# NEON USER GUIDE FOR SURFACE WATER MICROBE CELL COUNT (DP1.20138.001)

| PREPARED BY | ORGANIZATION |
| --- | --- |
| Stephanie Parker | AQU |
| Caren Scott | AQU |

# CHANGE RECORD

| REVISION | DATE | DESCRIPTION OF CHANGE |
| --- | --- | --- |
| A | 06/21/2017 | Initial Release |
| B | 05/25/2020 | Included general statement about usage of neonUtilities R package and statement about possible location changes. Included spatial sampling design changes for seepage lakes. |
| C | 02/08/2022 | Updated section 4.3 Data Revision with latest information regarding data release. |
| C.1 | 09/12/2024 | Added field blank samples to section 3.3 Sampling Design Changes and added new field blank QA table to section 3.9 Data Relationships. |
| D | 04/17/2025 | Add equation for surface water blanks. Added information about the new neonUtilities Python package. |

# TABLE OF CONTENTS

**1 DESCRIPTION** **1**
1.1 Purpose . . . 1
1.2 Scope . . . 1

**2 RELATED DOCUMENTS AND ACRONYMS** **2**
2.1 Associated Documents . . . 2
2.2 Acronyms . . . 2

**3 DATA PRODUCT DESCRIPTION** **3**
3.1 Spatial Sampling Design . . . 3
3.2 Temporal Sampling Design . . . 6
3.3 Sampling Design Changes . . . 6
3.4 Variables Reported . . . 7
3.5 Spatial Resolution and Extent . . . 7
3.6 Temporal Resolution and Extent . . . 7
3.7 Associated Data Streams . . . 7
3.8 Product Instances . . . 8
3.9 Data Relationships . . . 8
3.10 Special Considerations . . . 9

**4 DATA QUALITY** **9**
4.1 Data Entry Constraint and Validation . . . 9
4.2 Automated Data Processing Steps . . . 11
4.3 Data Revision . . . 11
4.4 Quality Flagging . . . 11
4.5 Analytical Facility Data Quality . . . 11

**5 REFERENCES** **12**

# LIST OF TABLES AND FIGURES

Table 1 Descriptions of the dataQF codes for quality flagging . . . 11

Figure 1 Generic aquatic site layouts with surface water cell count sampling locations highlighted in red boxes for wadeable streams, rivers, and both seepage and flow-through lakes. . . . 5
Figure 2 Lake and river sampling depths in non-stratified and stratified water columns. . . . 6
Figure 3 Schematic of the applications used by field technicians to enter water chemistry and surface water microbe cell count field data . . . 10

# 1 DESCRIPTION

## 1.1 Purpose

This document provides an overview of the data included in this NEON Level 1 (L1) data product, the quality controlled product generated from raw Level 0 (L0) data, and associated metadata. In the NEON data products framework, the raw data collected in the field, for example, the dry weights of litter functional groups from a single collection event are considered the lowest level (Level 0). Raw data that have been quality checked via the steps detailed herein, as well as simple metrics that emerge from the raw data are considered Level 1 data products.

The text herein provides a discussion of measurement theory and implementation, data product provenance, quality assurance and control methods used, and approximations and/or assumptions made during L1 data creation.

## 1.2 Scope

This document describes the steps needed to generate the L1 data product Surface water microbe cell count - the count of bacterial cells per liter of surface water in streams, rivers, and lakes - and associated field and external lab metadata. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the file, NEON Data Variables for Surface Water Microbe Cell Count (DP1.20138.001) (AD[05]), provided in the download package for this data product.

This document also describes the field data used in surface water microbial sequencing data products NEON Data Variables for Surface Water Microbe Marker Gene Sequences (DP1.20282.001) (AD[06]), NEON Data Variables for Surface Water Microbe Metagenome Sequences (DP1.20281.001) (AD[07]), NEON Data Variables for Surface Water Microbe Group Abundances (DP1.20278.001) (AD[08]), NEON Data Variables for Surface Water Microbe Community Composition (DP1.20141.001) (AD[09]). Details on the publication of these data products can be found in their respective user guides.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the data collected in the field pertaining to AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[10]). The raw data that are processed in this document are detailed in the file NEON Raw Data Validation for Surface Water Microbe Cell Count (DP0.20138.001) (AD[04]), provided in the download package for this data product. Please note that raw, L0 data products (denoted by 'DP0') may not always have the same numbers (e.g., '20138') as the corresponding L1 data product.

# 2 RELATED DOCUMENTS AND ACRONYMS

## 2.1 Associated Documents

| | | |
| --- | --- | --- |
| AD[01] | NEON.DOC.000001 | NEON Observatory Design (NOD) Requirements |
| AD[02] | NEON.DOC.001152 | NEON Aquatic Sampling Strategy |
| AD[03] | NEON.DOC.002652 | NEON Data Products Catalog |
| AD[04] | Available with data download | NEON Raw Data Validation for Surface Water Microbe Cell Count (DP0.20138.001) |
| AD[05] | Available with data download | NEON Data Variables for Surface Water Microbe Cell Count (DP1.20138.001) |
| AD[06] | Available with data download | NEON Data Variables for Surface Water Microbe Marker Gene Sequences (DP1.20282.001) |
| AD[07] | Available with data download | NEON Data Variables for Surface Water Microbe Metagenome Sequences (DP1.20281.001) |
| AD[08] | Available with data download | NEON Data Variables for Surface Water Microbe Group Abundances (DP1.20278.001) |
| AD[09] | Available with data download | NEON Data Variables for Surface Water Microbe Community Composition (DP1.20141.001) |
| AD[10] | NEON.DOC.003041 | AOS Protocol and Procedure: Aquatic Microbial Sampling |
| AD[11] | NEON.DOC.000008 | NEON Acronym List |
| AD[12] | NEON.DOC.000243 | NEON Glossary of Terms |
| AD[13] | NEON.DOC.002905 | AOS Protocol and Procedure: Water Chemistry Sampling in Surface Waters and Groundwater |
| AD[14] | NEON.DOC.004825 | NEON Algorithm Theoretical Basis Document: OS Generic Transitions |
| AD[15] | Available on NEON data portal | NEON Ingest Conversion Language Function Library |
| AD[16] | Available on NEON data portal | NEON Ingest Conversion Language |
| AD[17] | Available with data download | Categorical Codes csv |

## 2.2 Acronyms

| Acronym | Definition |
| --- | --- |
| PI | Propidium iodide |

# 3 DATA PRODUCT DESCRIPTION

This data product contains the quality-controlled, field sampling metadata and associated total bacterial cell count data provided by a contracted lab. Field samples are collected in the water column of wadeable streams, rivers, and lakes in conjunction with surface water chemistry samples. Bulk water samples are collected and subsampled for microbial cell counts, genetic analyses, and archive.

Surface water microbes are collected 12 times per year in wadeable streams and 6 times per year in lakes and non-wadeable streams, at the same time and location as standard recurrent (monthly) water chemistry samples (AD[13]). Details on sampling locations and timing are provided in NEON Raw Data Validation for Surface Water Microbe Cell Count (DP0.20138.001) (AD[10]) and the Surface Water Chemistry Sampling in Aquatic Habitats protocol (AD[13]). Samples are collected as grab samples from the water column at sampling locations near the S2 sensor set in streams, near the buoy in rivers, near buoy, inlet, and outlet sensor sets in flow-through lakes, and near the buoy sensor sets in seepage lakes. In lakes and rivers with a stratified water column, samples are collected at multiple depths.

Water samples are processed in the field or in the lab within 4 hours of collection if field conditions are not conductive to subsampling (e.g., freezing conditions). An 18 mL unfiltered aliquot of field sample is preserved with 2 mL of 10% formaldehyde (final concentration of ~1% formaldehyde). Samples for genetic analysis and archive are filtered on a 0.2 um Sterivex filter and flash-frozen.

Cell count samples are preserved with formaldehyde in the field, kept in the dark at 4$^{o}$C. Samples are sent to an external analytical lab within 60 days of collection and analyzed using a propidium iodide (PI) staining method and epifluorescence microscopy (Boulos et al. 1999). Cell counts are enumerated using an image analysis program calibrated for NEON samples. Additional quality assurance data related to the automated counts, including counts of a standard reference photo, are available in the expanded package.

Genetic filter samples are stored at -80$^{o}$C until analysis at an external facility. Archive filter samples are stored at -80$^{o}$C until use. Details of lab analyses are included in the user guides for NEON Data Variables for Surface Water Microbe Marker Gene Sequences (DP1.20282.001) (AD[06]), NEON Data Variables for Surface Water Microbe Metagenome Sequences (DP1.20281.001) (AD[07]), NEON Data Variables for Surface Water Microbe Group Abundances (DP1.20278.001) (AD[08]), NEON Data Variables for Surface Water Microbe Community Composition (DP1.20141.001) (AD[09]).

## 3.1 Spatial Sampling Design

Aquatic surface water microbial samples are collected at all NEON aquatic sites at the same time and location as surface water chemistry samples (AD[13]). At stream sites, 1 sample is collected from the thalweg <1 m downstream of the downstream sensor set (S2) on each sampling date. Samples represent the water column, so care is taken to avoid stirring up sediments that may contaminate the sample.

At river (non-wadeable stream) sites, surface water microbe samples are collected just downstream of the sensor set or profiling buoy (station = ‘c0’) (Figure 1). If the river is non-stratified, samples are collected at 0.5 m depth. If the river is stratified, an epilimnion sample is collected at 0.5 m (station = ‘c1’) and an integrated sample is collected from the hypolimion (station = ‘c2’). Care is taken to avoid contamination from sediments suspended by the boat motor or anchor.

At flow-through lake sites, samples are collected near the profiling buoy, the inlet sensor, and outlet sensor (Figure 1). In seepage lakes with no defined inlet and outlet, samples are only collected near the profiling buoy. Near the buoy, sampling depth is dependent on the presence or absence of lake stratification (Figure 2). In an unstratified lake, the sample is collected near the surface at 0.5 m depth. In a stratified lake, additional samples are collected from the hypolimnion, in addition to the surface water sample. In lakes with a shallow hypolimnion (<4 m), the sample is collected from the midpoint of the hypolimnion. In lakes with a deeper hypolimnion (>4 m), an integrated sample is collected throughout the hypolimnion. Samples collected near the inlet and outlet sensor sets are collected near the surface at 0.5 m depth. See AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[10]) and AOS Protocol and Procedure: Water Chemistry Sampling in Surface Waters and Groundwater (AD[13]) for additional details.

As much as possible, sampling occurs in the same locations over the lifetime of the Observatory. However, over time some sampling locations may become impossible to sample, due to disturbance or other local changes. When this occurs, the location and its location ID are retired. A location may also shift to slightly different coordinates. Refer to the locations endpoint of the NEON API for details about locations that have been moved or retired: https://data.neonscience.org/data-api/endpoints/locations/

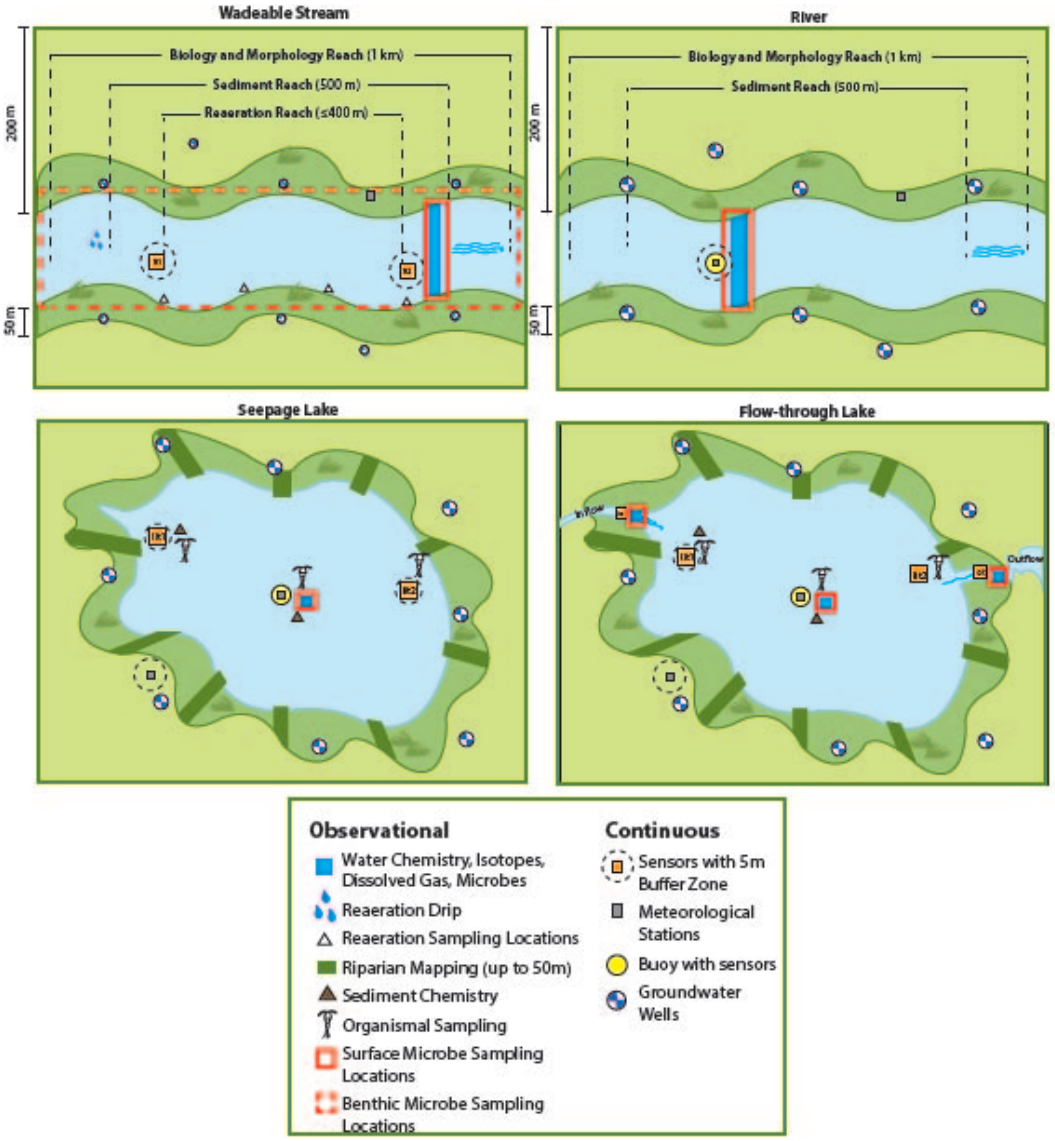


Figure 1: Generic aquatic site layouts with surface water cell count sampling locations highlighted in red boxes for wadeable streams, rivers, and both seepage and flow-through lakes.

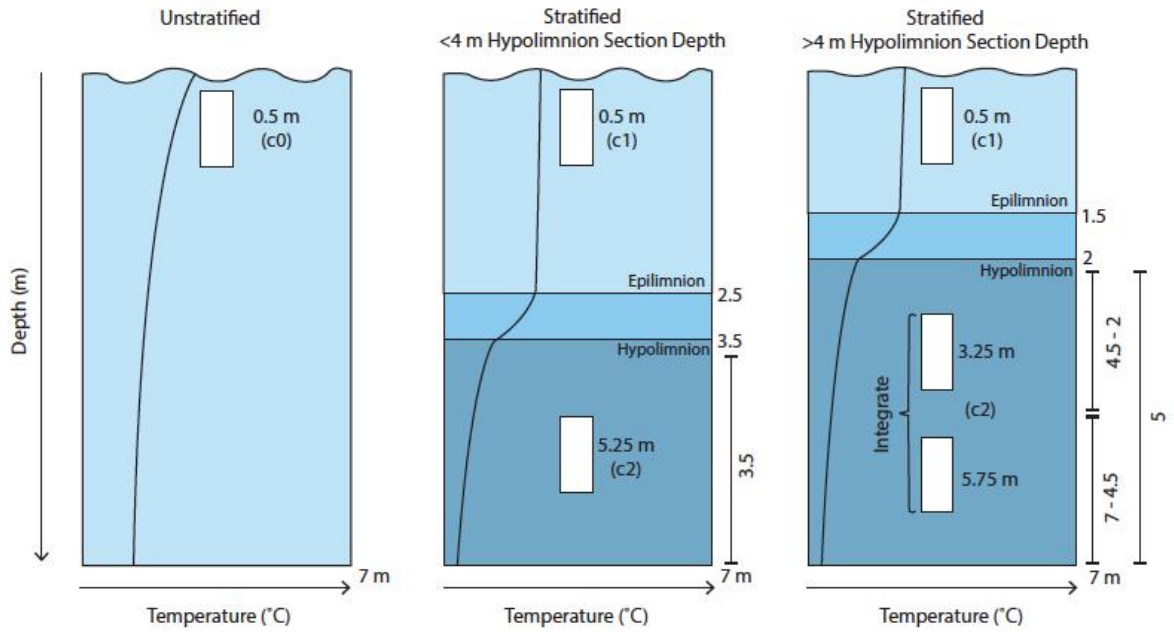


Figure 2: Lake and river sampling depths in non-stratified and stratified water columns.

## 3.2 Temporal Sampling Design

Surface water microbe samples are collected at the same time and location as standard recurrent water chemistry sampling: once per month in wadeable streams (12 times per year). At streams sites, samples are collected year-round including when the stream is frozen over if the ice can be broken by hand. When the ice becomes too thick, sampling is suspended and noted as **samplingImpractical** in the field data. At lake and river sites, microbe samples are collected every other month with standard recurrent water chemistry samples. At northern sites, samples are collected year round and and collected under the ice during winter. See NEON Aquatic Sampling Strategy (AD[02]), AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[10]) and AOS Protocol and Procedure: Water Chemistry Sampling in Surface Waters and Groundwater (AD[13]) for additional details.

## 3.3 Sampling Design Changes

2014-2017: During the first four sampling years, samples were collected at three locations in all lakes (seepage or flow-through). Starting in 2018, samples from seepage lakes are only collected near the buoy sensor sets.

2024: Field blank samples are collected starting in 2024. Field blank samples are collected once per year in mid-summer and analyzed by the external lab.

### 3.4 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Surface Water Microbe Cell Count (DP0.20138.001) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the file, NEON Data Variables for Surface Water Microbe Cell Count (DP1.20138.001) (AD[05]).

Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 16 February 2014), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 16 February 2014), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), where applicable. NEON OS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and Geoid12A geoid model for its vertical reference surface. Latitudes and longitudes are denoted in decimal notation to six decimal places, with longitudes indicated as negative west of the Greenwich meridian.

Some variables described in this document may be for NEON internal use only and will not appear in downloaded data.

### 3.5 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is at a single station within a site. For example, data may be collected at a specific depth in the water column of a lake.The basic spatial data included in the data downloaded include the latitude, longitude, and elevation of the named location at the aquatic site (e.g., the aquatic location) or the latitude and longitude of an alternate location if the named location is not suitable for sampling.

**namedLocation** (unique ID given to the location within the site) → **siteID** (ID of NEON site) → **domainID** (ID of a NEON domain)

### 3.6 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is at **collectDate**, the date and time of day when the sample was collected in the field.

The NEON Data Portal provides data in monthly files for query and download efficiency. Queries including any part of a month will return data from the entire month. Code to stack files across months is available here: https://github.com/NEONScience/NEON-utilities.

### 3.7 Associated Data Streams

Surface water microbe samples are related to water chemistry samples collected at the same time and location. Water chemistry data are available in the 'Chemical properties of surface water' data product (DP1.20093.001).

Cell count samples are also related to aquatic microbe sequencing data generated from subsamples of the same parent sample (linked with the field **parentSampleID**), including Surface water microbe community composition (DP1.20141.001), Surface water microbe group abundances (DP1.20278.001), Surface

water microbe marker gene sequences (DP1.20282.001), and Surface water microbe metagenome sequences (DP1.20281.001).

### 3.8 Product Instances

At each stream site, there will be 12 samples collected per year. At a lake or river site, there will be a minimum of 6 samples and a maximum of 9 samples collected per year (maximum if water the column is stratified). Each sample generates one cell count record at the external lab.

### 3.9 Data Relationships

A record in amc_fieldCellCounts must have a corresponding record in amc_fieldSuperParent describing measurement depth and abiotic variables during sample collection. Each record in amc_fieldCellCounts may be linked to a record in amc_cellCounts, which contains data from the external laboratory. Duplicates and/or missing data may exist where protocol and/or data entry aberrations have occurred; users should check data carefully for anomalies before joining tables.

amc_fieldSuperParent.csv - > One record is created for each cell count that comes from a surface water sample, and contains metadata that applies to cell counts.

amc_fieldCellCounts.csv - > One record is created by field personnel for each cell count sample. Each field record has a corresponding amc_fieldSuperParent **parentSampleID**. The field **cellCountSampleID** is also created here and will be used to track the sample through to the external lab.

amc_cellCounts.csv - > One record is created by the external lab for each cell count sample, linked to amc_fieldCellCounts by the field **cellCountSampleID**. If samples need to be re-analyzed for QA reasons, there may be more than 1 record per **cellCountSampleID**.

amc_cellCountsBlankQA.csv - > Records in this table start in 2024. One record is created by the external lab for each cell count sample, linked to amc_fieldCellCounts by the field **fieldBlankCellCountSampleID**. If samples need to be re-analyzed for QA reasons, there may be more than 1 record per **cellCountSampleID**.

amc_cellCountLabSummary.csv - > QA data are recorded using the fields **labSpecificStartDate** and **labSpecificEndDate**. The QA data pertain to the lab analysis if the field **testedDate** in amc_cellCounts falls between this date range.

Data downloaded from the NEON Data Portal are provided in separate data files for each site and month requested. The neonUtilities package in R and the neonutilities package in Python contain functions to merge these files across sites and months into a single file for each table. The neonUtilities R package is available from the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/web/packages/neonUtilities/index.html) and can be installed using the install.packages() function in R. The neonutilities package in Python is available on the Python Package Index (PyPi; https://pypi.org/project/neonutilities/) and can be installed using pip. For instructions on using the package in either language to merge NEON data files, see the Download and Explore NEON Data tutorial on the NEON website: https://www.neonscience.org/download-explore-neon-data.

### 3.10 Special Considerations

The cell count data results comes from an external lab, in the field **rawMicrobialAbundance**. **rawMicrobialAbundance** is NOT corrected for preservative volume, so data users will need to apply this correction using data from **cellCountPreservantVolume** from the amc_fieldCellCounts table for an accurate cell count value.

**Cell count samples**

$$microbialAbundancePerMl_i = amc\_cellCounts.rawMicrobialAbundance_i \times (amc\_fieldCellCounts.cellCountSampleVolume_i + \frac{amc\_fieldCellCounts.cellCountPreservantVolume_i)}{amc\_fieldCellCounts.cellCountSampleVolume_i}$$

Where 'i' is a unique **cellCountSampleID** See the external lab SOP (referenced in amc_cellCounts.csv) for calculations applied to the data by the external laboratory.

**Cell count blank samples**

$$blankMicrobialAbundancePerMl_i = amc\_cellCountsBlankQA.rawMicrobialAbundance_i \times (amc\_fieldCellCounts.fieldBlankCellCountSampleVolume_i + \frac{amc\_fieldCellCounts.fieldBlankCellCountPreservantVolume_i)}{amc\_fieldCellCounts.fieldBlankCellCountSampleVolume_i}$$

Where 'i' is a unique **fieldBlankCellCountSampleID**

## 4 DATA QUALITY

### 4.1 Data Entry Constraint and Validation

Many quality control measures are implemented at the point of data entry within a mobile data entry application or web user interface (UI). For example, data formats are constrained and data values controlled through the provision of dropdown options, which reduces the number of processing steps necessary to prepare the raw data for publication. The data entry workflow for collecting surface water microbe cell count data as part of the water sampling is diagrammed in Figure 3.

An additional set of constraints are implemented during the process of ingest into the NEON database. The product-specific data constraint and validation requirements built into data entry applications and database ingest are described in the document NEON Raw Data Validation for Surface Water Microbe Cell Count (DP0.20138.001), provided with every download of this data product. Contained within this file is a field named 'entryValidationRulesForm', which describes syntactically the validation rules for each field built into the data entry application. Data entry constraints are described in Nicl syntax in the validation file provided with every data download, and the Nicl language is described in NEON's Ingest Conversion Language (NICL) specifications ([AD[15]).

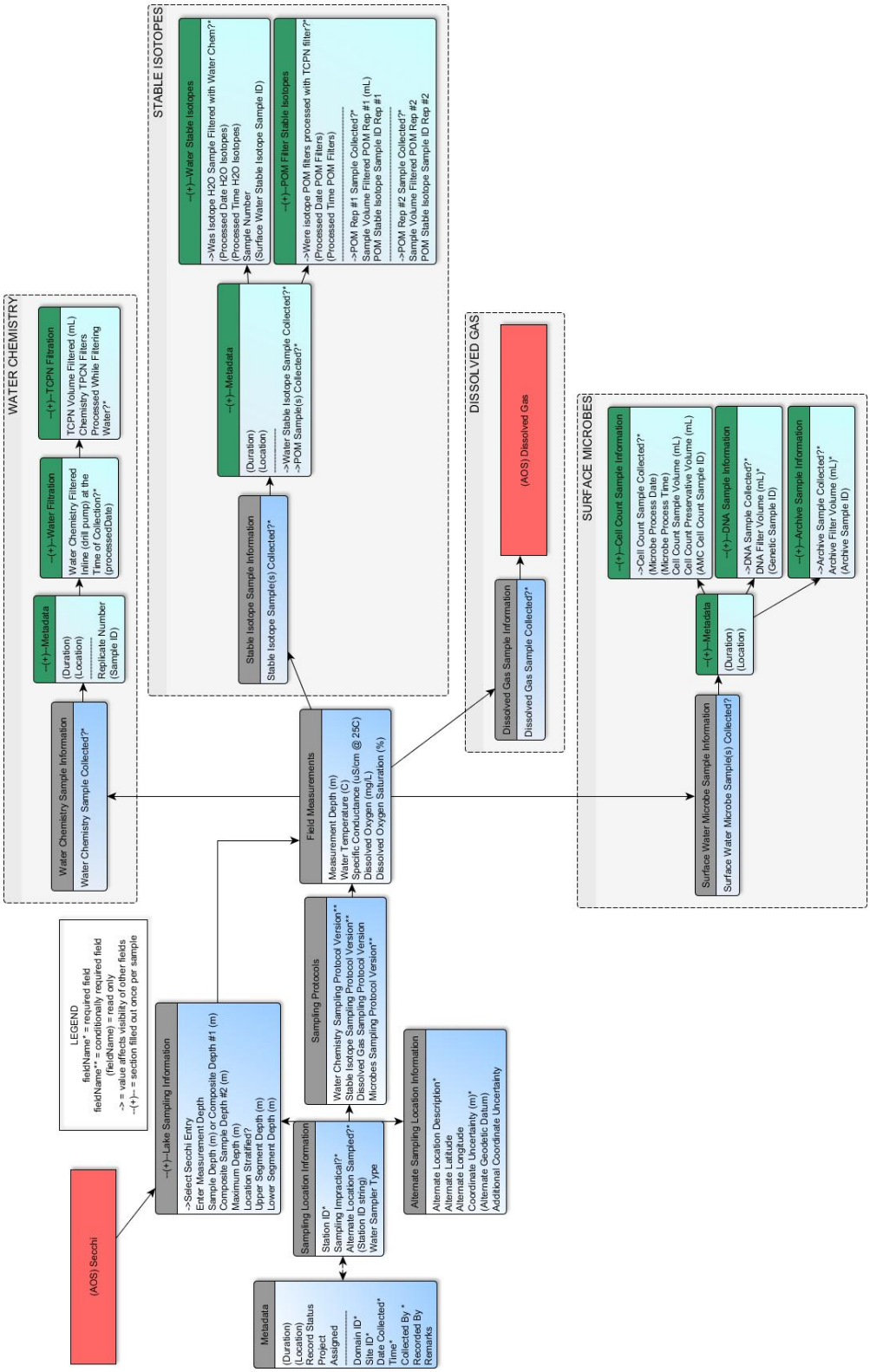


Figure 3: Schematic of the applications used by field technicians to enter water chemistry and surface water microbe cell count field data

## 4.2 Automated Data Processing Steps

Following data entry into a mobile application of web user interface, the steps used to process the data through to publication on the NEON Data Portal are detailed in the NEON Algorithm Theoretical Basis Document: OS Generic Transitions (AD[14]).

## 4.3 Data Revision

All data are provisional until a numbered version is released. Annually, NEON releases a static version of all or almost all data products, annotated with digital object identifiers (DOIs). The first data Release was made in 2021. During the provisional period, QA/QC is an active process, as opposed to a discrete activity performed once, and records are updated on a rolling basis as a result of scheduled tests or feedback from data users. The Issue Log section of the data product landing page contains a history of major known errors and revisions.

## 4.4 Quality Flagging

The **dataQF** field in each data record is a quality flag for known errors applying to the record. Please see the table below for an explanation of **dataQF** codes specific to this product.

Table 1: Descriptions of the dataQF codes for quality flagging

| fieldName | value | definition |
| --- | --- | --- |
| dataQF | legacyData | Data recorded using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow |

Records of land management activities, disturbances, and other incidents of ecological note that may have a potential impact are found in the Site Management and Event Reporting data product (DP1.10111.001)

## 4.5 Analytical Facility Data Quality

Data analyses conducted on microbial cell count data conform to the current data quality standards used by practitioners. Ten percent of all samples are quality checked for taxonomic difference between two taxonomists at the external facility. These records are indicated by the fields **qaqcStatus** and **enumerationDifference** indicating whether the sample has undergone internal lab quality checks. Samples are checked against a standard QC image and will be analysed for percent difference in enumeration (**enumerationDifference**, PDE) against the analyzed samples. The standard image is recounted by the technician running the samples monthly. Details on the lab QA/QC process can be found in the external lab SOP.

# 5 REFERENCES

Boulos, L., M. Prevost, B. Barbeau, J. Coallier, and R. Desjardins. 1999. LIVE/DEAD®*Bac*Light™: application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water. Journal of Microbiological Methods 37: 77-86.