. (2021). Our future in the Anthropocene biosphere. Ambio, 50(4), 834–869.

Huang, J. (2021). Characterizing the river water quality in China: Recent progress and on-going challenges. Water Research, 201, 117309.

Lin, J., Bryan, B. A., Zhou, X., Lin, P., Do, H. X., Gao, L., et al. (2023). Making China's water data accessible, usable and shareable. Nature Water. https://doi.org/10.1038/s44221-023-00039-y

Lin, J., Wang, P., Wang, J., Zhou, Y., Zhou, X., Yang, P., et al. (2023): A water quality dataset in China. figshare. https://doi.org/10.6084/m9.figshare.22584742.

Ma, T., Sun, S., Fu, G., Hall, J. W., Ni, Y., He, L., et al. (2020a). Pollution exacerbates China's water scarcity and its regional inequality. Nature Communications, 11(1), 650.

Ma, T., Zhao, N., Ni, Y., Yi, J., Wilson, J. P., He, L., et al. (2020b). China's improving inland surface water quality since 2003. Science Advances, 6(1), eaau3798.

Mariño, J., Kasbohm, E., Struckmann, S., Kapsner, L. A., & Schmidt, C. O. (2022). R Packages for Data Quality Assessments and Data Monitoring: A Software Scoping Review with Recommendations for Future Developments. Applied Sciences, 12(9), 4238.

https://doi.org/10.5194/essd-2023-151
Preprint. Discussion started: 15 May 2023

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., et al. (2017). Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resources Research, 53(2), 1735–1745.

Regnier, P., Resplandy, L., Najjar, R. G., & Ciais, P. (2022). The land-to-ocean loops of the global carbon cycle. Nature, 603(7901), 401–410.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., et al. (2020). Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth-Science Reviews, 205, 103187.

Shogren, A. J., Zarnetske, J. P., Abbott, B. W., Bratsman, S., Brown, B., Carey, M. P., et al. (2022). Multi-year, spatially extensive, watershed-scale synoptic stream chemistry and water quality conditions for six permafrost-underlain Arctic watersheds. Earth System Science Data, 14(1), 95–116.

UNESCO-WWAP (2019). The United Nations World Water Development Report: Leaving no one behind. Available online at: https://unesdoc.unesco.org/ark:/48223/pf0000367306

Virro, H., Amatulli, G., Kmoch, A., Shen, L., & Uuemaa, E. (2021). GRQA: Global River Water Quality Archive. Earth System Science Data, 13(12), 5483–5507.

Virdis, S. G. P., Xue, W., Winijkul, E., Nitivattananon, V., & Punpukdee, P. (2022). Remote sensing of tropical riverine water quality using sentinel-2 MSI and field observations. Ecological Indicators, 144, 109472.

Yin, Z., Li, J., Zhang, B., Liu, Y., Yan, K., Gao, M., et al. (2023). Increase in chlorophyll-a concentration in Lake Taihu from 1984 to 2021 based on Landsat observations. Science of The Total Environment, 873, 162168.

Zhang, F., Lin, L., Li, W., Fang, D., Lv, Z., Li, M., et al. (2022). Long-Term Study of Monitoring History and Change Trends in Surface Water Quality in China. Water, 14(13), 2134.