https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Open Access Earth System Science Data Discussions

The Surface Water Chemistry (SWatCh) database: A standardized global database of water chemistry to facilitate large-sample hydrological research

Lobke Rotteveel[1], Shannon M. Sterling[1]

[1] Sterling Hydrology Research Group, Dalhousie University, Halifax, B3H 4R2, Canada

Correspondence to: Shannon M. Sterling (shannon.sterling@dal.ca)

Abstract. Openly accessible global scale surface water chemistry datasets are urgently needed to detect widespread trends and problems, to help identify their possible solutions, and identify critical spatial data gaps where more monitoring is required. Existing datasets are limited in availability, sample size/sampling frequency, and geographic scope. These limitations inhibit the answering of emerging transboundary water chemistry questions, for example, the detection and understanding of delayed recovery from freshwater acidification. Here, we begin to address these limitations by compiling the global surface water chemistry (SWatCh) database, available on Zenodo (DOI: 10.5281/zenodo.4559696) We collect, clean, standardize, and aggregate open access data provided by six national and international agencies to compile a database consisting of three relational datasets: sites, methods, and samples, and one GIS shapefile of site locations. We remove poor quality data (for example, values flagged as "suspect"), standardize variable naming conventions and units, and perform other data cleaning steps required for statistical analysis. The database contains water chemistry data across seven continents, 17 variables, 38,598 sites, and over 9 million samples collected between 1960 and 2019. We identify critical spatial data gaps in the equatorial and arid climate regions, highlighting the need for more data collection and sharing initiatives in these areas, especially considering freshwater ecosystems in these environs are predicted to be among the most heavily impacted by climate change. We identify the main challenges associated with compiling global databases – limited data availability, dissimilar sample collection and analysis methodology, and reporting ambiguity – and provide recommendations to address them. By addressing these challenges and consolidating data from various sources into one standardized, openly available, high quality, and trans-boundary database, SWatCh allows users to conduct powerful and robust statistical analyses of global surface water chemistry.

1 Introduction

Globally, 159 million people are reliant on untreated surface water, with only one in three people having access to safely-managed drinking water services (World Health Organization and United Nations Children's Fund, 2017). With two-thirds of the global population (4.0 billion people) already experiencing water shortages at least one month per year (Mekonnen and Hoekstra, 2016), a number projected to increase to 4.8-5.7 billion people by 2050 (Burek et al., 2016), maintaining the

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

quality of these resources is paramount to human health and society. One of the main obstacles to achieving this goal is a lack of openly available, high quality, transboundary data (World Health Organization and United Nations Children's Fund, 2017). Existing large-sample water quality datasets have: 1) limited availability, for example, raw data may not be published with journal articles (Alsheikh-Ali et al., 2011); 2) limited sample size, for example, datasets may only include one water body type (Hartmann et al., 2014); or 3) limited geographic scope, for example, national datasets only include data for one country.

Delayed acidification recovery is an example of a transboundary problem which would benefit from a large-sample dataset. Ecosystem acidification and associated elevated aluminium (Al) concentrations are responsible for the loss of economically-significant fish species (Committee on the Status of Endangered Wildlife in Canada, 2011; Dennis and Clair, 2012), reductions in crop success (Collignon et al., 2012), reduced forest health (Collignon et al., 2012; DeHayes et al., 1999; de Wit et al., 2010) and therefore carbon sequestration, increased cost of water treatment (Letterman and Driscoll, 1988), and may contribute to human osteological and neurological diseases (World Health Organization, 2010). Prior large-sample (Björnerås et al., 2017; Monteith et al., 2007), and global scale (Weyhenmeyer et al., 2019) studies on freshwater acidification indicate that recovery is delayed in some regions. But, so far, there is no openly available global scale database of acidification related water chemistry which includes Al, increased concentrations of which are one of the most biotically toxic effects of acidification (Gensemer and Playle, 1999).

There is a need for harmonized large-sample hydrological research (Blöschl et al., 2019). The majority of water quality research has focussed on catchment scale datasets, which limits our understanding of transboundary processes. Catchment scale analyses make valuable contributions to our understanding of hydrochemical processes, but phenomena observed at the catchment scale may not generalize across regions. For example, with freshwater acidification, catchment response to acid deposition may be altered by its geology and land use/land cover, thus observations made in one watershed may not generalize others. Specifically, regions with slow-weathering, base cation (C_B) poor, bedrock are more strongly effected by acid deposition than those with high C_B geology (Stoddard et al., 1999), and watersheds with high-intensity forest harvesting may be more strongly affected by acid deposition than those with less disturbance (Aherne et al., 2008; Feller, 2005).

Obtaining and consolidating water chemistry datasets for transboundary hydrological research is challenging due to limited data access, and disparate (that is, dissimilar) data collection programs and data reporting formats. Access may be limited because data is not published and/or kept confidential, as is the case for some sites within the United Nations International Centre for Water Resources and Global Change's Global Water Quality Database and Information System (GEMStat). Data collection programs are dissimilar largely due to a lack of international variable and analysis method definitions (World Health Organization and United Nations Children's Fund, 2017). For example, Al measurements may not be comparable across different functional, operational, and classical species definitions (Namieśnik and Rabajczyk, 2010; Ščančar and Milačič, 2006). Lastly, disparate variable naming conventions, units, and censored data notation complicates consolidation of datasets from different sources, as these notations must first be standardized.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Here, we aim to address the above limitations by contributing an openly available, standardized, easy-to-use, global water chemistry database. We focus on providing data to address the problem of delayed freshwater acidification recovery by collecting, cleaning, standardizing, and compiling datasets of acidification related water chemistry variables. Specifically, our research goals are 1) to develop a global database of acidification related surface water chemistry, 2) to identify the main limitations associated with compiling this database, 3) to identify and characterize critical spatial data gaps within existing datasets, and 4) to provide recommendations for data reporting and storage to facilitate its easy access and use by other researchers.

2 Methods

2.1 Data Sources

We obtained input data for SWatCh from openly available datasets published by national and international agencies and from datasets available on open-access servers (Table 1). Our search terms were "water chemistry data" or "water quality data" and "global" or a country name, as listed in the United Nations member countries (United Nations, 2009). We assume that water chemistry data available from these reputable sources have undergone standard laboratory quality assurance and control; spot-checks of available methodology information support this assumption. Our data search did not have a geographic focus, although our sources were limited to datasets available in English. Datasets likely missed by this approach include those hosted on servers or websites without (or without English) Search Engine Optimization (SEO); that is, those which have not been optimized with keywords identifiable by search engines to provide results (Google, 2020). For example, the search "water quality data AND Sweden" does not return a website with Swedish water quality data. This data does exist; it is hosted by the Swedish University of Agricultural Sciences at https://miljodata.slu.se/MVM/, but cannot be found using our English search terms. Please note that this data is included in SWatCh, as it is included in the European Environment Agency's Waterbase, one of our data sources. All datasets were downloaded in September 2019.

2.2 Data Inclusion

SWatCh includes 17 water chemistry variables collected in untreated surface water bodies. We define "untreated" as water that is not wastewater or receiving treatment plant effluent near to the sample collection site. The included water chemistry variables are metals: Al, and iron (Fe); C_B's: calcium (Ca), magnesium (Mg), potassium (K), and sodium (Na); acid anions: sulphate (SO_4), nitrate (NO_3), and nitrite (NO_2); other anions: fluoride (F), and chloride (Cl); nutrients: phosphorus (P), phosphate (PO_4), and ammonium (NH_4); physical parameters: pH, and temperature; total organic carbon (TOC), and dissolved organic carbon (DOC). The included water body types are streams, rivers, canals, ponds, lakes, reservoirs, and impoundments. We screened out sites identified as confidential or with other publication restrictions (**Error! Reference source not found.**).

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021


Open Access Earth System Science Data Discussions

2.2.1 Removal of Low Quality Data

We removed low quality data; for example, samples flagged as "unreliable", "suspect", or "poor quality". Additionally, we removed values below zero for all variables except temperature; these values are assumed to be entered incorrectly.

2.2.2 Removal of Duplicates

We removed duplicate site and sample data. Three of our source databases, GEMStat, the Global River Chemistry Database (GloRiCh), and Waterbase are compilations of water chemistry data from several sources, and thus repeat some measurements. We removed duplicated sites based on the site identification code and the country the site was located in. We removed duplicated samples based on the site identification code, country, date, variable name, variable fraction, and sample value. We define "variable fraction" as the component part of a water sample, such as total (unfiltered sample) and dissolved (filtered sample). Country is included as a parameter in the duplicate removal process, as some site codes are replicated across different countries; this primarily occurs for numeric or single-letter site codes.

2.3 Data Standardization

2.3.1 Database Format

We formatted the SWatCh database to reduce storage requirements and simplify use. To reduce storage requirements, we provide SWatCh as a relational database containing three datasets: 1) sites, 2) methods, and 3) samples. These three datasets are linked via site and method identification codes. We formatted each dataset after the input dataset we found the most straightforward to analyse and manipulate; that is, the sites and methods datasets are modelled after the United Nations' GEMStat, and the samples dataset is modelled after the European Environment Agency's Waterbase.

2.3.2 Variable Naming and Measurement Units

We standardized variable naming conventions to prevent confusion due to inconsistent spelling and abbreviation (Table 2). For example, aluminium (British spelling) and aluminum (American spelling) are both abbreviated to Al. We keep variable names separate from variable fractions to simplify analysis examining different fractions simultaneously. In the input datasets, the fractions are not specified for all variables; for these, we denote the fraction as "unspecified".

We simplified and standardized measurement units to prevent analysis and encoding errors (Table 2). Several input datasets did not include their encoding type, causing corrupted characters and measurement unit ambiguity. To prevent these errors, we omit non-ASCII (American Standard Code for Information Interchange) characters; for example, micrograms (µg), are denoted as ug. Measurements were reported in different units in the input datasets; we standardized them to the most common International System of Unit (SI unit) we observed for each variable. For example, Ca was reported in µg L^{-1}, mg L^{-1}, eq L^{-1}, and Mol, but was most commonly reported as mg L^{-1}, thus, we standardized the measurement unit to mg L^{-1}.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021


Open Access Earth System Science Data Discussions

2.3.3 Censored Data Notation

We standardized censored data notation to facilitate easier handling of these values. Censored data notation varied across the input datasets and included abbreviations such as "bdl", "<", or the numeric value of the detection limit. The input datasets did not distinguish between samples measured at or below the detection limit. Detection limits differed across and within datasets; thus, we standardized below detection limit values by flagging them and providing the detection limit in separate columns, allowing for various approaches of handling these results.

2.4 Mapping

We standardized the coordinate reference systems (CRSs) of the sample site locations to simplify geographic analysis. Site location coordinates are provided in various CRSs in the input datasets; thus, we first mapped the sites in their original coordinate systems, then re-projected them to the World Geodetic System 1984 (WGS 84) geographic CRS. We selected WGS 84, as this provides good mean solution across the globe and can easily be projected to local datums (Bajjali, 2018).

3 Results

The SWatCh database contains water chemistry data across 17 variables, six fractions, 38,598 sites, and 9,608,026 samples collected between 1960 and 2019 (Table 3). Sample collection frequency ranges from daily to one-time samples, depending on the data source. Not all samples included collection and analysis methodologies; for the samples where this information was available, there are over 600 different methods.

Sites in SWatCh are located across the globe, but are concentrated in North America, South America, and Europe, and encompass a variety of bedrock and land use types (Figure 2).

4 Discussion

Here, we discuss the main limitations we encounter when compiling and analysing datasets and provide recommendations for data sharing to facilitate more large-sample and global scale water chemistry research.

4.1 Data Availability and Spatial Gaps

Some variables have smaller sample sizes. The number of reported measurements differs greatly per variable, with metals (Fe and Al) and F having the smallest sample sizes and pH and temperature having the largest. This discrepancy is possibly due to the cost of measurement, where pH and temperature can be measured with a variety of field or laboratory-based multiparameter probes, whereas metals and anions require laboratory analysis. What is currently unknown, is if analysis results are under-reported for some variables. Prior research on one of the main variables with low sample size (Fe), includes an openly available research dataset of 340 water bodies in Europe and eastern North America (Björnerås et al., 2017).

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Despite the quality of this dataset, it is not included in SWatCh due to missing site identification codes, variable fraction information, and analysis methodology information. These types of published research datasets are uncommon (Alsheikh-Ali et al., 2011) and highlight the potential contribution of unpublished raw research data.

Critical data gaps exist on the African, Asian, Australian, and Antarctic continents, representing mainly the equatorial, arid, and polar climate zones (Kottek et al., 2006). The zones of missing data represent regions where climate change induced alteration of freshwater discharge regimes is projected the greatest by 2050 (Döll and Zhang, 2010). Concentrations of many water chemistry variables are discharge dependant (Moatar et al., 2017); thus, these data gaps may inhibit the detection– and therefore treatment – of emerging climate change induced water quality problems. The observed lower data availability may be because of our reliance on English datasets, less data sharing in these regions due to concerns about "parachute research" (where researchers abscond with local data to their home countries) (Serwadda et al., 2018), a lack of funding for scientific research (Serwadda et al., 2018), a lack of national data sharing regulations (Serwadda et al., 2018; Thu and Wehn, 2016), outdated information management systems (Thu and Wehn, 2016), or preferential research focus. For example, research on freshwater acidification predominantly focusses on Europe and North America (for example, Björnerås et al., 2017; Holland et al., 2005; Stoddard et al., 1999) where this is an established environmental issue, and focusses less on other regions such as China, where this is an emerging concern (for example, Li et al., 2019).

Alleviating the issue of data availability is complex (Serwadda et al., 2018), but can be facilitated through journals more consistently implementing and enforcing data sharing policies (Alsheikh-Ali et al., 2011), ensuring coherence of and balance between data sharing policies and protecting national interests (Thu and Wehn, 2016), and engaging and crediting the peoples and organizations who collected the data (Serwadda et al., 2018).

4.2 Methodology Changes and Dissimilarity

The analysis of timeseries and intercomparison of data collected at different sites is challenging due to dissimilarity of sample collection programs and methodology changes. Methodology changes throughout a timeseries may result in spurious trend test results. For example, spurious negative Al trends may result from changing from inductively-coupled plasma optical emission spectroscopy (ICP-OES) to inductively-coupled plasma mass spectrometry (ICP-MS) if the measured values are at or near the detection limit, as ICP-MS has a lower detection limit than ICP-OES. Spurious positive Al trends may result from changing from extractable Al (Al_{ext}; comprising the dissolved fraction and weakly bound or sorbed molecules) to Al_t (comprising dissolved, weakly bound or sorbed, and particulate molecules), as was done by Environment and Climate Change Canada in Atlantic Canada in 2011. Similarly, disparate analysis methods across geographic regions may hinder comparability and consolidation of data collected by different sources (World Health Organization and United Nations Children's Fund, 2017). For example, in the USA, Al samples may be analysed by United States Environmental Protection Agency (US EPA) method 200.7, with an estimated detection limit of 45 µg L^{-1} (US EPA, 2015), whereas in Europe, Al samples may be analysed by International Organization for Standardization (ISO) method 15586:2003, with an estimated detection limit of 1 µg L^{-1} (ISO/TC 147 SC2, 2003); samples analysed by these two methods cannot be compared

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021


Open Access Earth System Science Data Discussions

if Al concentrations below are 45 μg L^{-1}. To facilitate intercomparison of data, the creation of internationally standardized variable definitions and cross-boundary analysis methodology is needed (World Health Organization and United Nations Children's Fund, 2017).

4.3 Ambiguity and Inconsistency

We encounter ambiguity and inconsistency in variable and faction naming conventions, reporting units, analysis methodology, and dataset encoding. Firstly, we find variable and fraction definitions and consistency to be lacking in most input datasets. For example, an Al_d sample may be filtered through a 0.45 or 0.10 μm filter; both samples are considered Al_d but represent a different set of Al molecules. Since naming conventions are variable, and there are no internationally standardized variable definitions (World Health Organization and United Nations Children's Fund, 2017), defining variables and their fractions is required to prevent confusion regarding comparability. Similarly, reporting units and censored data notation should be defined and consistent throughout the dataset; this includes spelling, abbreviations, and capitalization. We also observe ambiguity regarding analysis methodology, where analysis methods are inadequately described or missing entirely. Ideally, analysis method reporting includes all of the following which are applicable: filter size and type, analysis instrument, acid preservative type, location of acid preservation (in field or laboratory), and the analysis/speciation method, method code, its publishing agency, and link to a reference document. Lastly, we encounter corrupted characters due to unknown dataset encoding; to prevent this ambiguity, the encoding of the dataset should be known and published, this is especially important for datasets not encoded in 8-bit Unicode (UTF-8), which preferred for data exchange (ISO/IEC JTC 1/SC 2, 2017).

4.4 Limitations and Future Work

In addition to the challenges noted above, the main limitations of SWatCh are a lack of discharge data and information on watershed land use and land cover. We did not include discharge information, as there are numerous openly available global scale river discharge datasets which cover many of the sites available in the SWatCh database. For example, the European Environmental Agency's Waterbase contains a water quality dataset (used in the SWatCh database) and a water quantity dataset. Further development is needed to integrate existing discharge datasets into SWatCh, allowing discharge-weighted water chemistry concentrations to be computed. Most of the input datasets to SWatCh do not include watershed land use and land cover information; thus, we do not include these data in the SWatCh database. Some of these data are available in the GloRiCh database (Hartmann et al., 2014). Delineating the watersheds and computing land use and land cover information is beyond the scope of this research; but is a key area of improvement for future research and updates to the database.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

5 Conclusion

Prior research demonstrates that despite variability in sample size, geographic coverage, and analysis methodology, large-sample datasets facilitate the understanding of global water chemistry processes and the identification of transboundary problems (for example, Björnerås et al., 2017; Monteith et al., 2007; Weyhenmeyer et al., 2019). Despite these clear benefits, there are few global scale water chemistry datasets. We created SWatCh to begin to fill this gap; it is a global database of surface water chemistry focussed on freshwater acidification-related variables. This database contains water chemistry data across 17 variables, six variable fractions, 38,598 sites, and 9,608,026 unique samples collected between 1960 and 2019. The numerous available variables and large sample sizes in SWatCh allows users to conduct powerful and robust statistical analyses to answer emerging global surface water chemistry questions. To facilitate data use in databases like SWatCh and by other researchers, we recommend making research data openly available, standardizing analysis methodology, and avoiding ambiguity/inconsistency in variable and fraction names, reporting units, censored data notation, analysis method descriptions, and dataset encoding. Future work should focus on filling the spatial data gaps identified in Asia, Africa, and Australia, adding discharge data, and adding catchment land use/land cover information. With more people experiencing decreased water quantity (Burek et al., 2016; Mekonnen and Hoekstra, 2016), maintaining water quality is paramount. By facilitating the global exchange of their data, researchers can contribute toward this goal.

Data and Code Availability

The SWatCh database is available on Zenodo: https://zenodo.org/record/4559696 (DOI: 10.5281/zenodo.4559696). The code used to generate the SWatCh database is published on Github: https://github.com/LobkeRotteveel/SWatCh. SWatCh is composed of third-party data, as listed in Table 1. GEMStat data, 3,034 sites (7.9 % of sites), are not available in SWatCh due to a publication ban (Supplement S1). Users may add these data by requesting the GEMStat dataset from the United Nations Environment Programme and running the SWatCh data processing scripts available from the GitHub repository listed above.

Author Contribution

LR conceived the original idea, compiled, and prepared the data, wrote the scripts, conducted the geospatial information systems (GIS) procedures, conceptualized and prepared the figures and tables, and was the principal author. SMS provided supervision and co-edited the manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Open Access Earth System Science Data Discussions

Disclaimer

While substantial efforts are made to eliminate errors from the SWatCh database, complete accuracy of the data and metadata cannot be guaranteed. All data and metadata are made available "as is". Neither Lobke Rotteveel and Dr. Shannon M. Sterling nor their current or future affiliated institutions, including the Sterling Hydrology Research Group and Dalhousie University, can be held responsible for harms, damages, or other consequences resulting from the use or interpretation of information contained within the SWatCh database.

Acknowledgements

Thank you to the United Nations Environment Programme and International Centre for Water Resources and Global Change, Environment and Climate Change Canada, the McMurdo Dry Valleys Long Term Ecological Research Team, United States of America National Science Foundation and National Water Quality Monitoring Council, the European Environment Agency, and Jens Hartmann, Ronny Lauerwald, and Nils Moosdorf for making the data collected by their contributing agencies, laboratories, researchers, and technicians openly available data for research. Thank you to Dr. Rob Jamieson for his feedback on the draft of this manuscript. Thank you to Abby Millard and Lilian Barraclough for assistance with compiling site data. This research was funded by the Nova Scotia Government through the Nova Scotia Graduate Scholarship program.

References

Aherne, J., Posch, M., Forsius, M., Vuorenmaa, J., Tamminen, P., Holmberg, M. and Johansson, M.: Modelling the hydro-geochemistry of acid-sensitive catchments in Finland under atmospheric deposition and biomass harvesting scenarios, Biogeochemistry, 88(3), 233–256, doi:10.1007/s10533-008-9206-7, 2008.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H. and Ioannidis, J. P. A.: Public Availability of Published Research Data in High-Impact Journals, PLOS One, 6(9), e24357, doi:10.1371/journal.pone.0024357, 2011.

Bajjali, W.: ArcGIS for environmental and water issues, Springer International Publishing, Cham, Switzerland., 2018.

Björnerås, C., Weyhenmeyer, G. A., Evans, C. D., Gessner, M. O., Grossart, H.-P., Kangur, K., Kokorite, I., Kortelainen, P., Laudon, H., Lehtoranta, J., Lottig, N., Monteith, D. T., Nõges, P., Nõges, T., Oulehle, F., Riise, G., Rusak, J. A., Räike, A., Sire, J., Sterling, S. M. and Kritzberg, E. S.: Widespread increases in iron concentration in European and North American freshwaters, Global Biogeochem. Cycles, 31(10), 1488–1500, doi:10.1002/2017GB005749, 2017.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., Amorim, P. B. de, Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y., Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., Barros, F. P. J. de, Rooij, G. de, Baldassarre, G. D., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Bevacqua, A. G., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., et al.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, Hydolog. Sci. J., 64(10), 1141–1158, doi:10.1080/02626667.2019.1620507, 2019.

Burek, P., Satoh, Y., Fischer, G., Kahil, M. T., Scherzer, A., Tramberend, S., Nava, L. F., Wada, Y., Eisner, S., Flörke, M., Hanasaki, N., Magnuszewski, P., Cosgrove, B. and Wiberg, D.: Water futures and solution - fast track initiative (final report), International Institute for Applied Systems Analysis, Laxenburg, Austria. [online] Available from: http://pure.iiasa.ac.at/id/eprint/13008/ (Accessed 27 May 2020), 2016.

Collignon, C., Boudot, J.-P. and Turpault, M.-P.: Time change of aluminium toxicity in the acid bulk soil and the rhizosphere in Norway spruce (*Picea abies (L.) Karst.*) and beech (*Fagus sylvatica L.*) stands, Plant Soil, 357(1–2), 259–274, doi:10.1007/s11104-012-1154-2, 2012.

Committee on the Status of Endangered Wildlife in Canada: COSEWIC assessment and status report on the Atlantic salmon, Salmo salar, Committee on the Status of Endangered Wildlife in Canada, Ottawa, Canada., 2011.

DeHayes, D. H., Schaberg, P. G., Hawley, G. J. and Strimbeck, G. R.: Acid rain impacts on calcium nutrition and forest health, BioScience, 49(10), 789–800, doi:10.2307/1313570, 1999.

Dennis, I. F. and Clair, T. A.: The distribution of dissolved aluminum in Atlantic salmon (*Salmo salar*) rivers of Atlantic Canada and its potential effect on aquatic populations, Can. J. Fish. Aquat. Sci., 69(7), 1174–1183, doi:10.1139/f2012-053, 2012.

Döll, P. and Zhang, J.: Impact of climate change on freshwater ecosystems: a global-scale analysis of ecologically relevant river flow alterations, Hydrol. Earth Syst. Sci., 14(5), 783–799, doi:10.5194/hess-14-783-2010, 2010.

Feller, M. C.: Forest harvesting and streamwater inorganic chemistry in western North America: a review, J. Am. Water Resour. Assoc., 41(4), 786–811, doi:10.1111/j.1752-1688.2005.tb03771.x, 2005.

Gensemer, R. W. and Playle, R. C.: The bioavailability and toxicity of aluminum in aquatic environments, Crit. Rev. Environ. Sci. Technol., 29(4), 315–450, doi:10.1080/10643389991259245, 1999.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Goldewijk, K. K., Beusen, A., van Drecht, G. and de Vos, M.: The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years, Global Ecol. Biogeogr., 20(1), 73–86, doi:10.1111/j.1466-8238.2010.00587.x, 2011.

Google: Search Engine Optimization (SEO) starter guide - search console help, Google Help [online] Available from: https://support.google.com/webmasters/answer/7451184?hl=en (Accessed 7 July 2020), 2020.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: a representation of rock properties at the Earth surface, Geochem. Geophy. Geosy., 13(12), 1–37, doi:10.1029/2012GC004370, 2012.

Hartmann, J., Lauerwald, R. and Moosdorf, N.: A brief overview of the GLObal RIver Chemistry Database, GLORICH, Proced. Earth Plan. Sci., 10, 23–27, doi:10.1016/j.proeps.2014.08.005, 2014.

Holland, E. A., Braswell, B. H., Sulzman, J. and Lamarque, J.-F.: Nitrogen deposition onto the United States and Western Europe: synthesis of observations and models, Ecol. Appl., 15(1), 38–57, doi:10.1890/03-5162, 2005.

ISO/IEC JTC 1/SC 2: ISO/IEC 10646:2017 Information technology — Universal Coded Character Set (UCS), International Organization for Standardization, Geneva, Switzerland. [online] Available from: https://standards.iso.org/ittf/PubliclyAvailableStandards/c069119_ISO_IEC_10646_2017.zip, 2017.

ISO/TC 147 SC2: ISO 15586:2003 Water quality - determination of trace elements using atomic absorption spectrometry with graphite furnace, Standard, International Organization for Standardization, Geneva, Switzerland. [online] Available from: https://www.iso.org/standard/38111.html, 2003.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F.: World map of the Köppen-Geiger climate classification updated, Meteorol. Z., 15(3), 259–263, doi:10.1127/0941-2948/2006/0130, 2006.

Letterman, R. D. and Driscoll, C. T.: Survey of residual aluminum in filtered water, J. Am. Water Works Assoc., 80(4), 154–158, 1988.

Li, R., Cui, L., Zhao, Y., Meng, Y., Kong, W. and Fu, H.: Estimating monthly wet sulfur (S) deposition flux over China using an ensemble model of improved machine learning and geostatistical approach, Atmos. Environ., 214, 116884, doi:10.1016/j.atmosenv.2019.116884, 2019.

Mekonnen, M. M. and Hoekstra, A. Y.: Four billion people facing severe water scarcity, Sci. Adv., 2, e1500323, doi:10.1126/sciadv.1500323, 2016.

Moatar, F., Abbott, B. W., Minaudo, C., Curie, F. and Pinay, G.: Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions, Water Resour. Res., 53(2), 1270–1287, doi:10.1002/2016WR019635, 2017.

Monteith, D. T., Stoddard, J. L., Evans, C. D., de Wit, H. A., Forsius, M., Høgåsen, T., Wilander, A., Skjelkvåle, B. L., Jeffries, D. S., Vuorenmaa, J., Keller, B., Kopácek, J. and Vesely, J.: Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry, Nature, 450(7169), 537–540, doi:10.1038/nature06316, 2007.

Namieśnik, J. and Rabajczyk, A.: The speciation of aluminum in environmental samples, Crit. Rev. Anal. Chem., 40(2), 68–88, doi:10.1080/10408340903153234, 2010.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Ščančar, J. and Milačič, R.: Aluminium speciation in environmental samples: a review, Anal. Bioanal. Chem., 386(4), 999–1012, doi:10.1007/s00216-006-0422-5, 2006.

Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F. and Wanyenze, R. K.: Open data sharing and the Global South - Who benefits?, Science, 359(6376), 642–643, doi:10.1126/science.aap8395, 2018.

Stoddard, J. L., Jeffries, D. S., Lükewille, A., Clair, T. A., Dillon, P. J., Driscoll, C. T., Forsius, M., Johannessen, M., Kahl, J. S., Kellogg, J. H., Kemp, A., Mannio, J., Monteith, D. T., Murdoch, P. S., Patrick, S., Rebsdorf, A., Skjelkvåle, B. L., Stainton, M. P., Traaen, T., van Dam, H., Webster, K. E., Wieting, J. and Wilander, A.: Regional trends in aquatic recovery from acidification in North America and Europe, Nature, 401(6753), 575–578, doi:10.1038/44114, 1999.

Thu, H. N. and Wehn, U.: Data sharing in international transboundary contexts: the Vietnamese perspective on data sharing in the Lower Mekong Basin, J. Hydrol., 536, 351–364, doi:10.1016/j.jhydrol.2016.02.035, 2016.

United Nations: Member States of the United Nations, Member States [online] Available from: http://www.un.org/en/members/index.shtml (Accessed 5 June 2020), 2009.

US EPA: EPA method 200.7 Determination of metals and trace elements in water and wastes by inductively coupled plasma-atomic emission spectrometry, Environmental Monitoring Systems Laboratory, Cincinnati, Ohio., 2015.

Weyhenmeyer, G. A., Hartmann, J., Hessen, D. O., Kopáček, J., Hejzlar, J., Jacquet, S., Hamilton, S. K., Verburg, P., Leach, T. H., Schmid, M., Flaim, G., Nõges, T., Nõges, P., Wentzky, V. C., Rogora, M., Rusak, J. A., Kosten, S., Paterson, A. M., Teubner, K., Higgins, S. N., Lawrence, G. B., Kangur, K., Kokorite, I., Cerasino, L., Funk, C., Harvey, R., Moatar, F., de Wit, H. A. and Zechmeister, T.: Widespread diminishing anthropogenic effects on calcium in freshwaters, Sci. Rep., 9(1), 10450, doi:10.1038/s41598-019-46838-w, 2019.

de Wit, H. A., Eldhuset, T. D. and Mulder, J.: Dissolved Al reduces Mg uptake in Norway spruce forest: results from a long-term field manipulation experiment in Norway, Forest Ecol. Manag., 259(10), 2072–2082, doi:10.1016/j.foreco.2010.02.018, 2010.

World Health Organization: Aluminium in drinking-water: Background document for development of WHO Guidelines for Drinking-water Quality, WHO Press, Geneva, Switzerland. [online] Available from: https://apps.who.int/iris/bitstream/handle/10665/75362/WHO_SDE_WSH_03.04_53_eng.pdf?sequence=1&isAllowed=y, 2010.

World Health Organization and United Nations Children's Fund: Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines, Geneva, Switzerland. [online] Available from: https://www.who.int/mediacentre/news/releases/2017/launch-version-report-jmp-water-sanitation-hygiene.pdf (Accessed 27 May 2020), 2017.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

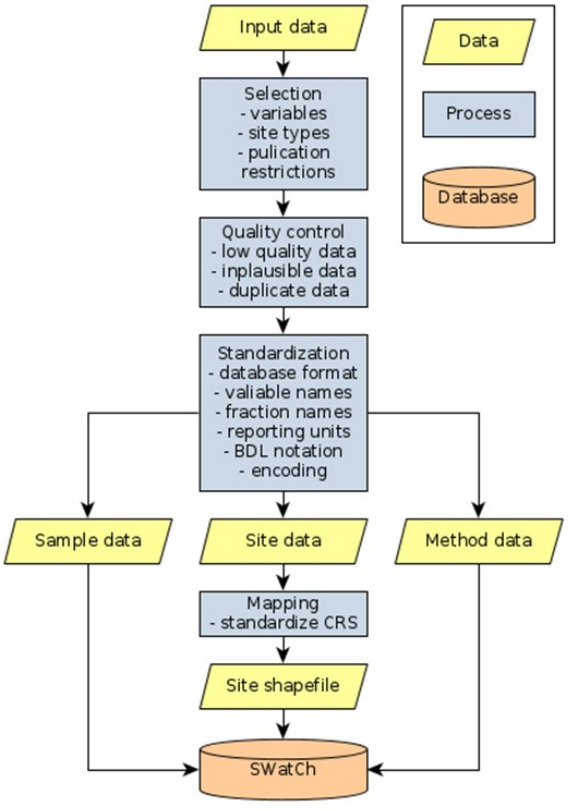
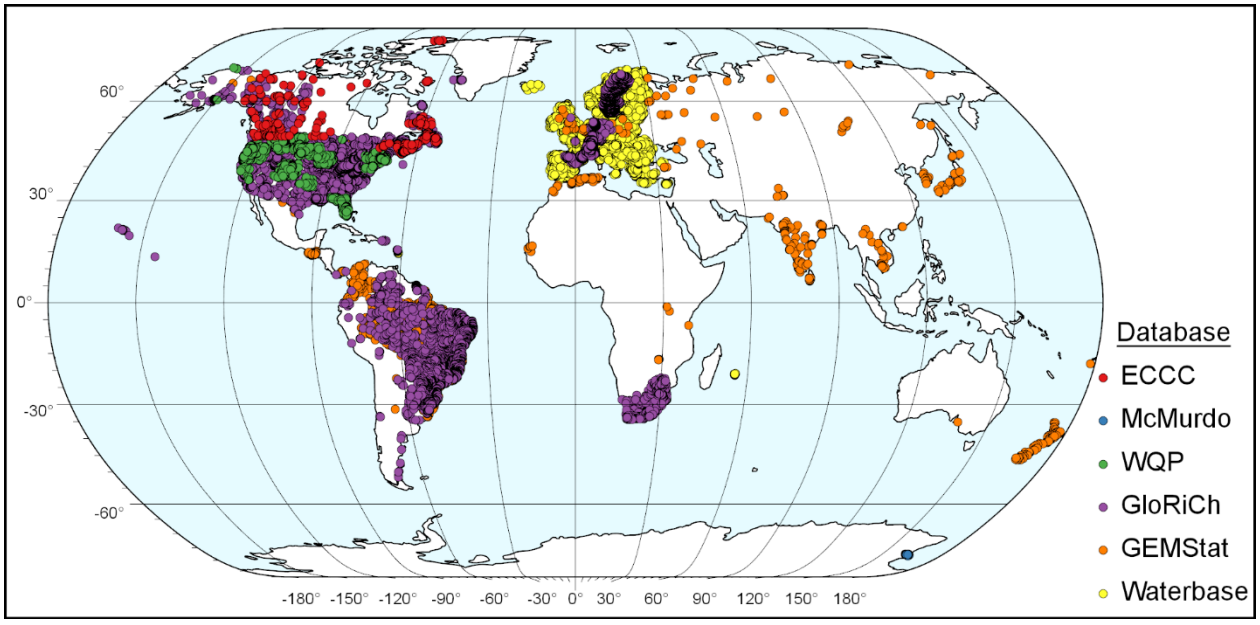


Figure 1: Workflow for creating SWatCh. Below detection limit is abbreviated as BDL and coordinate reference system is abbreviated as CRS.

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021



370 **Figure 2: The SWatCh database sample site locations, coloured by data source. Points overlap where sites are in close vicinity. Projection: Natural Earth, scales: 1:250,000,000.**

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Table 1: Data sources.

Dataset/Database	Source
Global Water Quality database and information system (GEMStat)	United Nations Environment Programme (2017). GEMStat database of the Global Environment Monitoring System for freshwater (GEMS/Water) Programme. International Centre for Water Resources and Global Change, Koblenz. Accessed 10 August 2019. Available upon request from GEMS/Water Data Centre: gemstat.org
Global River Chemistry Database (GloRiCh)	Hartmann, J., Lauerwald, R., Moosdorf, N. (2019). GLORICH - Global river chemistry database. PANGAEA. Accessed 18 August 2019. Available from: https://doi.org/10.1594/PANGAEA.902360. Supplement to: Hartmann, J. et al. (2014). 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.
National Long-Term Water Quality Monitoring Database	Environment and Climate Change Canada (2019). National Long-term Water Quality Monitoring Data. Accessed 8 September 2019. Available from: http://data.ec.gc.ca/data/substances/monitor/national-long-term-water-quality-monitoring-data/
Water Quality Database	National Water Quality Monitoring Council (2019). Water Quality Portal. Accessed 7 September 2019. Available from: https://www.waterqualitydata.us/apps_using_portal/.
Waterbase	European Environment Agency - European Environment Information and Observation Network (Eionet) (2019). Waterbase - Water Quality. Accessed 8 September 2019. https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-2.
McMurdo Dry Valleys Long Term Ecological Research Network	Lyons, W., Welch, K. (2014). McMurdo Dry Valleys LTER: Limnological Chemistry, Ion Concentrations and Silicon. Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/2cbb9e62342bdf6118b20553be7b922f.
	Priscu, J. (2014). McMurdo Dry Valleys LTER: Macronutrient Concentrations (NH4 - NO3 - NO2 - PO4) in Lakes. Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/ddafa14d91edc1092d4463d66d157fb9.
	Lyons, W., Welch, K. (2015). McMurdo Dry Valleys LTER: Stream Chemistry and Ion Concentrations. Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/be9f781814330116f68844a8957962e4.
	Lyons, W. (2015). McMurdo Dry Valleys LTER: Stream Chemistry - Dissolved Organic Carbon. Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/578d31fa0e142b6b7a6c80095d29b968.
	Lyons, W. (2015). McMurdo Dry Valleys LTER: Stream Nutrients (nitrate, nitrite, ammonium, reactive phosphorus). Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/81cfcad244f7d3f6a6fcdab83ea75849.
	Priscu, J. (2018). McMurdo Dry Valleys LTER: Dissolved Organic Carbon Concentrations in Lakes. Environmental Data Initiative. Accessed 8 September 2019. doi: 10.6073/pasta/95adee9e18f9487a373b51adced40875.

375

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Open Access Earth System Science Data Discussions

Table 2 Variable naming and measurement unit conventions in the SWatCh database.

SWatCh variable notation	Variable name	SWatCh unit notation	Unit
Al	aluminium	ug/l	µg L^{-1}
Fe	iron	ug/l	µg L^{-1}
Ca	calcium	mg/l	mg L^{-1}
Mg	magnesium	mg/l	mg L^{-1}
K	potassium	mg/l	mg L^{-1}
Na	sodium	mg/l	mg L^{-1}
Cl	chloride	mg/l	mg L^{-1}
F	fluoride	mg/l	mg L^{-1}
SO4	sulphate	mg/l	mg L^{-1}
NO3	nitrate	mg/l	mg L^{-1}
NO2	nitrite	mg/l	mg L^{-1}
NH4	ammonium	mg/l	mg L^{-1}
P	phosphorus	mg/l	mg L^{-1}
PO4	phosphate	mg/l	mg L^{-1}
OC	organic carbon	mg/l	mg L^{-1}
pH	pH	unit	unit
temperature	temperature	deg_c	ºC

380

https://doi.org/10.5194/essd-2021-43
Preprint. Discussion started: 17 March 2021

Open Access Earth System Science Data Discussions

Table 3 The SWatCh database sample sizes by variable and fraction. Field measurements are only applicable to pH and temperature. Organic carbon is abbreviated as OC and temperature is abbreviated as temp.

	Dissolved		Extractable		Recoverable		Total		Unspecified		Field	
	sites	samples	sites	samples	sites	samples	sites	samples	sites	samples	sites	samples
Al	2,515	27,113	170	5,959	642	2,604	4,144	120,216				
Fe	103	9,866	164	5,941	89	766	277	29,905				
Ca	12,199	591,313	155	5,694	797	5,094	1,433	19,340				
Mg	12,734	590,961	155	5,678	802	5,079	7,708	96,790				
Na	11,639	560,572	161	5,686	589	2,910	6,599	86,296				
K	11,980	549,954	161	5,689	490	2,657	6,205	99,386	189	257		
Cl	12,414	668,247					9,357	144,919	368	6,340		
F	5,567	424,997					901	6,771	25	3,335		
SO_4	13,347	637,383					8,840	117,968	108	3,805		
NO_3	8,434	228,049					4,022	74,668	2,566	65,234		
NO_2	9,148	205,258					2,606	36,721	2,139	51,176		
NH_4	11,241	506,116					5,677	146,000				
P	2,967	103,980					10,306	361,460	35	270		
PO_4	10,351	494,427					968	21,929	13	96		
OC	11,128	301,467					6,642	184,455				
pH									27,183	1,144,310	135	5,853
temp									27,296	849,130	195	6,272