The purpose of this document is to provide an order and general description of the steps taken to access, download, clean, and tag PFAS surface water and tissue data supplied by the US EPA’s Water Quality Portal. Using this documentation and prepared code, individuals will be able to interpret national or regional trends in PFAS contamination by analyzing data from the Water Quality Portal. Only two PFAS compounds, perfluorooctanoic acid (PFOA) and perfluorooctanesulfonate (PFOS) are evaluated within the code described herein, though other compounds are available in the dataset for expanded analytical efforts. This document demonstrates the use of the R package “EPATADA” to assist with coding processes that were used to generate the results of a nationwide study of ambient surface water PFAS concentrations in the US. EPATADA is free to download package for the open source R software, and was designed by the US EPA and detailed here: <https://github.com/USEPA/EPATADA/blob/develop/vignettes/TADAModule1.Rmd>. The code, data, and supporting documentation are all available free of charge. The document is broken down into sections that outline how each attached .R file can be used in sequence to evaluate surface water data in a reproducible manner.

Included .R files: data\_pull.R, data\_processing.R

Authors: Hannah Ferriby1, Kateri Salk1, Matt Dunn1, Chris Wharton1, Tammy Newcomer-Johnson2

Affiliations: 1. Tetra Tech Inc. 2. United States Environmental Protection Agency

Corresponding Author: Tammy Newcomer-Johnson – newcomber-johnson.tammy@epa.gov

**Outline of data\_pull.R:**

1. Setup and Package Installation: The following packages are required to perform the analysis presented in this study.
   1. Check and Install Dependencies:
      1. Checks for the *remotes* package and installs it if not already present.
      2. Installs the *EPATADA* package from GitHub.
   2. Load Required Libraries:
      1. Required R libraries (*EPATADA*, *tidyverse, scatterpie, ggplot2, sf, scales*) loaded.
2. Data Retrieval: EPATADA package is used to query and pull data from US EPA’s Water Quality Portal based on the characteristicName.
   1. Using TADA\_BigDataRetrieval:
      1. Pulls data for PFAS-related compounds from a specified source using predefined characteristics:
         1. characteristicType: Includes broad PFAS categories (e.g., "PFAS, Perfluorinated Alkyl Substance").
         2. characteristicName: Targets specific PFAS compounds (e.g., “Perfluorooctanoic acid”, “Perfluorooctanesulfonate”).
         3. Write Results of Data Pull to CSV: The unfiltered, uncleaned dataset is exported as a CSV file (data\_pull.csv) for further processing or analysis.
3. Initial Plotting to Create Map of all matrices and data
   1. Filter by States with data: Create new dataset containing only states with available surface water and/or tissue data
   2. Create *scatterpie* map with circle radius representing number of samples present per state
   3. Including all matrices and all unfiltered data for PFOA and PFOS

**Outline of data\_processing.R:**

1. Setup and Initial Data Import: Loading new packages and the recently exported data for further analysis.
   1. Load Packages: Required R libraries (*EPATADA*, *tidyverse, sf, ggplot2, scatterpie, scales*) loaded.
   2. Data Import: PFAS dataset from data pull (data\_export.csv) imported.
   3. Change characteristicName to figure-friendly abbreviations
2. Harmonize state name, location, and state codes
   1. Load state codes: state\_codes.txt
   2. Load in shape file (.shp) of United States for later plotting
3. Filter Data for PFAS Compounds: Pull only samples we can identify as surface water and tissue for the pre-selected PFAS
   1. Filter by Media and MediaSubdivisionName: Retain only records from surface water samples and tissue samples.
   2. Filter by Compound: Ensure samples are from a predefined list of PFAS compounds (perfluorooctanoic acid and perfluorooctanesulfonate).
   3. Harmonize Names: Abbreviate specific PFAS compound names for consistency.
4. Preparing Initial Summary Maps with Unfiltered Data
   1. Filter by States with data: Create new dataset containing only states with available surface water and/or tissue data
   2. Create *scatterpie* map with circle radius representing number of samples present per state
   3. Including only unfiltered surface water and tissue data
5. Quality Control (QC) Tagging: Add tags for quality control parameters of specific concern to our analysis such as units and methods used. The following steps includes processes designed by EPATADA’s development team to automate quality control and cleaning processes for data downloaded from the Water Quality Portal.
   1. Result Unit Validity: Adds a tag (TADA.ResultUnit.Flag) to validate measurement units.
   2. Sample Fraction Validity: Tags sample fraction (e.g., dissolved, total) issues (TADA.SampleFraction.Flag).
   3. Method Speciation Validity: Tags issues with method speciation (TADA.MethodSpeciation.Flag).
   4. Harmonization: Harmonizes synonyms across records for characteristic/fraction/speciation/unit, adding tags like TADA.Harmonized.Flag.
6. Tagging for Outliers and Thresholds: Add tags for evaluation of potentially erroneous data.
   1. Unrealistic Values: Tags values above upper or below lower thresholds.
   2. Continuous Data: Identifies continuous data (commented out in this code).
7. Method Check: Identify which method was applied to measure PFAS in each sample
   1. Analytical Methods: Tags data based on the validity of analytical methods used (TADA.AnalyticalMethod.Flag).
8. Duplicates Check: Identify any duplicate samples in dataset.
   * 1. Tags potential duplicates across multiple organizations (TADA.MultipleOrgDupGroupID).
     2. Tags single-organization duplicates (TADA.SingleOrgDup.Flag).
9. Quality Control Activity Identification: Check for any quality control flags that were uploaded by the original data source.
   1. QC Samples: Tags data from QC-related activities (TADA.ActivityType.Flag).
10. Coordinate Validation: Confirm that coordinates of sampling location are accurate/realistic.
    1. Invalid Coordinates: Tags records with problematic coordinates, e.g., locations outside the United States (TADA.InvalidCoordinates.Flag).
11. Other QC and Non-Detect Processing: Identify which samples were uploaded as non-detects or suspect by the original data source.
    1. Suspect Samples: Tags records with suspect qualifier codes (TADA.MeasureQualifierCode.Flag).
12. Non-Detect Values: Adds columns for censored data tags (TADA.CensoredData.Flag) and replaces non-detects with calculated values (e.g., 50% of the detection limit).
13. Column and Data Cleanup: Remove NA columns to tidy up the result .CSV file and dataset.
    1. Remove All-NA Columns: Drops columns that contain only NA values.
14. Filter Negative Values: Removes records with negative result values.
15. Detection Limit QC: Manual QC check to review units used for detection limits
16. Harmonizing detection limit units. Also evaluating if any samples were reported at a concentration that is lower than their corresponding detection limit.
    1. Detection Limit Tags:
       1. Tags samples based on detection limits and units.
       2. Converts and harmonizes detection limit units (e.g., converting µg/L to ng/L).
    2. Detection Limit Statistics:
       1. Calculates statistical summaries for detection limits (average, median, standard deviation).
       2. Tags outliers based on calculated thresholds.
17. Detection Limit User Error: Tag samples that are entered as ‘uncensored’ but are lower than the reported detection limit.
    1. Labeling:
       1. Uncensored: Reported concentration is higher than reported detection limit
       2. Non-detect: Concentration is reported as non-detect or below detection limit
       3. Unknown: Reported concentration is lower than reported detection limit
18. EPA Method Validation: Add a tag if the sample can be identified as using an EPA analytical method for PFAS in water.
    1. EPA Method Tags: Tags records based on accepted EPA analytical methods.
19. Export Processed Data: Export this processed, cleaned, and tagged data into shareable file.
    1. With Tags: Exports the dataset with all tags to a CSV (EPATADA\_Original\_data\_with\_flags\_tags.csv).
    2. Filtered Data: Applies a series of filters to exclude invalid data, retaining only acceptable samples. Exports this filtered dataset (EPATADA\_priority\_filtered\_data.csv).
20. Preparing Initial Summary Maps with Filtered Data
    1. Filter by States with data: Create new dataset containing only states with available surface water and/or tissue data
    2. Create *scatterpie* map with circle radius representing number of samples present per state