NEON USER GUIDE TO SOIL MICROBE BIOMASS (NEON.DP1.10104)

PREPARED BY	ORGANIZATION	DATE
Lee Stanish	TOS	03/21/2018

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	3/9/2018	Initial Release

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 temperature 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 for Soil microbe biomass, and associated metadata, from input data on terrestrial samples. 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 Soil Microbe Biomass (NEON.DP1.10104) (AD[04]), provided in the download package for each of these three data products.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the laboratory data from samples generated by the field sampling protocols TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[06]) for upland soil samples, and with TOS Standard Operating Procedure: Wetland Soil Sampling (AD[07]) for wetland soil samples. The raw data that are processed as described in this document are detailed in the file, NEON Raw Data Validation for Microbe Biomass (NEON.DP0.10104) (AD[03]), 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.002652	NEON Level 1, Level 2 and Level 3 Data Products Catalog
AD[03]	NEON.DP0.10104.001 _dataValidation.csv	NEON Raw Data Validation for Microbe Biomass (NEON.DP0.10104)
AD[04]	NEON.DP1.10104.001 _variables.csv	NEON Data Variables for Soil Microbe Biomass (NEON.DP1.10104)
AD[05]	NEON.DOC.000908	TOS Science Design for Microbial Diversity
AD[06]	NEON.DOC.014048	TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling
AD[07]	NEON.DOC.004130	TOS Standard Operating Procedure: Wetland Soil Sampling
AD[08]	NEON.DOC.000913	TOS Science Design for Spatial Sampling
AD[09]	NEON.DOC.000008	NEON Acronym List
AD[10]	NEON.DOC.000243	NEON Glossary of Terms
AD[11]	OS_Generic_Transitions .pdf	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[12]		NEON's Ingest Conversion Language (NICL) specifications

2.2 Acronyms

Acronym	Definition
PLFA	Phospholipid Fatty Acid
qPCR	Quantitative Polymerase Chain Reaction

3 DATA PRODUCT DESCRIPTION

The Soil Microbe Biomass data product provides quantitative estimates of total microbial biomass in soil samples. NEON measures the abundances of numerous lipid biomarkers that are found in soil microbiota. Data are generated using the high-throughput phospholipid fatty acid (PLFA) analysis, in which the total phospholipid content of a soil sample is extracted, and discrete lipid molecules are quantified using Gas Chromatography and Mass Spectrometry (Buyer and Sasser 2012, Gomez et al. 2014). While there is no perfect method for quantifying microbial biomass in soils, PLFA analysis is widely considered to be a reliable (Zelles 1999) and sensitive (Allison and Martiny 2008) proxy. The sample plan implements the guidelines and requirements in the Science Designs for TOS Terrestrial Microbial Diversity (AD[05]). Information on sample collection methods such as frequencies per sample type can be found in the Soil Sampling Protocol (AD[06]) and Wetland SOP for wetland soils (AD[07]), and in the NEON User Guide to Soil Physical Properties, Distributed Periodic (NEON.DP1.10086).

Microbial biomass samples are a subset of the homogenized soil sample collected as part of the soil microbial diversity and biogeochemistry sampling. After field collection, bulk soil is stored on wet ice and transported to the NEON field laboratory. Within 24 hours, the field-moist, bulk soil is passed through a 2 mm sieve (for mineral horizons) or picked of rocks, roots and coarse debris (for organic horizons), and then a representative subsample (5-10 grams) is placed into a vial and stored at -80°C. Samples are shipped to an analytical laboratory where sample processing and analysis occurs.

3.1 Spatial Sampling Design

Sampling for soil microbe biomass analysis is executed at all NEON terrestrial sites, with data reported at the resolution of a single sampling location. This equates to a randomly-assigned X,Y coordinate ($\pm$ 0.5 meters) within a NEON plot. Ten plots are sampled at 3 randomly selected locations within each plot (Figure 1). In general, only the surface horizon is sampled to a maximum depth of 30cm, and horizons are broadly defined as either organic (O) or mineral (M).

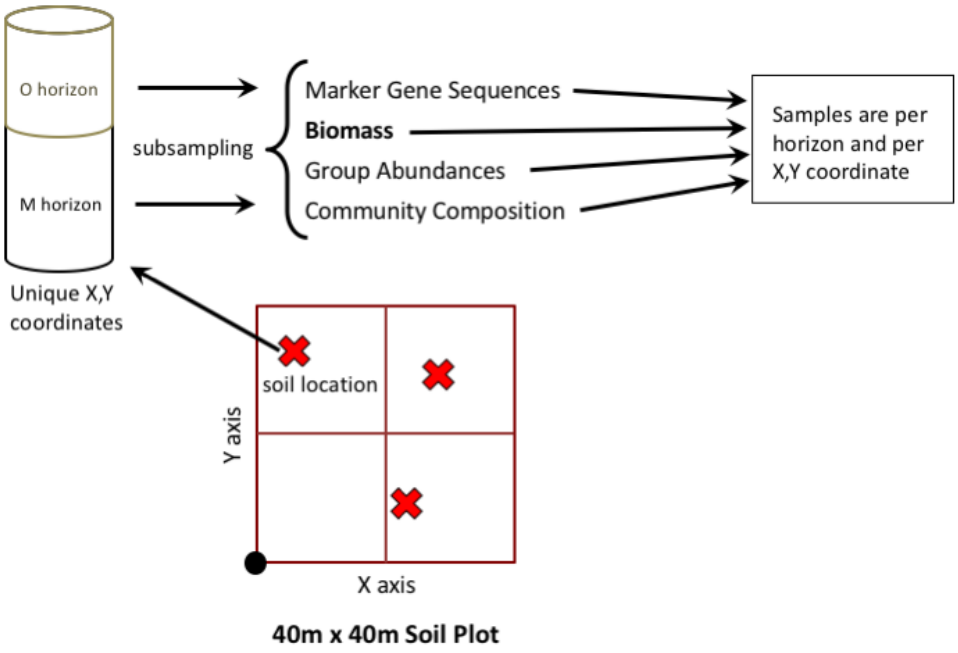

Figure 1: Overview of soil microbial field sampling, spatial design, and analysis workflow.

The spatial design for the microbial biomass data product is described in more detail in the Data Product User Guide for Soil Physical Properties (NEON.DP1.10086). For a description of the methods used in terrestrial plot selection, refer to the TOS Science Design for Spatial Sampling (AD[08]).

3.2 Temporal Sampling Design

Soil sampling for microbial biomass analysis occurs during a 'coordinated' bout, in which additional biogeochemical and isotopic measurements are made (DP1.10078), along with measurements of nitrogen transformation rates (DP1.10080). At most terrestrial sites, sampling occurs 3 times per year in conjunction with the soil physical properties data product (DP1.10086). Two sampling bouts occur during periods of seasonal transitions (e.g. winter-spring or wet-dry), and one during the period of peak greenness (as measured by remote sensing data). Only one sampling bout takes place at sites with short growing seasons (e.g. tundra and taiga), during peak greenness.

Up to 2 soil horizons (organic and mineral) are sampled for microbial analyses to a maximum depth of 30 cm.

For all samples, the temporal resolution is that of a single collection date. For a comprehensive description of field methods, refer to TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[06]). Descriptions of the upstream field data for soil (NEON.DP1.10086) sampling can be found in the Data Product User Guide for Soil Physical Properties.

3.3 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Microbe Biomass (NEON.DP0.10104) (AD[03]). All variables reported in the published data (L1 data) are also provided separately in the files within NEON Data Variables for Soil Microbe Biomass (NEON.DP1.10104) (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), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), where applicable.

Lipid names typically follow the common lipid nomenclature, with the definition including both the common and scientific names of the compound.

NEON TOS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and GEOID09 for its reference gravitational ellipsoid. 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.

3.4 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a single sampling location. This corresponds to a single X,Y coordinate location within a plot (Figure 1). The spatial hierarchy is as follows:

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

The spatial data are located in the data product Soil Physical Properties, distributed periodic (DP1.10086), in the table ***sls_soilCoreCollection***. The spatial data are measured at the plot *centroid*, and have an accuracury of ± 20 m. However, a more precise measurement may be determined by calculating the offset from the plot centroid using the variables **coreCoordinateX** and **coreCoordinateY**. Refer to the User Guide for Soil Physical Properties, distributed periodic, for more information and instructions.

3.5 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the **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 combine ("stack") files across months is available here: https://github.com/NEONScience/NEON-utilities

3.6 Associated Data Streams

This section describes the data products that are directly linked or closely related to the soil microbe biomass data product.

Soil data are derived from subsamples collected during soil biogeochemical and microbial sampling and include numerous related data products:

- Soil Physical Properties, distributed periodic (DP1.10086) - includes all field data associated with a soil sample. These data are linked to the microbial biomass data by the **biomassID** in the table ***sls_soilCoreCollection***.
- Soil microbe marker gene sequences (NEON.DP1.10108) - Microbial 16S and ITS sequence data. Use the Soil Physical Properties data product (NEON.DP1.10086, table ***sls_soilCoreCollection***) to obtain the **geneticSampleID** corresponding to the **biomassID**. The **geneticSampleID** is used for marker gene sequence analyses and links to a **dnaSampleID** in the table ***mmg_soilDnaExtraction***. The ***dnaSampleID*** is associated with the marker gene sequencing metadata found in the data tables ***mmg_soilPcrAmplification16S*(ITS)_** and ***mmg_soilMarkerGeneSequencing_16S(ITS)***.
- Soil microbe community composition (NEON.DP1.10081) - Microbial community composition data derived from marker gene sequencing. As for the marker gene sequences data, the **geneticSampleID** variable in the tables ***mcc_soilTaxonTable_16S*** and ***mcc_soilTaxonTable_ITS*** may be used to link data in this product to the microbial biomass data.
- Soil microbe group abundances (NEON.DP1.10109): Bacterial/archaeal and fungal abundances as measured by qPCR. The **geneticSampleID** variable in the table ***mga_soilGroupAbundances*** can be used to link data in this product to the microbial biomass data.
- Soil inorganic nitrogen pools and transformations (NEON.DP1.10080) - Measurements derived by field incubations of soil cores or buried bags. These data are linked to the microbial biomass data by the **sampleID** in the table ***sls_soilCoreCollection***.
- Soil chemical properties (Distributed periodic) (NEON.DP1.10078) - Measurements of soil carbon and nitrogen content. As with soil inorganic nitrogen data, the corresponding **sampleID** can be used to link data.
- Soil stable isotopes (Distributed periodic) (NEON.DP1.10100) - Measurements of soil carbon and nitrogen stable isotopes. As with soil inorganic nitrogen data, the corresponding **sampleID** can be used to link data.

3.7 Product Instances

A maximum of 10 plots will be sampled at every site one to three times per year. Up to 2 soil horizons will be collected as separate samples from each unique coordinate location. For each soil horizon sampled, 3 unique locations are collected at each plot, for up to 6 samples per plot and per sampling event. Thus, there will be 30-120 product instances generated per site per year.

3.8 Data Relationships

The protocol dictates that each X,Y location sampled yields a unique **sampleID** per horizon per collectDate (day of year, local time) in the table ***sls_soilCoreCollection*** for the data product Soil Physical Properties, distributed periodic (NEON.DP1.10086). Every bout type that includes biomass (e.g. the variable **boutType** includes the string

'biomass') should sample for microbial biomass analysis. A record from ***sls_soilCoreCollection*** may have zero or one child records in the table ***sme_microbialBiomass*** of this data product.

Each **biomassID** is a subsample of the parent **sampleID** in the table ***sls_soilCoreCollection***, and is sent for microbial biomass analysis. The PLFA results data appear in the table ***sme_microbialBiomass***, and are linked by the **biomassID**. One **biomassID** is expected per record. Duplicate records for an individual **biomassID** should not exist.

For each batch of samples recorded in ***sme_microbialBiomass***, one record is expected in the batch-level table ***sme_batchResults***. This table includes run-level data on standard and control samples for a batch of samples.

The table ***sme_labSummary*** is expected to have one record annually or when an update to the analytical methods occurs. This table includes long-term accuracy and variation data for lipid analytes used as analytical standards.

Duplicates and/or missing data may exist where protocol and/or data entry abberations have occurred; *users should check data carefully for anomalies before joining tables*.

Soil Physical Properties (NEON DP1.10086)

sls_soilCoreCollection.csv - > One record expected per **sampleID**. Generates samples used in Soil microbe biomass (NEON.DP1.10104), Soil microbe marker gene sequences (NEON.DP1.10108), Soil microbe community composition (NEON.DP1.10081), and Soil microbe group abundances (NEON.DP1.10109). Additionally, subsamples generated from soil sampleIDs are used in Soil inorganic nitrogen pools and transformations (NEON.DP1.10080).

Soil Microbe Biomass (NEON.DP1.10104)

sme_microbialBiomass.csv - > One record expected per **biomassID**. A biomassID will represent one sample per plot/horizon/X,Y coordinate combination and per collectDate (day of year, local time). There will be only one biomass sample per **biomassID**. For each batch of samples run, a **batchID** is generated.

sme_batchResults.csv - > One record is expected per **batchID**, which corresponds to the **batchID**, in the table ***sme_microbialBiomass***.

sme_labSummary.csv - > The laboratory reports long-term accuracy and variation in known analytes every 6 months to 1 year, or when analytes, methods or instrumentation changes. Thus, at least 1 record is expected from a unique laboratory per year. The **labSpecificStartDate** and **labSpecificEndDate** can be used to apply the lab summary data to individual batches of data.

3.9 Special Considerations

4 DATA QUALITY

4.1 Data Entry Constraint and Validation

Many quality control measures are implemented on the laboratory data at the point of data ingest into the NEON database. For example, data formats are constrained and data values are controlled through the provision of con-

trolled list of values (LOV's), which reduces the number of processing steps necessary to prepare the raw data for publication. An additional set of constraints is 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 Microbe Biomass (NEON.DP0.10104). This document is provided with every download of this data product. Contained within this file is a field named 'entryValidationRulesParser', which describes syntactically the validation rules for each field built into the data ingest validation. 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[12]).

4.2 Automated Data Processing Steps

Following laboratory submission of metadata into the NEON automated data ingest process, 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[11]).

4.3 Data Revision

All data are provisional until a numbered version is released; the first release of a static version of NEON data, annotated with a globally unique identifier, is planned to take place in 2020. 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 Change Log section of the data product readme, provided with every data download, 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.

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

4.5 Analytical Facility Data Quality

All analytical labs that generate microbial biomass data include standards run as unknowns alongside NEON samples in order to gauge run acceptability. Long-term analytical precision and accuracy of these standard analyses are reported for each lab to allow users to interpret and analyze lipid concentrations in the context of their uncertainty ranges. The data table sme_labSummary (available in the expanded package) contains the long-term precision (**analyteStandardDeviation**) and accuracy (**analyteAccuracy**) of lab analyses.

In addition, labs communicate record-level issues with samples or measurements using the quality flags described in the Quality Flagging section. In general, a null entry in a quality flag field means there is no issue to report.

For further information about individual laboratory QA procedures, refer to the lab SOPs, found on the NEON Data Portal (http://data.neonscience.org/home) in the Resources > Document Library > External Lab Protocols section.

5 REFERENCES

1. Allison, S.D. and J.B.H. Martiny. 2008. Resistance, resilience, and redundancy in microbial communities. Proceedings of the National Academy of Sciences, USA **105**:11512-11519.
2. Buyer, J.S. and M. Sasser. 2012. High throughput phospholipid fatty acid analysis of soils. Applied Soil Ecology **61**:127-130.
3. Gomez, J.D., K. Denef, C.E. Stewart, J. Zheng, and M.F. Cortrufo. 2014. Biochar addition rate influences soil microbial abundance and activity in temperate soils. European Journal of Soil Science **65**:28-39.
4. Zelles, L. 1999. Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation of microbial communities in soil: a review. Biological Fertility in Soils **29**:111-129.