**Protocol for downloading, cleaning, and selecting water quality data**

Current sources supported to varying degrees: National water quality monitoring council (“Portal”), Water Atlas. If another data source is to be used, please add the protocol for data collection to this file and notify repo admin of the additional source to update the compilation and selection R code.

***Please note, mapping projects are currently undergoing reorganization. File references may be inaccurate depending on where in the phase the document is reviewed. Please address all questions to repo admin.***

1. **Determine boundaries of desired area/estuary**
   1. Refer to Table 1 for boundaries. Add updates to Table 1 as needed. Broad boundaries can be found in the “Existing data tracking” file on the “StateGrid Assignment” tab. Adjust as necessary for best cover the estuary of focus.
   2. Confirm existence of KML area file in the Water-Quality-Processing/KML folder. Create file if needed by following Site\_Section\_file\_creation\_SOP document and naming the file using the code in Table 2.
   3. Ensure files are named appropriately based on estuary, data source (Table 2), and years of data.
   4. Determine water quality data sources and refer to the following sections as needed:
      1. National water quality monitoring council (Section II)
      2. (Section III)
2. **Retrieve water quality data from the National water quality monitoring council (“Portal”)**
   1. Go to <https://www.waterqualitydata.us/portal/> to download the water quality data for the desired area.
   2. Select the “Advanced” option near the top of the screen.
   3. In the “Select Location Parameters” section:
      1. A screenshot of a computer

         Description automatically generated with medium confidenceWithin the ‘Bounding Box’ area, enter in the bounding box coordinates using decimal degrees. Do not put anything in the country, state, or county boxes.
      2. Under ‘Site Type’ select “Estuary”, “Ocean”, “Wetland”, and “Stream”.
   4. Under the “Filter Results” section:
      1. In the “Sample Media” box, select “Water”.
      2. In the “Date Range” section, enter the starting and ending dates for the time period (5-year increments is best).
      3. In the “Minimum Sample Activities per Site” and “Minimum Results per Site” enter “12”.
   5. A screenshot of a computer

      Description automatically generatedUnder the “Download the Data” section:
      1. Make sure all items under “Data Source” are selected.
      2. File format: MS Excel 2007+
   6. Select “Site Data Only” under the “Data Profiles” area and select “Download” at the bottom of the page. Select “Continue” in the pop-up window.
   7. Navigate to where the downloaded file is saved, rename the file using the estuary and data source naming convention from Table 2, “Site data”, starting year, and ending year separated by underscores (ex. CR\_Portal\_Site data\_2012\_2022). Save the file to the WQ “Raw\_data” folder.
      1. Name files with years indicated by 4 digits (e.g., 2020, 2025).
      2. If data is not for complete years, a two-digit month can be added to the file name. Do NOT include underscores within the data range (e.g., 062020\_082025).
   8. Return to the webpage and select “Sample Results (narrow)” under “Data Profiles”. Select “Download” at the bottom of the page. Select “Continue” in the pop-up window. If the results are too large, modify the date range by cutting them into 2 or more sections. Repeat 9 with the updated date ranges in separate files.
   9. Navigate to where the downloaded file is saved, rename the file using the estuary code and data source naming convention from Table 2, “Results”, starting year, and ending year separated by underscores (ex. CR\_Portal\_Results\_2012\_2022). Make sure to use the same date naming convention as mentioned in step g.i-ii. Save the file to the WQ “Raw\_data” folder.
3. **Retrieve water quality data from the water atlas** (counties included: Lake, Manatee, Orange, Polk, Sarasota, Seminole, Pinellas, & CHENP).
   1. Go to the primary water atlas web page at <https://wateratlas.usf.edu/> and select one the desired atlases. All atlases can be downloaded together in the data selection process, so it isn’t as important which is chosen in this step.
   2. Under the “Maps/Data” section, chose “Data Download”.
   3. Accept the disclaimer then select the data type:
      1. Surface Water Quality – primary data
      2. Surface Water Hydrology – flow and water level data
   4. Select the following parameters then select Submit:
      1. By Location: Water atlas, and watershed
      2. By Site Info: Water body type
      3. By Sample Info: Date range, Parameter
   5. Complete the selection page as follows:
      1. Water Atlas: select desired data sources
      2. Watershed: select desired watersheds from list
      3. Water Body Type: select Bay and Stream/River
      4. Date Range: Specify date range for data
      5. Parameter: Select parameters. Include all the following plus any additional as needed. Parameters can be searched for by name using the search box. Select submit once all parameters have been selected.

* Depth, bottom, ft
* Secchi disk depth, ft
* Salinity, PSS and PPT
* Dissolved oxygen (DO)
* Dissolved oxygen saturation (%)
* Temperature, water (deg C & deg F)
* pH
* Turbidity (NTU)
* Chlorophyll a (probe relative fluorescence)
* Chlorophyll a, corrected for pheophytin
* Total suspended solids
  + 1. Select “Give me all station data” on the next screen (Step 4 of 5).
    2. Compile data to download by specifying “Excel” as the file type and “Row” as the file format and selecting “Generate file to download”. Then “Download File” once the file has been generated.
  1. Navigate to where the downloaded file is saved, rename the file using the estuary code and data source naming convention from Table 2, “Site data”, starting year, and ending year separated by underscores (ex. CR\_WA\_Site data\_2012\_2022). Save the file to the WQ “Raw\_data” folder.
     1. Name files with years indicated by 4 digits (e.g., 2020, 2025).
     2. If data is not for complete years, a two-digit month can be added to the file name. Do NOT include underscores within the data range (e.g., 062020\_082025).
     3. Open the file to make sure the data is converted properly: Select Yes to open and read the file, then select Convert in the 2 pop-up windows.
     4. Once the data is converted, select columns A (WBodyID) through column F (Actual\_StationID) and change cell type to ‘Text’.
     5. Save the file as a .xlsx file using the same name.

1. **Clean raw data and map station locations**
   1. The water quality R Project should be run from the “Water quality” folder.
   2. Save all “Site data” and “Results” files to the “Raw\_data” folder in your local drive. If working on the network, add files to “Raw\_data” folder while working on cleaning the data but consider moving to a local storage location once data is cleaned unless the data will need to be accessed often.
   3. A screenshot of a computer

      AI-generated content may be incorrect.Open the Water Quality Processing R project located in the main folder.



* 1. Open the WQ\_data\_compilation R code file by navigating to the ‘Code’ folder in the view window and the selecting the code file.
     1. Run the *Load required packages* section of code and update the Estuary\_code, Data\_source, and years of data range as specified in the raw data files.

Text

AI-generated content may be incorrect.

* + 1. If working with FIM data, skip the “Location\_data” and “Results” lines. If working with Portal or Water Atlas data, in the *Load Files* section, run the “Location\_data” line without changing the code. Next run the “Results\_data” line if using only one results file. If more than one results file is needed, run the “Results1” line through “ResultsN” for the number of files needing to be uploaded.
    2. Continue by running the “Estuary area” and “State outline” code.
    3. In the *Select data columns* section, run the code without changing anything unless different columns are needed in the output. Items listed below are the current filtering performed on the data. If additional data or columns are required, the code can be updated and saved for estuary- or project-specific use.
       1. Minimum column requirements for location data (“keep\_site”):
* MonitoringLocationIdentifier
* OrganizationIdentifier
* OrganizationFormalName
* MonitoringLocationName
* MonitoringLocationTypeName
* MonitoringLocationDescriptionText
* LatitudeMeasure
* LongitudeMeasure
* HorizontalCoordinateReferenceSystemDatumName
* StateCode
* CountyCode
* ProviderName
  + - 1. Minimum column requirements for results data (“keep\_results”):
         1. Portal:
         * MonitoringLocationIdentifier
         * ResultIdentifier
         * ActivityStartDate
         * ActivityStartTime/Time
         * ActivityStartTime/TimeZoneCode
         * CharacteristicName
         * ResultMeasureValue
         * ResultMeasure/MeasureUnitCode
         1. Atlas: (TBD)
    1. In the *Combine data by station* section, run the code without changing anything unless different characteristic/parameters are needed in the output.
       1. Minimum parameter “Characters” to keep:
          - Salinity
          - Temperature, water
          - Depth, bottom
          - Depth, Secchi disk depth
          - Temperature, air, deg C
          - Turbidity
          - Conductivity
          - Specific conductance
          - pH
          - Dissolved oxygen (DO)
          - Dissolved oxygen saturation
          - Chlorophyll a, corrected for pheophytin
          - Chlorophyll a
          - Total dissolved solids
          - Total suspended solids
          - Zooplankton
          - Diatoms
          - Stream flow, instantaneous
          - Flow, severity (choice list)
          - Stream stage
          - Flow
          - Stream flow, mean. Daily
    2. In the *Map of stations and limitation of stations to estuary area* section, run the code without changing anything unless different mapping is needed. Runing the “saveWidget()” line will save the map to the “Maps” folder for future reference.
    3. In the *Clean parameter data* and *Save filtered data* sections, run the code as is unless other corrections are needed. This will combine and clean the data, then save the cleaned data to the “Raw\_cleaned” folder as a “combined\_filtered” Excel file.
    4. Additional notes:
       1. Data compilation will have to be performed for each data source independently. Data from different sources can be joined together later.
  1. Once finished with the “WQ\_data\_compilation” R code file, determine what point locations (i.e., monitoring stations, central points of areas, etc.) will be needed for mapping or data selection. Fixed monitoring stations can be displayed in maps and used for selecting data stations in the next section. Additional point locations can be added for selecting data stations in the next section. Check for required point locations in the Data/ Reference-data/Stations\_area\_selections file and add as necessary. This file is maintained on the network (or in the main branch if working with git) for other use.
  2. Open the “WQ\_data\_selection” R code file to select desired data. This code will help output the selected data and a map of selected WQ stations to the “Compiled\_data” and “Maps” folders, respectively.
     1. Run the *Load require packages* section and updated parameters in the *Setup* section.
     2. Run the next two sections of code (*Limit date ranges* and *Station map*) without changes to load combined data, filter by dates, and output WQ station map to view station locations.
     3. In the *Station Selection* sections, data can be limited according to the sections listed below. Follow in-text code annotations for guidance on how to limit data. Within each section is code that will save the filtered data to the “Final\_data” folder.
        1. *Specified buffer distance –* Buffer criteria specifying distance from stations using on one or two “buffer” zones – specified locations as listed in the Excel file found in Data/Reference\_data/Stations\_area\_selections.
        2. *Closet N stations –* Closest N number of stations to specified locations as listed in the Excel file found in Data/Reference\_data/Stations\_area\_selections. A maximum distance for consideration can also be set.
        3. *By station name or designated boundary –* Specific boundary or locations via:
           1. Station names
           2. Bounding box area
     4. All data can be saved by using the *Output all stations* section.

1. **Combination of data from different sources**
   1. Make sure that all data selection methods have been used for all data sources required for data combination and that all output files are located in the “Compiled\_data” folder.
   2. Open the “WQ\_data\_combination” R code file to combine data. Establish set up specs referring to notes below as needed:
      1. Estuary\_code – two letter estuary code
      2. Data\_sources – list of all data sources to combine into final file: "Portal", "FIM"
      3. Project\_codes – list of any project code IDs used. This is the short code used in naming the compiled file and comes before the start and end years
      4. Data\_selection\_method – list of data selection methods used. If used all data, enter NA without quotation marks. Other selection methods should match wording in data file names. Selection methods can be combined together and can be combined with NA (all data)
      5. Final\_code – short code to distinguish what data compiled is used for
      6. Start\_year – first year of data included
      7. End\_year – last year of data included
   3. Run the “Load files” code to load all files that fit the setup parameters. Code will provide a list of the data loaded, summary information for data, and a list of file names that were loaded. Warning messages will show file name combinations search for but not located-this can be used as a secondary check that all desired data was checked for.
   4. Run the “Combine and output” code to combine all data into one data file saved to an Excel file with basic summary of data included. Data columns in the final data file include:

* Estuary – estuary two-letter code
* Date – date of sample
* WQ\_ID – unique water quality station ID (from source)
* Longitude
* Latitude
* KML – point information as “In” KML shape area or “Out” of KML shape area
* Parameter – measurement type
* Measurement – value of measurement
* Result\_Unit – unit of measurement
* Station – station relation if selected to be near fixed locations
* Buffer – buffer distance if data selected by buffer zone
* Source – source of data

***Not all data sources have been added yet. Please contact E Williams for any requested updates.***

**TABLES**

Table . North, south, east, and west boundaries of estuary or region indicated. Long name of area is given (Estuary), and existence of KML file noted by two-letter code for the area. Update as appropriate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Estuary** | **North** | **South** | **East** | **West** | **Estuary KML** |
| Caloosahatchee | 27.1 | 26.1 | -81.4 | -82.4 | CR |
| St Lucie | 27.3 | 27.1 | -80.1 | -80.3 | SL |
| Lake Worth | 26.84362 | 26.527 | -80.03102 | -80.05384 | LW |
| Loxahatchee | 27.0 | 26.9 | -80.0 | -80.2 | LX |
| Tampa Bay | 28.7 | 27.45 | -82.3 | -82.9 | TB |
| St Andrews | 30.3 | 30 | -85.35 | -85.9 | SA |
| Pensacola | 30.7 | 30.66 | -86.8 | -87.4 | PE |

Table 2. Naming conventions to be used for file names. Estuary or data source indicated by shorthand can be found in column 2. Update as appropriate.

|  |  |
| --- | --- |
| **Naming convention for files** | **Estuary or data source** |
| *Estuaries* | |
| CR | Caloosahatchee |
| SL | St. Lucie |
| LW | Lake Worth |
| LX | Loxahatchee |
| SA | St. Andrews |
| PE | Pensacola |
| *Data sources* | |
| Portal | National water quality monitoring council |
| WA | Water atlas sources |

Table . Description of folder, file, and data types of water quality data created using the R Project.

|  |  |
| --- | --- |
| **Folder Name** | **Description** |
| Compiled\_data | Cleaned data (from Raw\_cleaned files) that has been:   * combined within data sources based on files specified, * limited to the desired date range, and either * selected to include all stations, * selected stations based on buffer areas, * selected stations based on proximity to locations, * selected stations by name, or * selected stations within a specified boundary.   Name of file output includes a short code to describe intended use/project of data (e.g., SLCAGE, LWQRT). Data output uses source formatting. *Output of WQ\_data\_selection R code.* |
| Final\_data | Data file of selected WQ data combined from various data types into one formatting. Output file used for HSM models.  *Output of data from WQ\_data\_combination R file.* |
| Raw\_cleaned | Data cleaned but still in the data source formatting.  Data is:   * limited to desired columns and parameters, * designated as located “In” the estuary KML shape or “Out”, * and has the same units for each parameter.   Data file is divided by estuary, data source, data type (site data vs. results), and years of data.  *Output of WQ\_data\_compilation R file.* |
| Raw\_data | Data downloaded from data sources without cleaning or modification.  Divided by estuary, data source, data type (site data vs. results), and years of data. |
| Reference\_data | Reference data for identifying fixed station locations for water quality station selection and mapping of stations in relation to each other. |