MassWateR: Improving quality control, analysis, and sharing of water quality data

by Marcus W. Beck, Benjamen Wetherill, Jillian Carr, and Pamela DiBona

Abstract An abstract of less than 150 words.

Introduction

Water quality measurements provide the foundation of environmental monitoring programs designed to protect or restore aquatic resources. In the United States, water quality monitoring programs are broadly guided by the federal Clean Water Act with the singular goal of restoring and maintaining the chemical, physical, and biological integrity of the nation's surface waters. Similarly, the Water Framework Directive provides a framework for the protection of aquatic resources in member states of the European Union. Numeric standards that define critical thresholds for protecting recreational, aquatic life, industrial, navigation and consumptive uses of the resource are often established, that, if exceeded based on water quality measurements, require additional regulatory action to ensure compliance. These standards and other regulatory assessments as applied at the state-level use information from long-term monitoring datasets (Schiff et al. 2016; Tango and Batiuk 2013), or data collected *ad hoc* from multiple assessment endpoints (Stein and Cadien 2009; Behmel et al. 2016; Kumpel et al. 2020), where the former is atypical for most surface water bodies. Many state or regional institutions that assess water quality rely on decentralized data sources, often combining datasets from local watershed groups or participatory science programs rather than a single database that contains adequate coverage for areas of interest [Buytaert et al. (2014); Kelly22]. Use of these monitoring data in a regulatory context is not possible unless standard operating procedures are adopted and the data fulfill quality control requirements.

Monitoring data of sufficient quantity and quality are critical to ensure precise and accurate representation of environmental conditions. A significant bottleneck in the use of monitoring data for environmental assessment of surface waters is the ability to clearly and efficiently indicate that the data fulfill applicable quality control (QC) criteria for inclusion in a consolidated database. Common QC checks for *in situ* field measurements or concentrations measured in the laboratory may include 1) comparison of the precision between replicate samples (duplicates), 2) comparison of a sample to a known concentration (spikes or instrument checks), and 3) precision of the measurement from an empty or blank sample (blanks). An adequate number of QC samples must also be included in the dataset as a measure of "completeness". These checks are often compiled in a single report for review by appropriate regulatory agencies. For example, precision of duplicate samples for a given parameter must not vary 5% and at least 10% of the data should be dedicated to these checks as a measure of completeness. For local monitoring groups that lack the resources to develop robust and repeatable workflows, QC reports are often prepared manually before submitting the data for review This process is time-consuming and prone to errors, often limiting the amount of useful information that is submitted to formal databases.

The R statistical programming language offers a valuable software platform for developing tools to improve the QC of water quality data. The use of R with document generation systems offered through packages like **knitr** (Xie 2015) and **rmarkdown** (Allaire et al. 2023) can be leveraged to generate QC reports that follow a standard format for review by regulatory agencies. These tools can also be used to format water quality data for submission to state or national water quality databases, such as the Water Quality Exchange (WQX) database maintained by the US Environmental Protection Agency (USEPA). This database is the largest source of monitoring data in the United States that includes information on hydrologic conditions and chemical, physical, and biological measurements from surface waters. Moreover, many environmental resource managers have the need to analysis status and trends in monitoring data and R packages such as **ggplot2** (Wickham 2016) offer useful approaches to visualize numerous water quality records in a single graph. Integrating this functionality into a single package is expected to have wide ranging utility for anyone collecting surface water data and is likely to improve the quality and insights obtained from these data.

This paper describes the MassWateR package developed to improve how environmental professionals perform quality control, analysis, and sharing of monitoring data for surface waters. The regional focus of the package is for monitoring data collected in Massachusetts, USA, with QC reports submitted to the Massachusetts Department of Environmental Protection and data submitted to the national WQX database. Although the initial conception of MassWateR was meant to address regional needs in Massachusetts, there is nothing specific in the package that prevents its use outside of the state

as the QC checks and analyses follow routine methods for data collected at locations elsewhere. As such, this paper is written with emphasis on how the tools are broadly applicable to anyone interested in improving efficiency and reproducibility of QC checks, in addition to analysis of water quality data and submission to WQX as the largest source of water monitoring data in the US.

Requirements for using MassWateR

To our knowledge, there are no existing R packages on CRAN that can be used to facilitate QC of water quality data, nor are any available that facilitate submission to existing databases. However, there are several that can be used to retrieve and analysis data from existing sources (see the CRAN *Hydrology* Task View). In particular, the **dataRetrieval** package (De Cicco et al. 2022) has been used widely to retrieve data from the USEPA Water Quality Portal (WQP), which is the counterpart of the WQX system for accessing data submitted using the latter. This package leverages a robust API to query existing water quality data in the WQP. As such, data retrieval using existing web services is much simpler than data submission to a similar resource, as data formatting requirements do not apply when retrieving data. Developing a robust tool that can facilitate the upload of data to WQX, in addition to streamlining QC processes, would further the value of packages like **dataRetrieval** by increasing the amount of data that can be accessed through the WQP. The **MassWateR** package was developed to provide this benefit.

Users can engage with **MassWateR** to achieve different goals. This design was purposeful based on likely differences in needs among the user community. Although increasing data submission and facilitating QC reporting was the primary goal, we also assumed that users may not want to do both. That is, QC reporting is not a requirement to submit to WQX, whereas state institutions require it for regulatory assessment. Users may also simply have a need to understand trends or to summarize their data, while also wanted to extend these analyses beyond **MassWateR** using additional R packages. Figure xx demonstrates how a user may apply the functions in **MassWateR** once the required data are imported. The functions allow a user to engage with their data several, including 1) screening data for quality control, 2) summarizing quality control results, 3) creating graphics for analysis and reports to stakeholders, and 4) formatting data for upload to WQX.

No matter the user need, all data inputs to MassWateR must follow a strict format.
Make a case for a unified format, cannot cater to all, input checks are meant assist in this effort

Read

Quality Control

Analysis

Data submission

Building a community and future work

All developers hope that their package is used by the intended audience. To ensure this for MassWateR, a community of practice was…

Building the community of practice…

MassWater in other locations, expansion elsewhere

Inclusion of historical data

Inclusion for continuous monitoring data, consider outlier anlaysis, drift, biofouling, etc.

Summary

Acknowledgments

References

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2023. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Behmel, Sonja, Mathieu Damour, Ralf Ludwig, and MJ Rodriguez. 2016. "Water Quality Monitoring Strategies—a Review and Future Perspectives." *Science of the Total Environment* 571: 1312–29. https://doi.org/10.1016/j.scitotenv.2016.06.235.

Buytaert, Wouter, Zed Zulkafli, Sam Grainger, Luis Acosta, Tilashwork C Alemie, Johan Bastiaensen, Bert De Bièvre, et al. 2014. "Citizen Science in Hydrology and Water Resources: Opportunities for Knowledge Generation, Ecosystem Service Management, and Sustainable Development." *Frontiers in Earth Science* 2: 26. https://doi.org/10.3389/feart.2014.00026.

De Cicco, Laura A., David Lorenz, Robert M. Hirsch, William Watkins, and Mike Johnson. 2022. *dataRetrieval: R Packages for Discovering and Retrieving Water Data Available from u.s. Federal Hydrologic Web Services* (version 2.7.12). Reston, VA: U.S. Geological Survey; U.S. Geological Survey. https://doi.org/10.5066/P9X4L3GE.

Kumpel, Emily, Clara MacLeod, Kara Stuart, Alicea Cock-Esteb, Ranjiv Khush, and Rachel Peletz. 2020. "From Data to Decisions: Understanding Information Flows Within Regulatory Water Quality Monitoring Programs." *Npj Clean Water* 3 (1): 38. https://doi.org/10.1038/s41545-020-00084-0.

Schiff, Ken, PR Trowbridge, ET Sherwood, Peter Tango, and Rich A Batiuk. 2016. "Regional Monitoring Programs in the United States: Synthesis of Four Case Studies from Pacific, Atlantic, and Gulf Coasts." *Regional Studies in Marine Science* 4: A1–7. https://doi.org/10.1016/j.rsma.2015.11.007.

Stein, Eric D, and Donald B Cadien. 2009. "Ecosystem Response to Regulatory and Management Actions: The Southern California Experience in Long-Term Monitoring." *Marine Pollution Bulletin* 59 (4-7): 91–100. https://doi.org/10.1016/j.marpolbul.2009.02.025.

Tango, Peter J, and Richard A Batiuk. 2013. "Deriving Chesapeake Bay Water Quality Standards." *JAWRA Journal of the American Water Resources Association* 49 (5): 1007–24. https://doi.org/10.1111/jawr.12108.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

Marcus W. Beck
Tampa Bay Estuary Program
263 13th Ave S
St. Petersburg, Florida, USA 33701
https://tbep.org
ORCiD: 0000-0002-4996-0059
mbeck@tbep.org

Benjamen Wetherill
ACASAK Consulting
Boston, Massachusetts, USA
https://www.acasak.com/
ORCiD: 0000-0002-0912-0225
bwetherill@acasak.co

Jillian Carr
Massachusetts Bays National Estuary Partnership
University of Massachusetts Boston, 100 Morrissey Blvd
Boston, Massachusetts, USA 02125
https://www.mass.gov/orgs/massachusetts-bays-national-estuary-partnership
Jillian.Carr@umb.edu

Pamela DiBona

Massachusetts Bays National Estuary Partnership
University of Massachusetts Boston, 100 Morrissey Blvd
Boston, Massachusetts, USA 02125
https://www.mass.gov/orgs/massachusetts-bays-national-estuary-partnership
pamela.dibona@state.ma.us