Chesapeake Data Explorer Data Dictionary

Water Quality Data Introduction

Water Quality data collected by the CMC includes baseline chemical and bacteria parameters, collected *in situ* or analyzed at a lab. All groups do not collect the same parameters across the watershed; instead the CMC has a universe of parameters to choose from and each group selects the specific parameters that match their data use goals and questions of interest. Equipment and methods are then selected for each parameter based on cost, quality assurance feasibility, and data use goal (desired tier level). Parameter codes are then assigned to data for upload to the Data Explorer. These codes are defined by the following criteria:

- General equipment/methods selected (ie. *In situ* probe, field kit, grab sample for lab analysis)
- Final tier level (based on the equipment and QA standards in place)
- Any other nuances based on the individual equipment selected – units measured, depth profile vs. surface sample, etc.

The goal being that each individual parameter may have multiple parameter codes for specific methods and equipment to account for various quality assurance protocols present or missing. For example, dissolved oxygen can be sampled with a Winkler titration kit that has the capability of being Tier 2 because it can be standardized. The individual group can choose to standardize the kit before each use and the data is Tier 2, or not standardize the kit and the data is Tier 1. Each piece of equipment has its own criteria for Tier 1 and Tier 2, but the intent is *all Tier 2 data across the dissolved oxygen parameter are equivalent in quality.*

Water Quality data are collected on (1) a monthly basis year-round, (2) biweekly/weekly basis during the summer (May through September), or (3) a combination of biweekly/weekly sampling during the summer and monthly sampling the rest of the year. Samples are collected from either the surface of the water (at either 0.3m, 0.5m or 1.0m depending on the state and equipment capabilities), or as a depth profile (measurements taken 1m off the bottom, in 1m intervals up through the water column, and ending at one of the surface sample depths). A complete dataset for a given year is generally defined as an 80% completion rate based on the number of samples anticipated. For example, if a monthly program anticipates collecting 12 samples a year, if they actually collect 10-12 samples the dataset is considered complete.

- Monthly, year-round sampling is typically used to track trends over time from year to year.
    - Samples are generally taken on the same day of the month, at roughly the same time, in order to capture a variety of weather and stream conditions. These sites are typically smaller non-tidal streams, sampled at the surface, and by individual monitors. The sample is collected one of three ways: by wading into the stream, collecting from a dock or bridge, or standing on the shoreline.
- Weekly or Biweekly summer sampling has two main purposes:
    - Tidal programs that sample main stem tributaries collecting depth profile data intended to track dissolved oxygen levels and dead zones. These are typically sampled by boat, on a regularly scheduled day of the week that collects data for multiple stations in a sampling run.
    - Bacteria programs that sample commonly used public access points. These programs are set up to monitor either on a Wednesday or Thursday, in order to process the sample (takes 24 hours) by Friday. This data is then used to make a recommendation to the public about contact with that waterway based on human contact standards set by the state.

Water Quality Data Download File (cmcWaterQualitySamples-date)

This file contains the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file.

Date: The date the sample was taken, formatted month/day/year
Time: The time the sample was taken, formatted hrs:mm:ss
StationName: The station ID or name assigned by the individual monitoring group, usually used on the field datasheets to identify the station.
StationCode: The unique code for the station assigned in the Chesapeake Data Explorer, formatted GroupCode.StationName. This is used to ensure the data is assigned to the correct group and station in the event that two groups have the same station name. This field can be used to find more information on each station in the stations metadata file.
Latitude/Longitude: The latitude and longitude coordinate for the station, formatted in decimal degrees.
GroupCode: The unique code that identifies the group that collected the data. This field can be used to find more information on each group in the groups metadata file.
SampleId: Indicates whether the sample is a duplicate/replicate, the first sample will have a 1 in this column and a duplicate/replicate will have a 2. Each duplicate/replicate will appear on separate rows.
SampleDepth: Indicates the depth the sample was taken, in meters. Each depth will appear on separate rows, air temperature is the only indicator that does not have an associated sample depth.
Parameters: Starting in Column J is the actual water quality data listed by parameter code. Each file will contain just the parameter codes included in that particular dataset, see Appendix 1 Table 1 for a full list of parameter codes. The parameters will appear in alphabetical order and each parameter code will have 6 associated columns in the same order:

1. Parameter value (Parameter Name (code)) – this is the actual data point value collected in the field.
2. Parameter code (Parameter Name (code) Code) – this is the parameter code that links that data point to the general parameter measured, equipment used and tier level.
3. Problem Code (Parameter Name (code) Problem) – this field indicates any QA issues or flags with that particular data point. See Appendix 1 Table 2 for a full list of problem codes and what they mean.
4. Qualifier Code (Parameter Name (code) Qualifier) – this field indicates if the value reported is actually above or below a detection limit for that piece of equipment. See Appendix 1 Table 3 for a full list of qualifier codes.
5. Tier Level (Parameter Name (code) Tier) – this field indicates the tier level of that parameter.
6. Units (Parameter Name (code) Units) – this field indicates the units used to measure that parameter.

Station Conditions: After the water quality data, if any observational or station condition data is recorded, it will appear at the end of the spreadsheet and will only contain the fields that contain data for that particular dataset. See Appendix 1 Table 4 for a full list of station conditions.
Modified Date: The date that the sampling event was upload to the Chesapeake Data Explorer or modified after upload.
Comments: This field includes any additional comments made by the monitor, program coordinator or CMC team about that sampling event. Can include QA, equipment, or sampling issues.

Water Quality Groups Metadata File (cmcGroups)

This file contains the group metadata information for the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file.

Code: The unique code that identifies the data collection group. This code can be used to reference the group associated with a data point in the water quality data download file.
Name: The full name of the data collection group.
ContactName: The main point of contact, usually a group coordinator, for the monitoring program.
ContactEmail: The email address for the main point of contact for the monitoring program.
Cell and Office Phone numbers: The phone numbers for the main point of contact for the monitoring program.
Description: A description of the monitoring program, usually includes details about the QAPP used and partner organizations.
URL: Website, if applicable
Address: The address for the office or point of contact for the monitoring program

Water Quality Stations Metadata File (cmcStations)

This file contains the station metadata information for the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file.

Code: The unique code for the station assigned in the Chesapeake Data Explorer, formatted GroupCode.StationName. This is used to ensure the data is assigned to the correct group and station in the event that two groups have the same station name. This code can be used to reference the station associated with a data point in the water quality data download file.
Name: The station ID assigned by the individual monitoring group.
NameLong: A longer station name that usually includes more details about the waterway or station.
Latitude/Longitude: The latitude and longitude coordinate for the station, formatted in decimal degrees.
***CBseg:** Refers to the Chesapeake Bay segmentation scheme.
***WaterBody:** Indicates the HUC12 watershed name associated with the station.
Description: Includes any additional information or descriptions about the site.
***Datum:** All stations are in NAD83 datum.
***CityCounty:** Includes information about the City or County the station is located in.
***Tidal:** Indicates whether the station is considered Tidal (True) or Non-tidal (False). This data is automatically populated by the Chesapeake Data Explorer by referencing the Chesapeake Bay Programs designation.
Comments: Any additional comments or information about the site not captured in the description.
***Fips:** The FIPS number for the Station.
***HUC12:** The HUC12 watershed number associated with the station.
***State:** Indicates the state in which the station is located.
Huc6Name: Indicates the HUC6 watershed the station is located within, based on the watershed name not the HUC number.

*Indicates fields that are automatically populated by the Chesapeake Bay Program through an API connected to the Chesapeake Data Explorer.

Water Quality Parameters Metadata File (cmcParameters)

This file contains the parameter metadata information for the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file.

Code: The parameter code associated with the parameter included in the water quality data download file. This code can be used to reference the parameter associated with a data point in the water quality data download file.
Name: The broad parameter name.
Units: The units associated with that parameter code.
Method: References the Chesapeake Bay Program method code for data upload to the CBP system. Any codes labelled 999 are not currently in the CBP system.
Tier: Indicates the tier level associated with that parameter code.
Matrix: Indicates the matrix (either air or water) where that parameter is sampled.
Tidal: True indicates that parameter is a tidal water quality parameter, false indicates that parameter is not a tidal water quality parameter.
NonTidal: True indicates that parameter is a non-tidal water quality parameter, false indicates that parameter is not a non-tidal water quality parameter.
Analytical Method: Indicates whether the parameter is measured with a probe, a kit or lab analysis. Each parameter can only have one analytical method.
ApprovedProcedure: References any approved method or procedure (ie. US EPA) used to measure that parameter. Some parameters may have more than one approved procedure.
Equipment: Details the type of equipment used to measure that parameter (ie. Individual Probe) and lists specific manufacturers as applicable. This list is not exhaustive, for more information on the specific equipment used for any given data point, contact a CMC service provider or the group coordinator.
Precision: Lists the precision range of the equipment listed.
Accuracy: Lists the accuracy range of the equipment listed.
Range: Lists the reading range of the equipment listed.
QcCriteria: Lists any additional QC criteria for the equipment listed.
InspectionFreq: Indicates the frequency that piece of equipment needs to be inspected in order to meet the QA requirements for the specified tier, either before each use, annually, or N/A (if no inspection is needed).
InspectionType: Indicates the type of inspection that needs to occur at the frequency listed above in order to meet the QA requirements for the specified tier.
CalibrationFrequency: Indicates the frequency that piece of equipment needs to be calibrated in order to meet the QA requirements for the specified tier, either before each use, annually or N/A (if no calibration is needed).
StandardOrCalInstrumentUsed: Indicates the type of standard or calibration needed at the frequency listed above in order to meet the QA requirements for the specified tier.
TierIIAdditionalReqs: Lists the justification for the tier level associated with that parameter code, this includes justifications for provisional tier status or lowering a tier level. These are generic justifications for tier levels, for more information on the specific tier justifications for any given data point, contact a CMC service provider or the group coordinator.
HoldingTime: The amount of time a sample can be held before analysis.
SamplePreservation: Indicates any sample preservation needs in order to process the sample accurately, mostly applies to lab samples.
Status: Indicates if the parameter is actively being used?

requiresSampleDepth: True indicates the parameter is used to measure depth profile samples, false indicates the parameter is only a surface sample.
isCalibrationParameter: True indicates the parameter code is associated with a calibration reading and not a field reading, false indicates the parameter code is associated with a field reading.
requiresDuplicate: True indicates the parameter can have duplicate or replicate samples taken, false indicates only one sample is recorded.
Description: Gives additional description for the "Name" field.
NonfatalUpperRange/NonfatalLowerRange: Indicates the upper and lower range for that parameter, this is used as a QA check in the Data Explorer. Values can be uploaded that are outside of this range, but a warning is given to the person uploading the data to check for errors.

Water Quality Calibration Samples Metadata File (cmcCalibrationSamples)

This file contains the calibration sample information (if available) for the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file. Calibration samples are not required to be upload with the field data, so this data may not exist in the Data Explorer for every field data point, but that does not mean the equipment was not calibrated. For more information on specific calibrations contact a CMC service provider or the program coordinator.

Date: The date the calibration sample was taken, formatted month/day/year
Time: The time the calibration sample was taken, formatted hrs:mm:ss
StationName: The station ID or name assigned by the individual monitoring group, usually used on the field datasheets to identify the station.
StationCode: The unique code for the station assigned in the Chesapeake Data Explorer, formatted GroupCode.StationName. This is used to ensure the data is assigned to the correct group and station in the event that two groups have the same station name. This field can be used to find more information on each station in the stations metadata file.
Latitude/Longitude: The latitude and longitude coordinate for the station, formatted in decimal degrees.
GroupCode: The unique code that identifies the group that collected the data. This field can be used to find more information on each group in the groups metadata file.
SampleId: Indicates whether the sample is a duplicate/replicate, the first sample will have a 1 in this column and a duplicate/replicate will have a 2. Each duplicate/replicate will appear on separate rows.
Parameters: Starting in Column I is the calibration data listed by parameter. Each file will contain just the calibration parameters included in that particular dataset, see Table X for a full list of calibration parameters. The parameters will appear in alphabetical order and each parameter will have 3 associated columns in the same order:

1. Parameter value (Parameter Name (code))
2. Parameter code (Parameter Name (code) Code) – all calibration codes start with C.
3. Units (Parameter Name (code) Units)

Station Conditions: After the water quality data if any observational or station condition data is recorded it will appear at the end of the spreadsheet and will only contain the fields that contain data for that particular dataset. See Table X for a full list of station conditions.
Comments: This field includes any additional comments made by the monitor, program coordinator or CMC team about that sampling event. Can include QA, equipment, or sampling issues.

Benthic Data Introduction

Benthic sampling is the collection of benthic macroinvertebrates (insects and crustaceans) that live on the bottom (benthos) of a stream bed. For the CMC data, all samples are collected in non-tidal streams. Different types of benthic macroinvertebrates are sensitive to pollution and water quality issues- looking at the composition of a benthic sample can give a holistic stream health score. Benthic data is collected seasonally, although some groups collect only once or twice a year in the spring and fall.

Benthic data is filtered into the following tier levels:

- Tier 1 – identifies bugs to order level
- Tier 2 – identifies bugs to family level

Most of the volunteers trained under the CMC fall under Tier 1 methods through the VASOS or ALLARM monitoring programs.

Common Benthic Metrics

Richness/ Diversity- How diverse is the composition?

Composition- what kind of make up is the composition? A common and effective metric is the % EPT (Ephemeroptera, Plecoptera, Trichoptera) which are the most sensitive benthic macroinvertebrate orders.

Functional Feeding Groups- Different macroinvertebrates fill different roles in the benthic ecosystem—they each fall into a functional feeding group (ie. , filtering, scraping, predator, collecting, gathering)- analysis of these roles can show how "balanced" the benthic ecosystem is

Tolerance- This metric is often used to look at what percentage of the benthic macroinvertebrates collected are tolerant to pollution- benthics usually fall into broad categories of "tolerant", "somewhat tolerant", and "sensitive", and these categories can be used for developing a stream health score

How this data can be analyzed

Benthic data is stored in two ways depending on the tier level. Tier 1 data is almost all order level data, but has been entered into the data explorer as one- or two-letter codes that are linked to the common names of the benthic macroinvertebrates, (see Appendix 2 Table 5), this is then scored using the VASOS metric. Tier 2 data is family level data and entered into the Data Explorer using the scientific names of the benthic macroinvertebrates (see Appendix 2 Table 6), this is not currently scored by any metrics the Data Explorer but can be scored based on the Chessie BIBI metric detailed below. In order to start to assess this data together, Tier 1 data needs to be converted to scientific names.

The index that the ICPRB and the Chesapeake Bay Program developed is called the Chessie BIBI. It scores stream sites and is calibrated to the ecology of the Chesapeake Watershed. It can be accessed as an R package here- https://rstudio-pubs-static.s3.amazonaws.com/462489_8d2734925de3492eb294b64f54a260af.html.

Benthic Data Download File (cmcBenthicSamples-date)

This file contains the water quality data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file.

Date: The date the sample was taken, formatted month/day/year
Time: The time the sample was taken, formatted hrs:mm:ss
StationName: The station ID or name assigned by the individual monitoring group, usually used on the field datasheets to identify the station.
Latitude/Longitude: The latitude and longitude coordinate for the station, formatted in decimal degrees.
GroupName: The unique code that identifies the group that collected the data. This field can be used to find more information on each group in the groups metadata file.
Benthic Classification Name: This will have the scientific names which are each unique identifiers, of either phylum, order, class, genus, or species.
Benthic Classification Code: For scientific names, this is going to be the same as the column before. For the one-or two-letter codes, these will represent the classification that volunteers identify to VASOS.
Count: The data here is arranged vertically—each row is not arranged by sampling event but by count of benthic macroinvertebrate on a date. To get a stream health score, all data from a single sampling event must be
Station Conditions: After the benthic data if any observational or station condition data is recorded it will appear at the end of the spreadsheet and will only contain the fields that contain data for that particular dataset. See Appendix 1 Table 4, for a full list of benthic station conditions, many of these are habitat conditions of the stream site.
Comments: This field includes any additional comments made by the monitor, program coordinator or CMC team about that sampling event. Can include QA issues, equipment issues or sampling issues.

Benthic Groups Metadata File (cmcGroups)

This file contains the group metadata information for the benthic data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file. Detailed information is the same as the Water Quality metadata file.

Benthic Stations Metadata File (cmcStations)

This file contains the station metadata information for the benthic data selected through the query function on the homepage of the Chesapeake Data Explorer, exported as a .csv file. Detailed information is the same as the Water Quality metadata file.