USGS
science for a changing world

Integrated Water Availability Assessments Program

Pooling Resources Across Organizations—Multisource Water-Quality Data for the Delaware River Basin

The U.S. Geological Survey (USGS) recently launched a pilot Integrated Water Availability Assessment (IWAA) in the Delaware River Basin to explore, test, and refine systems and processes for assessing water availability for human and ecological uses based on water monitoring data. Water-quality monitoring provides citizens, managers, and scientists with the information needed to evaluate the health of aquatic ecosystems and the safety and availability of water for drinking, agriculture, recreation, and other uses. Many organizations collect water-quality data at various sites and sampling frequencies to meet their assessment needs. The result is multiple individual datasets suitable for the specific organization's needs that also hold great potential if pooled into a much larger dataset sourced from multiple organizations (multisource data). A multisource dataset increases the value and power of multiple single datasets and expands the breadth and depth of available water-quality data to ultimately increase the number and types of questions that can be answered. This fact sheet describes the process of "harmonizing" water-quality data from multiple organizations and presents a recently developed dataset for surface-water quality in the Delaware River Basin. This harmonized multisource surface-water-quality dataset will serve as a resource for analysis and modeling of surface-water quality to support IWAA efforts in the basin. Furthermore, this harmonization process can be expanded and applied to other regional IWAA basins or applied nationally.

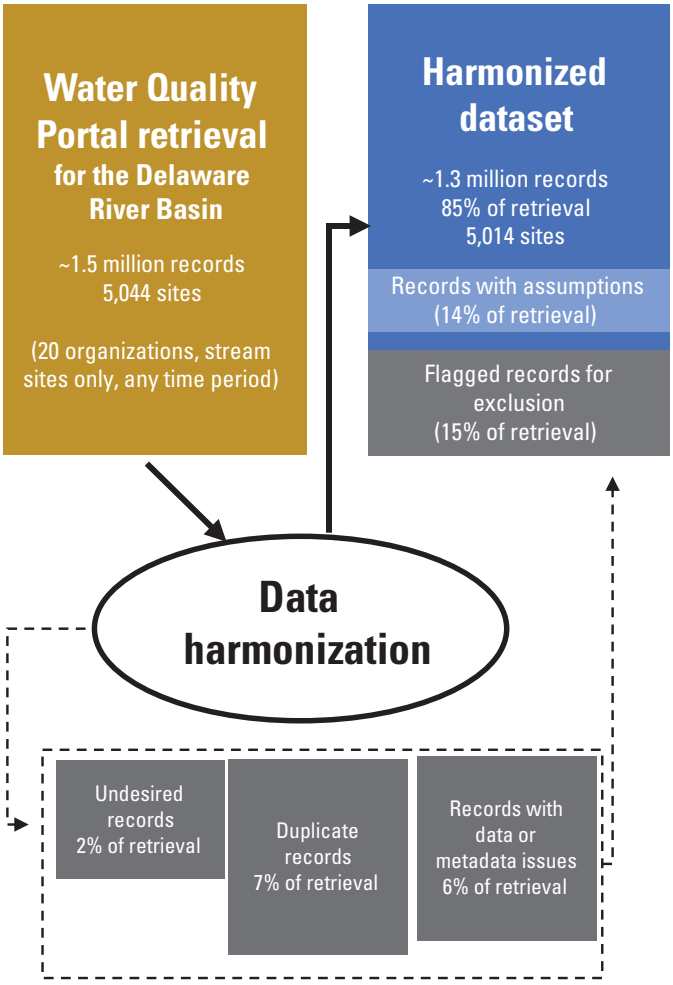

Figure 1. Data processing workflow and record types for multisource surface-water-quality data in the Delaware River Basin.

Data Harmonization

The Water Quality Portal (WQP) is a widely used repository for water-quality data (National Water Quality Monitoring Council, 2019). Data from multiple organizations can be retrieved directly from the WQP, which is a substantial improvement to obtaining and compiling data from multiple organizations individually. Considerable time and expertise are still necessary, however, to create a "harmonized" dataset—meaning a quality-controlled dataset with consistent nomenclature that can be used for analysis and modeling.

The harmonization process includes the following steps: selecting and combining appropriate water-quality parameter names, identifying comparable and problematic methods, "cleaning" text fields, converting data to common units, identifying proper fractionation, assessing missing, zero, or unrealistic result values, synthesizing remark codes and data quality comments, and resolving duplicate records. These steps build on the original data from the WQP and create a harmonized dataset for use in water-quality investigations. The original data retrieved from the WQP are preserved during the harmonization process. Flags are added to specific records to indicate duplicates, potentially erroneous data, metadata issues, or assumptions that were made to account for missing or ambiguous information. Records for undesired parameters or methods are also flagged. Users can decide which data to omit by evaluating and selecting these flags.

Metadata Issues

As with many large complicated datasets, metadata issues present challenges for secondary data use that require intensive review and quality control. Out of the more than 1.5 million water-quality records retrieved from the WQP for the Delaware River Basin, approximately 6 percent of the

The Water Quality Portal

The Water Quality Portal (WQP; https://www.waterqualitydata.us/) is a publicly accessible data repository containing physical, chemical, and biological water-quality data collected at more than 1.5 million sites by over 400 Federal, State, Tribal, and local organizations across the Nation. This effort allows researchers and water managers to store and access their data for use in their planned studies (considered primary data use) and facilitates secondary use of the data by the originators and others, thus adding substantial value to the data by increasing its discoverability.

retrieved records had some sort of metadata or data issue, mostly ambiguous units of measure, leading to their exclusion from the harmonized dataset (fig. 1). About 7 percent of the retrieved records were identified as duplicates, defined as having the same collecting organization, site, sample date, sample time, parameter, and fraction. Records containing ambiguous metadata were also common; 14 percent of the records were ambiguous but were able to be resolved by using an assumption. A common assumption made during the harmonization process was the meaning of the word "total." For example, "total nitrogen" was assumed to mean an unfiltered sample of all forms of nitrogen if a collecting organization only reported "total" and "dissolved" sample fractions. However, if an organization also reported "unfiltered" fractions or only reported "total," then it was unclear if "total" meant an unfiltered sample or the sum of all forms of nitrogen. In these cases, it was ambiguous if "total" nitrogen determinations were from a filtered or unfiltered sample, and these samples were flagged for exclusion. In total, about a third of the retrieved records required an assumption or were undesired, duplicate, or flagged with a metadata or data issue (fig. 1).

Harmonized Water-Quality Data for the Delaware River Basin

To leverage existing multisource data for the Delaware River Basin for IWAA analyses, the USGS retrieved surface-water-quality data for nutrients, salinity (total dissolved solids and specific conductance), major ions, sediment (suspended-sediment concentration and total suspended solids), field parameters (temperature, dissolved oxygen, pH, and turbidity), and bacteria (total coliforms, fecal coliforms, *Escherichia coli* and Enterococcus) from the WQP. These data were harmonized, and water-quality sites were matched to nearby USGS streamgages. The dataset, streamgage pairings, and related information are available in the data release associated with this fact sheet (Shoda and others, 2019).

Twenty organizations contributed to the harmonized dataset through the WQP (fig. 2). The timeframe of the harmonized dataset is from the early 1900s to present day, with data from more than 14,000 days when a water sample was collected or field parameter was measured in a river or stream in the Delaware River Basin. State and local organizations were the source for about a third of the samples in the harmonized dataset, with Federal agencies supplying the rest. More than 90 percent of the samples were collected by one of five organizations:

- Delaware Department of Natural Resources and Environmental Control
- Delaware River Basin Commission
- New Jersey Department of Environmental Protection
- Pennsylvania Department of Environmental Protection
- USGS

The harmonized dataset was used to identify preliminary lists of sites with recent and long-term data based on various sampling frequency and duration scenarios. Recent sampling, characterized as either monthly or quarterly sampling for at least 2 of the 3 years from 2016 to 2018, occurred at approximately 3 percent of the sites in the harmonized dataset. About 50–60 percent of these sites also had long-term sampling, characterized as quarterly sampling for at least 80 percent of a decadal or multidecadal period (ranging from 2008–18 to 1978–2018).

Streamflow data are essential to understanding how water quality varies and changes over time. Of the nearly 5,000 water-quality sites in the harmonized dataset, only about a third were matched to a USGS streamgage with a drainage area that differed from the water-quality site by no more than 10 percent. About 40 percent of the sites with recent water-quality sampling and about 50 percent of the sites with 1 to 2 decades of water-quality data have a nearby and recently active (as of 2016–18) streamgage. Most long-term sites with 30- to 40-year records have a nearby and recently active streamgage.

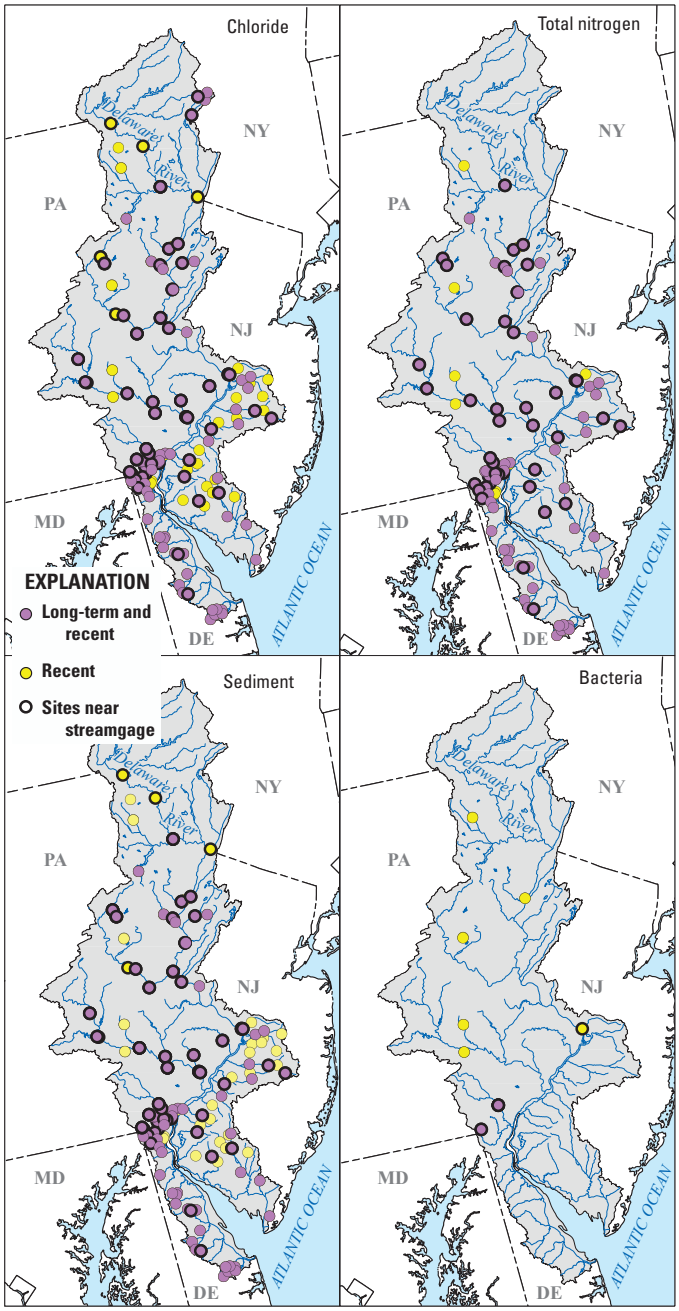

Figure 2. Delaware River Basin stream sites from the harmonized dataset meeting recent (2016–18) and long-term sampling frequency and duration screens, with and without a nearby and active (as of 2016–18) U.S. Geological Survey streamgage. Base from U.S. Geological Survey digital data, 1:100,000 USA Albers Equal Area Conic USGS version projection, North American Datum of 1983.

References Cited

National Water Quality Monitoring Council, 2019, Water Quality Portal, accessed February 2019 at https://www.waterqualitydata.us/.

Shoda, M.E., Murphy, J.C., Falcone, J.A., and Duris, J.W., 2019, Multisource surface-water-quality data and U.S. Geological Survey streamgage match for the Delaware River Basin: U.S. Geological Survey data release, https://doi.org/10.5066/P9PX8LZO.

By Jennifer C. Murphy and Megan E. Shoda

ISSN 2327-6932 (online) https://doi.org/10.3133/fs20203006