NEON USER GUIDE TO SOIL PHYSICAL AND CHEMICAL PROPERTIES, PERIODIC (DP1.10086.001)

PREPARED BY	ORGANIZATION
Lee Stanish	TOS
Samantha Weintraub-Leff	TOS

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	04/20/2017	Initial Release
B	09/01/2020	Included statement about use of neonUtilities R package and possible sampling location changes. Added text on publication of soil chemistry and stable isotope data along with soil field collections and pH/moisture measurements as part of this data product. Sections 3 and 5: Updated reference for pH method; Section 3.2: Revised temporal design description to match current design; Section 3.3: Added Sampling Design Changes section and included changes to sampling frequency for microbial analyses; Section 3.5: Updated instructions on obtaining sample location data using the NEON API; Section 3.6: Introduced eventID; Section 3.8: Added information for missed sampling; Section 4.4: Added new dataQF option for early Alaska samples.
C	03/02/2022	Section 3.3: Added details for sample identifier changes that occurred in 2021 and 2022; Section 4.3: Updated Data Revision with information regarding data release; Minor text and correction updates
D	12/07/2022	New Section 3.3.1: Description of pH method change and rationale; Section 3.9: New field to capture frozen soil archive container info and how to interpret missing values; Section 3.10: Clarified when litterDepth or soilTemp might still be recorded in wetland plots with standing water; Section 4.4: New dataQF value for 'missing-pH-Data', new fields geneticSamplePrepMethod and geneticArchiveSamplePrepMethod and how to interpret missing values; minor text clarifications throughout.
D.1	04/18/2023	Section 3.3.1: Updated the description of 2020 pH method change including possible site-specific impacts on pH values.
E	03/01/2024	Figure 1: updated to use more detailed gridded plot map. Section 3.5: Explanation of change in subplot naming convention. Section 3.10: Added note for resampled locations. Section 4.4: Added two new dataQF options.

REVISION	DATE	DESCRIPTION OF CHANGE
F	04/27/2025	Section 3.5: Updated link for NEON TOS shapefiles and added instructions for using the getLocByName function in the geoNEON R package to derive precise sampling locations. Section 3.10 and Section 4.4: Added details regarding how 'location-resampled' dataQF flags are determined, and added 'pit-location-resampled' dataQF flag for samples collected at or near the location of a soil characterization pit. Added information about the new neonUtilities Python package.
G	03/10/2026	Section 3, 3.1, 3.2: Expanded details of sampling design description including spatial and temporal components and included link to NEON Biorepository to request archive samples. Figures 1, 2, and 4 updated for accuracy. Sections 3.1 and 3.5: Revised documentation for tracking sampling location changes. Section 3.3: Added Design Change text for 2026 sampling suspension. Section 4.2: Included the OS Data Quality Control Algorithm Theoretical Basis Document. Section 4.4: Revised the list of dataQF values. Related microbial data product updates: included a DPID at each mention, clarified that group abundances (DP1.10109.001) was discontinued, replaced reference to community composition (DP1.10081.001) with community taxonomy (DP1.10081.002) since that is the recommended product for derived soil microbial community analyses.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

This document provides an overview of the data included in this NEON Level 1 data product, the quality controlled product generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field - for example, soil core depth from a single sample - 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 Soil physical and chemical properties, periodic (DP1.10086.001). The data product encompasses field metadata for soil samples, generation of subsamples used for microbial analyses, measurements of soil temperature, moisture, and pH, and measurements of soil carbon (C) and nitrogen (N). As of August 2020, the Soil chemical properties (DP1.10078.001), Soil stable isotopes (DP1.10101.001), and Soil inorganic nitrogen pools and transformations (DP1.10080.001) data products are published as part of Soil physical and chemical properties, periodic. However, this User Guide is focused on the subset of tables dealing with field collection, subsampling, and moisture/pH measurements, while two related Data Product User Guides detail the biogeochemical measurements included in this data product. Data from the microbial subsamples can be found in related data products that will be described later.

This document provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the files NEON Data Variables for Soil physical and chemical properties, periodic (DP1.10086.001) (AD[04]) and and NEON Categorical Codes for Soil physical and chemical properties, periodic (AD[06]), provided in the download package for this data product.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the data collected in the field pertaining to TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[09]), or TOS Standard Operating Procedure: Wetland Soil Sampling (AD[10]) if the site is a wetland. The raw data that are processed in this document are detailed in the file, NEON Raw Data Validation for Soil physical properties, distributed periodic (DP0.10086.001) (AD[05]), provided in the download package for this data product. Please note that raw data products (denoted by 'DP0') may not always have the same numbers (e.g., '10033') 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.000913	TOS Science Design for Spatial Sampling
AD[03]	NEON.DOC.002652	NEON Data Products Catalog
AD[04]	Available with data download	Variables csv
AD[05]	Available with data download	Validation csv
AD[06]	Available with data download	Categorical Codes csv
AD[07]	NEON.DOC.000906	TOS Science Design for Terrestrial Biogeochemistry
AD[08]	NEON.DOC.000908	TOS Science Design for Terrestrial Microbial Diversity
AD[09]	NEON.DOC.014048	TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling
AD[10]	NEON.DOC.004130	TOS Standard Operating Procedure: Wetland Soil Sampling
AD[11]	NEON.DOC.000008	NEON Acronym List
AD[12]	NEON.DOC.000243	NEON Glossary of Terms
AD[13]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[14]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[15]	Available on NEON data portal	NEON Ingest Conversion Language
AD[16]	NEON.DOC.005424	Algorithm Theoretical Basis Document: OS Data Quality Control

3 DATA PRODUCT DESCRIPTION

The Soil physical and chemical properties, periodic (DP1.10086.001) data product is derived from TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[09]), or TOS Standard Operating Procedure: Wetland Soil Sampling (AD[10]) if the site is a wetland. The sampling plan implements the guidelines and requirements described in the TOS Science Design for Terrestrial Biogeochemistry (AD[07]) and TOS Science Design for Terrestrial Microbial Diversity (AD[08]). Most field and laboratory data are reported at the spatial resolution of a single soil sampling location, e.g., an X,Y coordinate (+/- 0.5 meters) within a subplot within a NEON plot, with a temporal resolution of a single collection date. The exception is microbial metagenomic samples, which represent a pooled, plot-level composite and are thus reported at the scale of a NEON plot and may have temporal resolution > 1 day if pooled subsamples were collected on different days.

Soils are sampled by horizon type (organic or mineral, Figure 1) to a maximum depth of 30 cm. This depth was chosen to capture dynamics in the upper layers of the subsoil or B horizon along with the topsoil, and to match the sampling depth used in the Root biomass and chemistry, periodic (DP1.10067.001) data product. In some cases the boundary between organic and mineral horizons is difficult to distinguish. NEON technicians do their best but some degree of horizon mis-classification is unavoidable. For certain bouts, users can validate technician designation of horizon type using measured carbon content in the sls_soilChemistry table. The U.S. Department of Agriculture (USDA) generally considers a horizon 'organic' when organic C content is greater than or equal to 20%.

In most cases NEON collects mineral soil cores using 2-2.5" diameter augers, however certain sites use a different approach due to local soil properties. The specific device used to collect soils at each site is listed in the appendix of TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[09]). At each X, Y sampling location, *in-situ* soil temperature at 10 cm depth is measured with a handheld probe and litter depth is recorded, then surface litter is brushed away and soil samples are removed and homogenized. Extra field homogenization occurs for locations with mature biological soil crust prior to subsampling. Subsamples for microbial community analyses and archive are created and immediately frozen on dry ice (except in rare circumstances, see Section 4.4 for more detail). The '-omics' datasets generated from these field-prepared subsamples can be found in the data products Soil microbe marker gene sequences (DP1.10108.001), Soil microbe metagenome sequences (DP1.10107.001), and Soil microbe community taxonomy (DP1.10081.002).

After field work is complete, samples are transferred to NEON domain support facilities where measurements of gravimetric soil moisture and pH are conducted, following the methods outlined in Robertson et al. (1999) and Burt (2014), respectively. For select bouts, subsamples are also prepared for analysis of inorganic N pools and transformation rates, total organic C and total N concentrations and stable isotope ratios, and microbial biomass via phospholipid fatty acid (PLFA) analysis. PLFA data are delivered as part of the Soil microbe biomass (DP1.10104.001) data product, while soil chemistry and inorganic N data are part of this data product and described in detail in related Data Product User Guides.

Soil samples from field collections in frozen and air-dried condition are stored at the NEON Biorepository and are available by request for further study and analysis. Contact the Biorepository for detailed information about sample availability.

Data on soil physical and chemical properties help to elucidate the controls on microbial activity, nutrient

cycling, and carbon storage in soils at the plot, site, and continental scales. They also provide essential data for understanding change in soil microbial and biogeochemical dynamics over time.

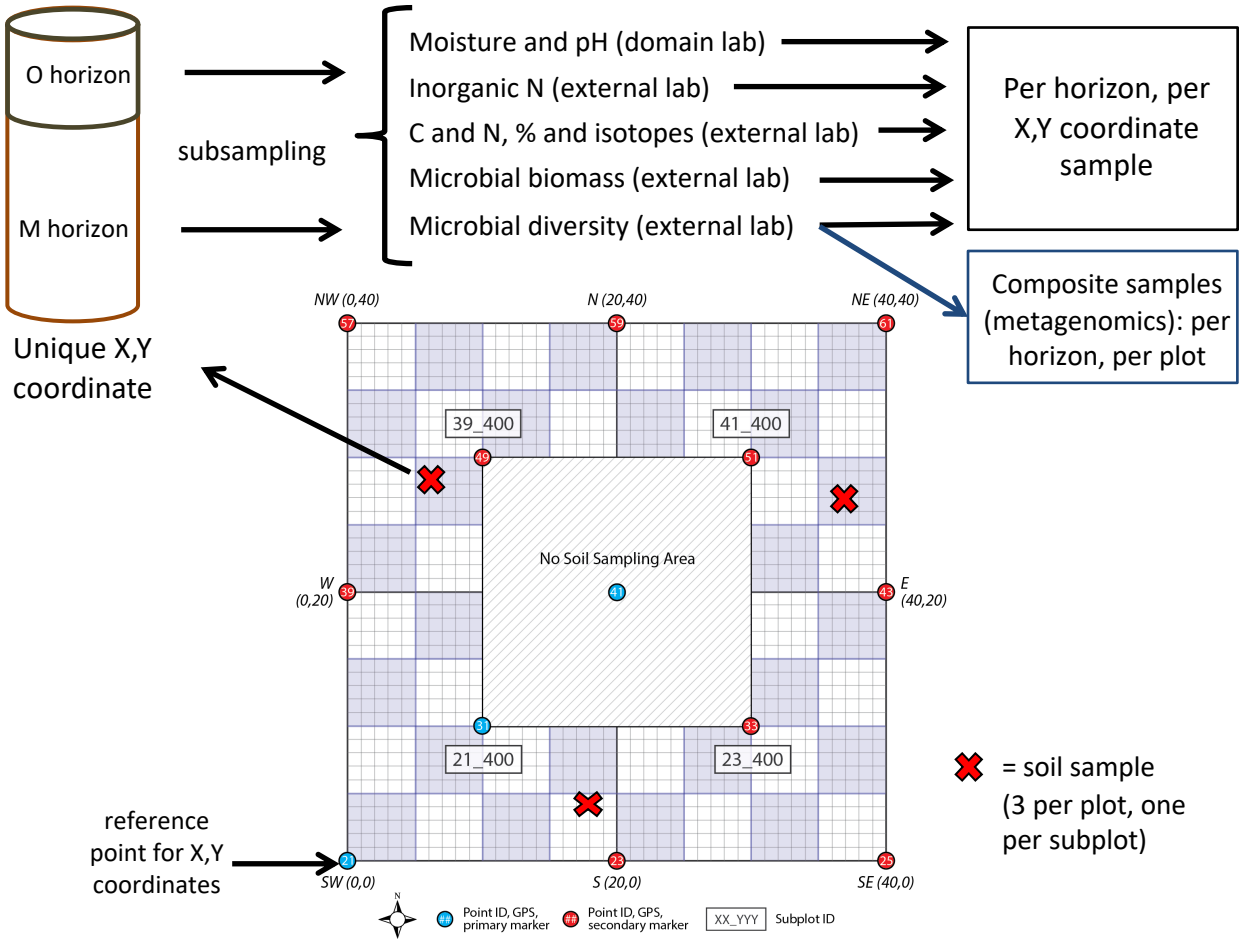

Figure 1: Overview of soil field sampling and analysis workflow.

3.1 Spatial Sampling Design

Soil biogeochemical and microbial sampling is conducted at all terrestrial NEON sites. Soils are sampled from three pre-determined, randomly assigned X,Y locations in 3 of 4 randomly selected subplots per 40 x 40 meter plot (Figure 1). If large rocks, tree roots, animal burrows, or other significant impediments or disturbances are encountered at a pre-determined sampling location, technicians proceed down a list of random locations until an acceptable one is found. Ten plots per site are sampled, four within the tower airshed and six others distributed across the landscape and located in dominant vegetation types (see Figure 2 for NEON site design). Soil Tower plots are selected using a random spatial design, while soil Distributed plots follow a random stratified design based on National Land Cover Database (NLCD) vegetation class. See AD[02] for further details on the NEON spatial design.

Sampling typically occurs in the same locations over the lifetime of the Observatory. However, sampling locations may become impractical to sample due to disturbance or other local changes. When this occurs, the location and its location ID are retired or shifted to slightly different coordinates. Refer to the TOS plot location changes spreadsheet found in the "Terrestrial Observation System Sampling Locations" download on the spatial-data-maps page at neonscience.org for details about locations that have been retired or added since the operations phase started in 2019. The same download also includes the 'versionedPoints" and "versionedSubplots" files, which document shifts in coordinates.

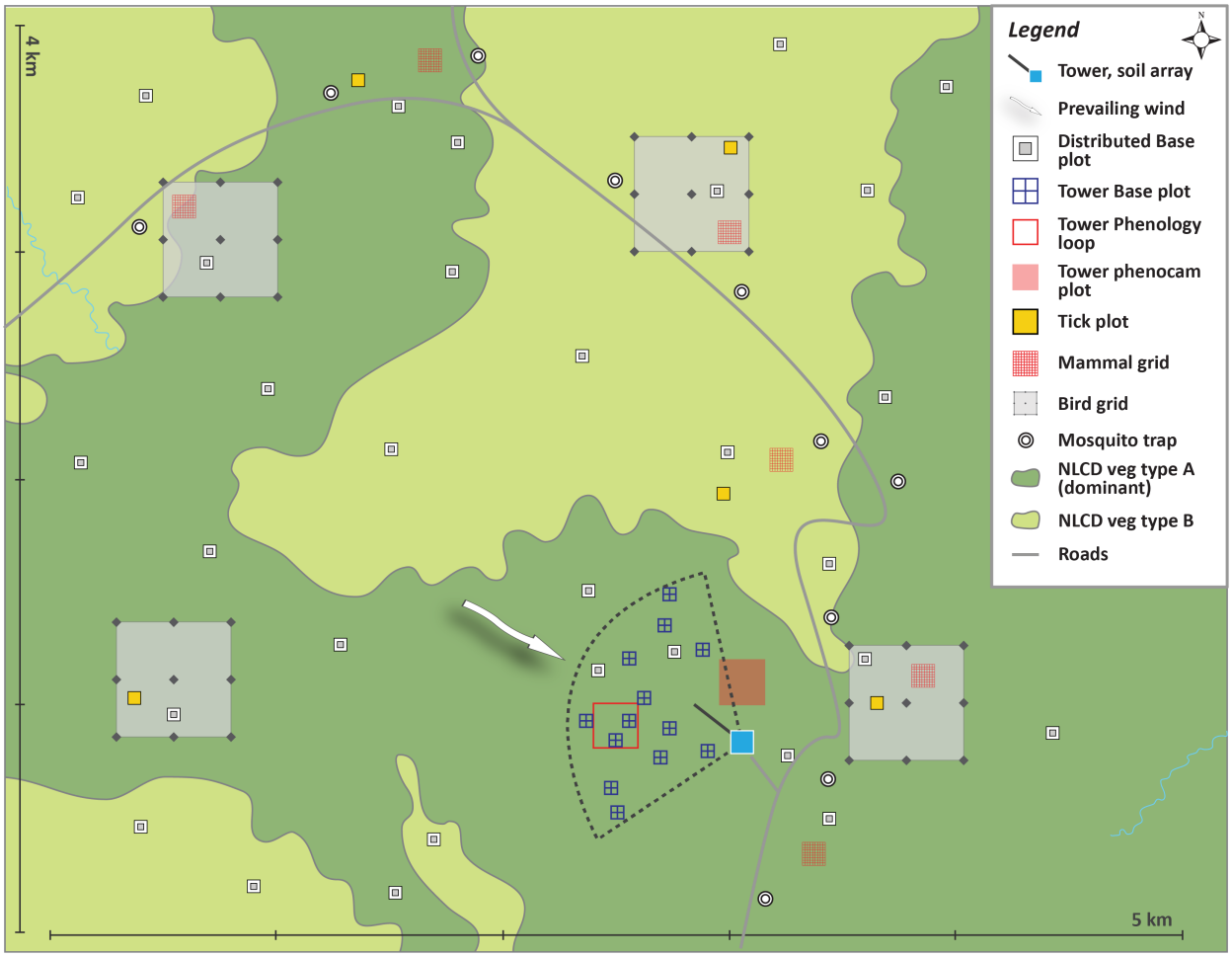

Figure 2: Representation of a NEON site with Tower and Distributed Base plots shown spanning different land cover types. Soil sampling occurs in 4 Tower and 6 Distributed Base plots.

3.2 Temporal Sampling Design

Soil sampling for a suite of physical, chemical and microbial analyses occurs up to three times per year on either an annual or inter-annual time frame, the frequency of collection varying by analysis type (Figure 3). Certain procedures are conducted at all sites during every sampling bout, namely basic field measurements, soil moisture, pH, and subsampling for microbial genetic archive. At all sites, sampling occurs

during the period of peak greenness/productivity as indicated either by remotely sensed vegetation indices (e.g. NDVI) or historic changes in precipitation. At most sites, soil sampling occurs two additional times per year, usually bracketing the peak greenness window and aimed at capturing seasonal transitions in microbial activity. Sampling frequency is reduced in arctic and boreal sites, which are only sampled during peak greenness.

	Off-Year			Coordinated bout			
N-trans Bout Type	No			T initial			T final
Sample Timing	T1	PG	T2	T1	PG	T2	T1, PG, T2
Bout Type	microbes/ microbes-Biomass	microbes/ microbes-Biomass	microbes/ microbes-Biomass	microbes-Biomass	microbes-BiomassBGC	microbes-Biomass	fieldOnly
Field-generated Samples	Bulk -gen* -gaX	Bulk -gaX -gen* -comp*	Bulk -gen* -gaX	Bulk -gen -gaX	Bulk -gen -gaX -comp	Bulk -gen -gaX	Bulk
Lab-generated Samples	-bm*	-bm*	-bm*	-bm -kcl	-bm; -kcl -cn; -ba	-bm -kcl	kcl
Lab measurements	pH moisture	pH moisture	pH moisture	pH moisture	pH moisture	pH moisture	moisture
*** Core sites only**							

Abbreviations

Sample Timing:

T1: Transition 1

PG: Peak Greenness

T2: Transition 2

Sample:

Bulk: Homogenized soil used for all subsamples and analyses

Subsamples:

-gen: microbial genetic analysis subsample

-gaX: microbial genetic archive subsample, X denotes subsample number of 1-5 (for up to 5 vials)

-comp: plot-level composited microbial metagenomics subsample

-bm: microbial biomass subsample

-kcl: KCl extraction sample for nitrogen transformation rate measurement

-cn: BGC analysis subsample

-ba: BGC archive subsample

Figure 3: Overview of soil field sampling and laboratory analyses based on bout type. At the highest level, within a year sampling bouts at a site are either Off-year or Coordinated. Off-year bouts (majority) perform routine measurements and collect microbial samples either for downstream analyses and archive or archive only depending on site type. Coordinated bouts (every 5 years) perform routine measurements, microbial collections regardless of site type, and a full suite of biogeochemical analyses.

Once every five years each NEON site will conduct Coordinated sampling in which soils collected during routine sampling bouts are utilized for additional downstream analyses. These analyses include incubation and processing of soil cores to determine inorganic N pools and net transformation rates, plus total soil organic C and total N measurements and creation of an air-dried subsample for the Biorepository (during the peak greenness bout only, Figure 4). Sites rotate on a pre-defined schedule such that every site will complete Coordinated sampling over a 5-year period, then this cycle repeats for the lifetime of the Observatory.

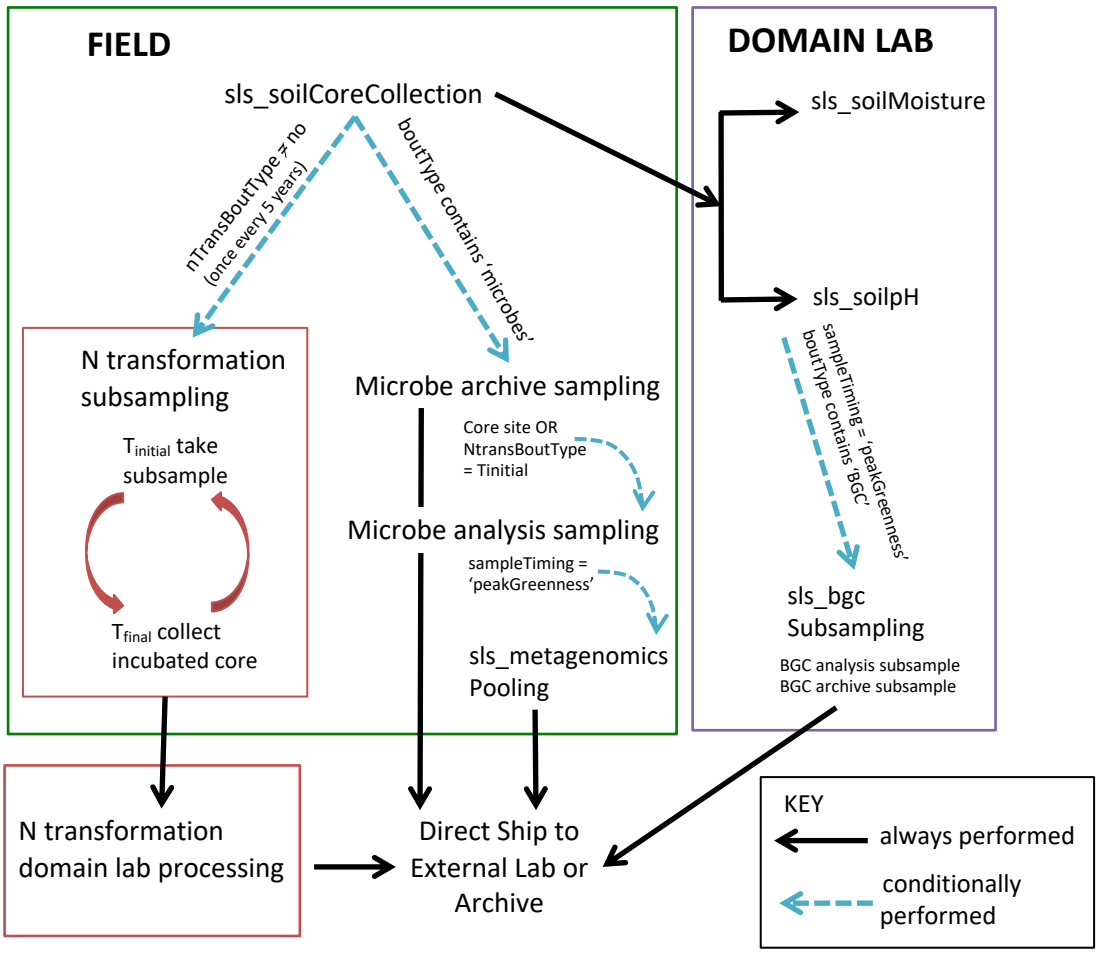

Figure 4: Soil field and laboratory workflows. Arrows indicate which data tables are generated by each step in the soil sampling process and under what conditions. Data table names follow the convention: module_tableName, for example sls_soilCoreCollection

A plot-level, pooled metagenomics sample is collected during peak greenness. At core terrestrial sites, this sample is generated every year, while gradient sites generate a metagenomics sample once every five years during Coordinated bouts.

Subsamples for additional microbial analyses are collected according to Figure 3. Subsamples are col-

lected for microbial marker gene sequencing (16S and ITS sequencing) and microbial biomass analyses during all bouts (all seasons) at core terrestrial sites, while gradient sites generate subsamples for these analyses once every five years, all bouts and seasons, during Coordinated sampling.

For each of the five data tables discussed in this User Guide,

- field collection, pH, and moisture tables will contain data for 1-3 sampling events per site per year
- metagenomics pooling table will contain data for 1 sampling event per core site per year as well as for gradient sites conducting Coordinated sampling
- biogeochemistry subsampling table will contain data for 1 sampling event per site, every 5 years

3.3 Sampling Design Changes

Over the course of early operations, the design for soil periodic sampling evolved. Below is a list of previous sampling strategies that differ from the current design, with applicable years indicated.

- 2013 - 2018: All terrestrial sites generated a metagenomics sample annually.
- 2013 - 2018: Subsamples were collected for microbial marker gene sequencing analyses (16S and ITS sequencing) during every bout and for all sites.
- 2013 - 2018: Subsamples were collected for microbial group abundance analysis during every bout and for all sites. Data are available from this period (see DP1.10109.001 Soil microbe group abundances) but new data generation was discontinued.
- 2015 - 2017: The field methods for distinguishing litter from organic soil were deemed not optimal for arctic tundra and boreal sites, and the method was refined in 2018. See Special Considerations for details.
- 2017 - 2019: Subsamples for microbial biomass analysis were collected once every five years only, during Coordinated sampling bouts.

Additionally, there have been changes to the way samples are uniquely identified.

- Pre-2021: Genetic archive samples had identifiers linked to the descriptive sampleID (ex: CPER_001-M-1-2.5-20200415-ga1). From mid-2021 onward they are identified by scannable barcodes only (ex: C00000234432).
- Pre-2022: The **sampleID** field contained the sampled X,Y coordinate (ex: WREF_001-O-35-4.5-20200715). From 2022 onward, it contains the pointID component of the subplot instead (ex: WREF_001-O-23-20200715).

Lastly, data collection for the soil nitrogen transformations and soil chemistry components of Soil physical and chemical properties, periodic (i.e., coordinated bouts) were suspended in 2026 due to unusual budget demands, leading to a 6-year interval between coordinated sampling events rather than the expected 5-year interval. The annual sampling of Soil physical and chemical properties, periodic is unaffected by this change and will be carried out in 2026 as usual. Coordinated bouts with nitrogen transformation and chemistry measurements originally scheduled for 2026 will be implemented in 2027, and the schedule originally intended for 2027 will be implemented in 2028 and so forth, leading to a consistent 6-year interval between sampling events across all sites in the schedule until 2032.

3.3.1 pH Method Change

Temporal consistency in methods for sample analysis is beneficial, but method changes may be adopted in order to realize efficiencies while still aligning with community standards. This was the case with soil pH measurements, where a modified method was adopted in 2020. Before implementing the change, tests from a small number of sites suggested there would be no impact to soil pH data. However, in reviewing data from all sites after the change was implemented, there may be site-specific effects on measured soil pH, especially in water-prepared mineral horizon samples. Users should keep this in mind while analyzing pH time series data.

- 2013 - 2019: pH measurements in water and CaCl2 utilized separate samples and both used a 1:2 soil to solution ratio for mineral horizons, following the methods in Robertson et al. (1999)
- 2020 and onward: The pH method of the USDA Natural Resource Conservation Service (Burt et al. 2014) was adopted, using a single container and soil sample to measure pH with both methods. First, water is added to soil and pH is measured at 1:1 soil to solution ratio. Then an equal volume of calcium chloride is added to the same container and pH is measured again at 1:2 soil to solution. These ratios generally apply to mineral soils, more solution is used for organic soils as well as very hygroscopic mineral soils. Soil to solution ratios are recorded in the sls_soilpH table in the 'waterpHRatio' and 'caclpHRatio' fields.

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 Soil physical properties, distributed periodic (DP0.10086.001) (AD[05]). All variables reported in the published data (L1 data) are also provided separately in the file, NEON Data Variables for Soil physical and chemical properties, periodic (DP1.10086.001) (AD[04]).

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), and the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), where applicable. NEON TOS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum andGeoid12A 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. These are indicated with downloadPkg = "none" in NEON Data Variables for Soil physical and chemical properties, periodic (DP1.10086.001) (AD[04]).

3.5 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a single X,Y sampling location.

sampleID (unique ID given to the individual soil sampling location and horizon) → **subplotID** (ID of subplot within plot) → **plotID** (ID of plot within site) → **siteID** (ID of NEON site) → **domainID** (ID of NEON domain).

During each bout, one location is sampled in three of the four subplots in each soil Base plot (Figure 1). The naming convention for subplots consists of the identity of the plot point in the southwest corner of that subplot and its size or scale. For example, subplot '21_400' is located with point 21 in the southwest corner and is 400 m2 (20 m x 20 m). In data releases prior to RELEASE-2024, subplot naming differed in that only the identity of the point in the southwest corner was used (e.g., '21' or '41'), omitting the scale. While this previous naming convention is retained as a component of soil sample identifiers, the subplotID field uses the corner_scale convention in RELEASE-2024 data and onward.

The basic spatial data included in the data downloaded include spatial location (latitude and longitude) and elevation of the *centroid* of the plot where sampling occurred, plus associated uncertainty due to GPS error and plot width. Shapefiles of all NEON Terrestrial Observation System sampling locations can be found on the NEON science spatial data webpage.

To derive a more precise estimate of the location of each soil sampling location, there are two options:

- Use the getLocTOS function from the geoNEON package, available here: https://github.com/NEONScience/NEON-geolocation
- Or follow these steps to perform the same calculation:

    1. Find the namedLocation field in the data; this is the named location of the plot. For example, 'MLBS_064.basePlot.bgc' is a complete named location for plotID 'MLBS_064'.
    2. Obtain easting, northing, and namedLocationCoordUncertainty for the plot centroid. This can be accomplished either by using the getLocByName function in the geoNEON R package, or using the NEON API (http://data.neonscience.org/data-api; e.g. https://data.neonscience.org/api/v0/locations/MLBS_064.basePlot.bgc.41?history=true) to query for centroid easting ("locationUtmEasting"), northing ("locationUtmNorthing"), and coordinateUncertainty ("Value for Coordinate uncertainty") of plot named locations. If the location data include more than one entry because spatial data changed, use the values corresponding to the date(s) when sampling occurred. The following instructions use the names provided by the geoNEON getLocByName function.
    3. Calculate the actual northing of the X,Y location. This can be accomplished by subtracting 20 meters from northing, then adding the y-coordinate value. Use the following formula:

$$yCoordNorthing = northing - 20m + coreCoordinateY$$

    4. Calculate the actual easting of the X,Y location. This can be accomplished by subtracting 20 meters from easting, then adding the x-coordinate value. Use the following formula:

$$xCoordEasting = easting - 20m + coreCoordinateX$$

    5. Increase namedLocationCoordUncertainty by an appropriate amount to account for variance in sampling location area (+/- 0.5 m), plus the error introduced by technicians navigating within plots, stretching meter tapes to navigate to X,Y locations, etc. This uncertainty, on

average, will be +/- 1 m, but will vary by site based on terrain heterogeneity and density of vegetation.

3.6 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is collectDate. All samples associated with a sampling event have collectDates within a ~14-day window and are categorized by the target season or seasonal transition (e.g. wet-dry transition, winter-spring transition, peak greenness). The total number of sampling events per year will vary among sites, based on the length of the growing season. It is expected that 3 sampling events will occur annually in most sites, except those in arctic and boreal regions.

In 2020 the **eventID** field was added in order to facilitate grouping of samples collected during the same sampling event or bout. The eventID is created by merging the following fields:

siteID + . + **sampleTiming** + . + **collectDate** (year portion only)

Example. ***WOOD.peakGreenness.2020***

For the handful of sites where sampling seasons straddle 2 calendar years (for example, Transition 1 in Oct, Peak Green in Jan, Transition 2 in March), the eventID contains both years separated by a dash.

Example. ***LAJA.peakGreenness.2021-2022***

3.7 Associated Data Streams

sampleID is the linking variable that can be used to join field metadata, moisture/pH, and soil chemistry data across tables. In addition, field and laboratory data from the Soil physical and chemical properties, periodic data product will be necessary to interpret and utilize several related soil microbial data products. For Soil microbe marker gene sequences (DP1.10108.001), Soil microbe community taxonomy (DP1.10081.002), and Soil microbe group abundances (DP1.10109.001, not actively collected), **geneticSampleID** in the sls_soilCoreCollection table is the variable name that links samples and their associated metadata from field to external laboratory data tables. For Soil microbe biomass (DP1.10104.001), this variable is **biomassID** published in the sls_soilCoreCollection table. For Soil microbe metagenome sequences (DP1.10107.001), the linking variable name is **compositeSampleID**, which is published in the sls_metagenomicsPooling table.

3.8 Product Instances

Soil samples are collected at all terrestrial NEON sites. A maximum of 10 plots are sampled at every site at a frequency of 3 times per year (except for arctic and boreal sites, which are sampled once). For each soil horizon (maximum of 2, organic or mineral), 3 samples per plot are collected, 1 in each of 3 randomly selected subplots. When organic and mineral horizons are present at a sampling location, both horizons are sampled only during Coordinated bouts, else only the top horizon is sampled.

Off-years: At most sites, expect soil sampling will result in 90 unique soil samples per site per year analyzed for moisture and pH. At boreal/arctic sites, expect 30 unique soil samples per site per year. Each unique sample may result in 0-8 sub-samples used for additional analyses and archiving.

Coordinated years: At most sites, expect soil sampling will result in 90-180 unique soil samples per site per year, analyzed for pH, moisture, and chemistry, and another 90-180 analyzed for net N transformation rates (incubated cores). At boreal/arctic sites, expect 30-60 unique soil samples analyzed for pH, moisture, and chemistry and another 30-60 analyzed for net N transformation rates (incubated cores). Each unique sample may result in 0-11 sub-samples used for additional analyses and archiving.

3.8.1 Product Instances for Missed or Incomplete Sampling

Beginning in 2020, missed or incomplete sampling that was not completed as scheduled is recorded using the **samplingImpractical** field located in the *sls_soilCoreCollection* data table. Any value for samplingImpractical that is not 'OK' indicates that sampling did not occur. A record is created for every expected sampling location (plot and subplot) that was unable to be sampled for a given year and sample timing. A **sampleID** will be generated for these records that consists of the following:

- plotID + "-" + pointID component of subplotID + "-" + collectDate (intended) + "-" + samplingImpractical value
- Example. "NIWO_013-23-20200501-locationFrozen"

The same number of records as outlined above are expected in sls_soilCoreCollection for an impractical sampling event. However, records are not expected in any downstream tables (e.g. *sls_Moisture*, *sls_pH*, *sls_bgcSubsampling*, *sls_metagenomicsPooling*, all chemistry data tables) if field samples were not practical to collect.

3.9 Data Relationships

The protocol dictates that each X,Y location within a randomly selected subplot yields a unique sampleID per horizon per **collectDate** (day of year, local time) in sls_soilCoreCollection. A record from sls_soilCoreCollection may have zero or one child records in sls_soilpH and sls_soilMoisture; a given sls_soilCoreCollection.sampleID is expected to be sampled only once. Depending on the type of bout and time of year (Figure 3), a record from sls_soilCoreCollection may have zero or one child records in sls_metagenomicsPooling and in sls_bgcSubsampling. Records for soil subsamples used in chemistry measurements will appear in associated downstream tables, described in the two related Data Product User Guides. 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*.

sls_soilCoreCollection.csv - > One record expected per **sampleID**. Depending upon **boutType**, each **sampleID** generates up to a single **geneticSampleID**, up to 5 interchangeable **geneticArchiveSamples**, and up to a single **biomassID**. Generates samples used in Soil microbe marker gene sequences (DP1.10108.001), Soil microbe community taxonomy (DP1.10081.002), Soil microbe group abundances (DP1.10109.001, not actively collected), and Soil microbe biomass (DP1.10104.001). Additionally, subsamples generated from soil **sampleID**s are used to measure soil inorganic N pools and transformations

sls_soilpH.csv - > One record expected per **sampleID**, generates a single **pHSampleID**. Missing records in this table indicate failure to carry out the measurement.

sls_soilMoisture.csv - > One record expected per **sampleID**, generates a single **moistureSampleID**. Missing records in this table indicate failure to carry out the measurement.

sls_metagenomicsPooling.csv - > One record expected per **plotID** per **horizon** per **eventID**. Record represents a mixture of the samples collected in a plot (listed in **toCompositeSampleIDList**). Each record generates a single **compositeSampleID**, used in Soil microbe metagenome sequences (DP1.10107.001).

sls_bgcSubsampling.csv - > One record expected per **sampleID**, generates a single **cnSampleID** and **bgcArchiveID** used for chemistry measurements and air-dried archive respectively. Chemistry values are reported in a separate carbon-nitrogen table.

sampleIDs will be generated for each unique physical soil sample created during a collection event, and **sampleBarcodes** may also be generated. Each **sampleID** will yield between one to five frozen archive subsamples, stored long term at -80C or in liquid nitrogen. These frozen archive samples vary in size, depending on the container type used. Beginning 2022-08-03, information on container type and size is tracked in a field called **geneticArchiveContainer**. Samples collected prior to this time will have missing entries for this field, but here is how missing values should be interpreted:

- 2013 to 2017, 2 oz (60 mL) whirl-pak bags
- 2018 to mid-2022, 5 mL cryovials
- Aug 2022 onward, either 2 mL or 5 mL cryovials depending on vendor availability, recorded in the data

When a collection event includes soil chemistry and isotope analyses, an air-dried archive subsample (**bgcArchiveID**) will also be created with an associated **bgcArchiveMass**. If not, following pH and moisture measurements, any remaining soil material will be discarded within one year of collection.

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

1. Every five years, incubation and processing of soil cores occurs in order to measure net nitrogen transformation rates. Incubated cores, designated with **nTransBoutType = tfinal**, will appear intermingled with non-incubated samples throughout the sls_soilCoreCollection and sls_soilMoisture data tables. However, they will not be measured for soil pH and will not be subsampled for microbial diversity or bulk soil chemistry and isotopic analyses. Refer to the related inorganic N Data Product User Guide for more details.
2. *Wetland sites*: At sites in which the Wetland SOP is carried out, standing water may be encountered during sampling. Litter depth and soil temperature are generally not measured for locations where **standingWaterDepth** > 0, unless technicians are certain there is no standing litter (e.g., flooded tundra) or if standing water is patchy such that temperature can be acquired from a non-flooded

location in the same 0.5 m radius as the X,Y sampling location.

3. *Alaska-specific consideration*: Beginning in 2018 the process for differentiating between organic soil and litter in arctic tundra and boreal sites was revised to more accurately reflect the slow rates of organic matter decomposition and the extensive fibric organic material at the TOOL, BARR, DEJU, BONA, and HEAL sites. Prior to 2018, instructions to technicians may have resulted in a portion of the top-most fibric material being included in measurements of litter depth, and excluded from O-horizon soil samples. From 2018 onward, the Protocol (AD[09]) dictates that all non-green, friable fibric material that has roots growing in it should be classified as organic soil and sampled as such, rather than being measured as litter. Users should use caution when comparing soil data from these five sites collected before and after 2018.

4. *Location resampled*: The soil protocol (AD[09]) mandates that each unique X,Y location within a NEON soil plot is sampled only once during the lifetime of the project. However, occasional errors lead to X,Y locations being resampled months to years apart. Resampled locations are identified by grouping multiple horizons and/or N-trans data for each sampling location within each sample timing and year; if sampling occurrences for an X,Y location are identified across multiple sample timings and years, a **dataQF** flag is added to 'sls_soilCoreCollection' data records for instances other than the first (intended) sampling. A flag is also added to **genomicsDataQF** for 'sls_metagenomicsPooling' data records if any of the composited samples are from resampled locations; individual samples from resampled locations are listed in the **remarks** field. Note: For certain older data (some of the oldest legacyData), workflows may have intentionally led to resampling of coordinates within the same sample timing and year; these resampling occurrences are not flagged.

5. *Pit location resampled*: This dataQF flag is added if samples were taken at or near the location of a soil characterization pit (see https://data.neonscience.org/data-products/DP1.10047.001). A flag is also added to **genomicsDataQF** for 'sls_metagenomicsPooling' data records if any of the composited samples are from resampled pit locations; individual samples from resampled locations are listed in the **remarks** field.

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. 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 Soil physical properties, distributed periodic (DP0.10086.001), 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 using a standardized data validation language (Nicl) internal to NEON. Please see AD[14]

and AD[15] for more information about the Nicl language.

Data collected prior to 2017 were processed using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow.

4.2 Automated Data Processing Steps

Following data entry into a mobile application or 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[13]). Additionally, the methods used to create calculated fields from the raw data (soil moisture, water-to-solution ratios for pH measurements) are detailed in NEON Raw Data Validation for Soil physical properties, distributed periodic (DP0.10086.001).

Published data are reviewed for completeness, timeliness, and validity using an internal set of tests and metrics, as detailed in the NEON Algorithm Theoretical Basis Document: OS Data Quality Control (AD[16]). These quality tests are used to guide process improvements, audits of analytical facilities, and data updates, but do not generate quality flags in published data.

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 catch-all quality flag for known errors applying to the record. The dataQF codes specific to the five soil tables discussed in this User Guide are detailed below. If multiple dataQF entries apply, they are presented as a pipe-delimited list. The issues with Alaska soils and resampled locations were detailed in the *Special Considerations* section above.

Table 2: 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
dataQF	alaskaDeprecatedMethod	Different methods used for measuring litter depth and the boundaries between soil horizons prior to 2018, use caution when comparing measurements to data collected in 2018 and later
dataQF	coordinatesUncertain	Exact location of soil sampling uncertain, precise location data may be inaccurate
dataQF	missing-pH-Data	Samples discarded before pH measurements could be taken, no pH data for the entire bout for associated soil samples. Currently a known issue for ORNL 2021 fallWinterTransition
dataQF or genomicsDataQF	location-resampled	The X,Y location where sample was collected has already been visited, deviates from NEON soil sampling design
dataQF or genomicsDataQF	pit-location-resampled	The X,Y location where sample was collected was at or near the location of a soil characterization pit, deviates from NEON soil sampling design
dataQF	not-all-biocrust-homogenized-in	mature biological soil crust was present at the sampling location but crust material was not mixed into the soil as thoroughly as instructed in the protocol.

Additionally, several other condition and quality fields have been added to the field collection and lab processing tables over time in order to communicate anomalous sampling conditions or method deviations. These include variables such as **horizonDetails**, **biophysicalCriteria**, **sampleCondition**, **geneticSamplePrepMethod**, and others. Definitions for the categorical codes used for these fields are included in the file NEON Categorical Codes for Soil physical and chemical properties, periodic (AD[06]), provided in the download package for this data product. Fields have been added over time and entries may be missing in older data.

For many of these fields, it is not possible to back-fill missing entries for older data. However for the **geneticSamplePrepMethod** and **geneticArchiveSamplePrepMethod** fields, which were added as of 2022-02-01, any missing values for samples collected before this date should be considered 'dry ice.' Starting in late 2021, NEON began to experience issues reliably sourcing dry ice in certain regions. This prompted creation of a field to record alternate storage conditions (e.g., ultra-cold ice packs) for -gen, -gaX, and -comp sample types. This issue was not encountered prior to late 2021 and all earlier microbial genetic analysis and archive samples were stored on dry ice.

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).

5 REFERENCES

Burt, R. (2014). Kellogg Soil Survey Laboratory Methods Manual. Report No. 42 Version 5.0. Soil Survey Laboratory Investigations. United States Department of Agriculture, Natural Resources Conservation Service. 1031 pp.

Robertson, G. P. (1999). Standard soil methods for long-term ecological research (Vol. 2). Oxford University Press on Demand.