NEON USER GUIDE TO MOSQUITOES SAMPLED FROM CO2 TRAPS (NEON.DP1.10043) AND MOSQUITO-BORNE PATHOGEN STATUS (NEON.DP1.10041)

PREPARED BY	ORGANIZATION	DATE
Katherine LeVan	SCI	03/01/2018
		03/01/2018

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	07/19/2017	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, 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 two L1 data products: Mosquitoes sampled from CO2 traps and Mosquito-borne pathogen status and associated metadata from input data. This document also 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 Mosquitoes sampled from CO2 traps (NEON.DP1.10043) (AD[05]) and NEON Data Variables for Mosquito-borne pathogen status (NEON.DP1.10041) (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: Mosquito Sampling (AD[09]). The raw data that are processed in this document are detailed in the file, NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[04]), 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., '10043') 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 Level 1, Level 2, and Level 3 Data Products Catalog
AD[04]	NEON.DP0.10043.001 _dataValidation.csv	NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043)
AD[05]	NEON.DP1.10043.001 _variables.csv	NEON Data Variables for Mosquitoes sampled from CO2 traps (NEON.DP1.10043)
AD[06]	NEON.DP1.10041.001 _variables.csv	NEON Data Variables for Mosquito-borne pathogen status (NEON.DP1.10041)
AD[07]	NEON.DOC.000910	TOS Science Design for Mosquito Abundance, Diversity and Phenology
AD[08]	NEON.DOC.000911	TOS Science Design for Vectors and Pathogens
AD[09]	NEON.DOC.014049	TOS Protocol and Procedure: Mosquito Sampling
AD[10]	NEON.DOC.000008	NEON Acronym List
AD[11]	NEON.DOC.000243	NEON Glossary of Terms
AD[12]	OS_Generic_Transi-tions.pdf	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[13]	Nicl Language.pdf	NEON's Ingest Conversion Language (NICL) specifications

2.2 Acronyms

Acronym	Definition
CDC	Centers for Disease Control and Prevention

3 DATA PRODUCT DESCRIPTION

mosquitoes are sampled in the field using CDC CO_2 light traps. Following collection, mosquito samples are sent to an external facility where they are sorted to remove bycatch and taxonomically identified (to species and sex, whenever possible). In the case of large field samples, a subsample of up to 200 individual mosquitoes is taxonomically identified but both total weights of the field collected sample and the subsample are provided to inform estimates of total abundance. Identifications for a subset of difficult taxa are verified by DNA barcoding. For additional details on the sampling design and associated protocol, see the TOS Science Design for Mosquito Abundance, Diversity and Phenology (AD[07]) and TOS Protocol and Procedure: Mosquito Sampling (AD[09]).

Mosquito-borne pathogen sampling involves the testing of all or a subset of collected mosquitoes for infection by viral pathogens by one or more external facilities. Only female mosquitoes identified to the species-level and captured in sufficient quantity over a season from likely vector species are eligible for pathogen testing. A set of up to 1000 individual mosquitoes per species per site per year are targeted for pathogen testing of arboviruses within the families Bunyaviridae, Alphaviridae, and Flavivirdae. Mosquito species are prioritized into three tiers, with highest priority given to Tier 1 species (*Aedes aegypti* and *Aedes albopictus*); followed by Tier 2 species: *Culex tarsalis*, *Culex pipiens*, and *Aedes triseriatus*; and then Tier 3 species: any other individuals identified to the species-level within the genera of *Aedes* and *Culex*. To be eligible for pathogen testing, a species must have a minimum of 100 (Tier 1 species) or 200 (Tiers 2 & 3) individuals present at a site in a year that are available for testing.

Following identification, mosquitoes are combined by species, sex, and bout (e.g., all female mosquitoes of species A collected at site C during sampling bout D). Groups of individuals combined for testing are assigned the same testingID. This large pool is then subdivided into vials (testingVialID), which contains a defined number of mosquitoes (generally a poolSize of 10-50 individuals). Each testingVialID is tested one or more times using a variety of methods which may include RT-PCR, Vero cell culture, and melt curve assays. These methods vary in target specificity, from general (e.g., Vero cell culture) to specific viral species (e.g., RT-PCR). Most pools of mosquitoes are negative because pathogens are rare; when pools are determined to be positive for any virus, the identit(ies) of the virus(es) are determined to the species-level, if possible. Test results yield data on the presence of important mosquito pathogens (e.g., West Nile virus, Eastern equine encephalitis virus, Dengue, etc) in a subset of species that are known vectors of disease. See the TOS Science Design for Vectors and Pathogens (AD[08]) for additional background on mosquito-borne pathogen sampling.

3.1 Spatial Sampling Design

Mosquito sampling is executed at all terrestrial NEON sites and follows a spatially-balanced stratified random design (AD[02]). Mosquitoes are sampled at 10 mosquito points per site. Points are randomly positioned within each National Land Cover Database (NLCD) class with representation within each NLCD class set as proportional to its representation at the site; NLCD classes with less than 5% representation are excluded from sampling. For ease of deployment, mosquito plots are always located between 5-45 meters distance from a road accessible to sampling by NEON technicians. Mosquito points must be separated by a minimum of 310m and must be 10m from the edge of other NEON sampling locations.

3.2 Temporal Sampling Design

When adult mosquitoes are active, sampling bouts will occur every two weeks at core sites and every four weeks at relocatable sites. The finest temporal resolution at which mosquito data (for the purposes of species richness, abundance and phenology) will be tracked is trapping night or trapping day. The finest level of temporal resolution at which mosquito-borne pathogen status will be tracked is at the level of a sampling bout. Two trapping nights and the intervening day for up to ten plots comprise a sampling bout of three separate samples per trap (30 samples per site; see Figure 1). The setDate (indicating when the trap was set) and collectDate (indicating when the trap was collected) will be recorded for each sample collected during a bout. Bouts are grouped using the **eventID** designation (a descriptor that includes the year of sampling, the site ID, and the calendar week in which a sampling bout occurred). Infrequently, a bout may be scheduled over 2 ISOweeks such that a bout of 30 samples will span 2 **eventID**s.

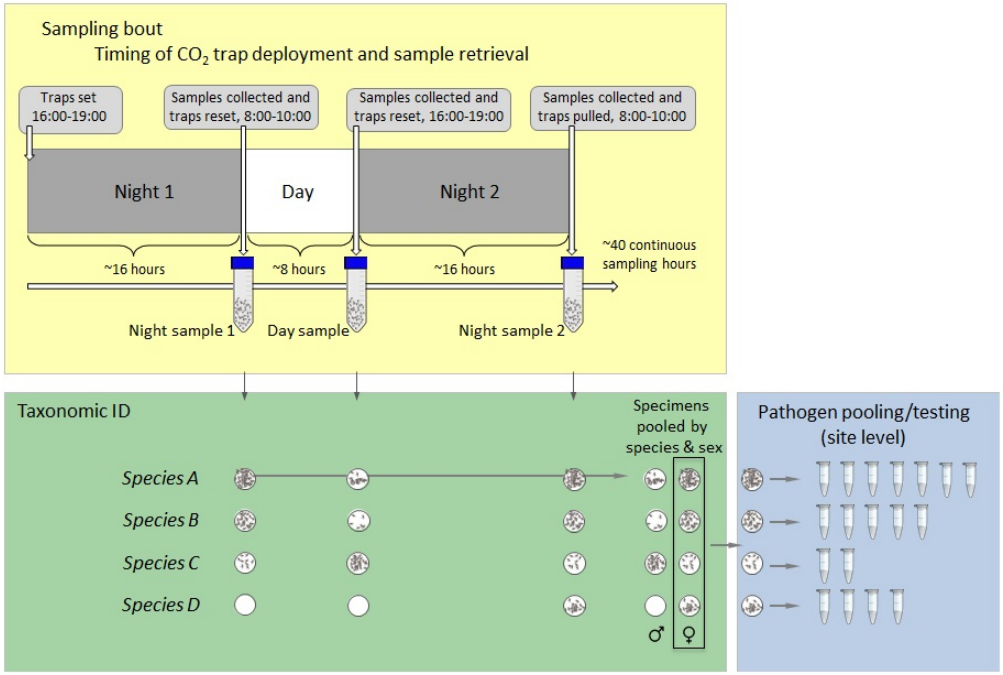

Figure 1: A workflow illustrating the process of data collection for the mosquito protocol

The total number of bouts per year varies among sites based on seasonality of each site (e.g., stopping during winter at temperate sites). During the time of year when mosquitoes are flying, sampling bouts occur every 2 weeks at the core site and every 4 weeks at each relocatable site, alternating between the core and a relocatable such that one site in the domain is sampled each week. After the mosquito season has ended (e.g., upon the onset of winter), weekly sampling at the core site will monitor for mosquito presence and help determine when the next mosquito sampling season should begin. A given sampling bout will be cancelled if minimum ambient temperature thresholds are not met. Additional details about sampling bout frequency can be found in the TOS Protocol and Procedure: Mosquito Sampling (AD[09]).

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 Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the files, NEON Data Variables for Mosquitoes sampled from CO2 traps (NEON.DP1.10043) (AD[05]) and NEON Data Variables for Mosquito-borne pathogen status (NEON.DP1.10041) (AD[06]).

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 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 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the **trapHours**, the range between **setDate** and **collectDate**.

collectDate (date an individual trap was collected) → **trapHours**

3.5 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a single trap (Figure 1).

plotID (unique ID given to the individual trap) → **siteID** (ID of NEON site) → **domainID** (ID of a NEON domain).

The basic spatial data included in the data downloaded include the latitude, longitude, and elevation of the plot marker where trapping occurred + associated uncertainty due to GPS error (trapping data) or the latitude, longitude and elevation of the NEON tower at each site (pathogenresults), since mosquitoes are typically pooled across plots within a site. Shapefiles of all NEON Terrestrial Observation System sampling locations can be found here: http://www.neonscience.org/science-design/field-sites/maps-spatial-data.

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

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

trapping: 1. Precise geocoordinates for the plot marker and associated coordinate uncertainly are provided in the downloaded data 2. Technicians are permitted to move up to 10m from the marked location to find a suitable place to install traps. Thus realized coordinate uncertainty on trap placement = **coordinateUncertainty** + 10m

pathogenresults: 1. Precise geocoordinates for the plot marker and associated coordinate uncertainly are provided in the corresponding trapping data records 2. Weighted averaging of trap locations for mosquitoes contributing to a given pathogenpooling **testingID**

3.6 Associated Data Streams

testingID is the linking variables that tie specific samples and associated metadata between the Mosquitoes sampled from CO2 traps data product (NEON.DP1.10043) and Mosquito-borne pathogen status data product (NEON.DP1.10041).

3.7 Product Instances

There are a maximum of 26 field season collection bouts per year, with mosquitoes collected from no more than 10 plots per bout. Each plot will yield no more than 3 samples per bout of collection. Thus, no single site should ever exceed 780 trapping data product instances in a given calendar year. The number of records for identification and pathogenresults varies with the diversity of the site and pathogen testing workflows.

3.8 Data Relationships

The protocol dictates that 3 samples are recovered (if present) from each trap per bout (roughly 30 expected records per eventID in mos_trapping). A record from mos_trapping may have zero or more child records in mos_sorting, depending on whether the trap contained mosquitoes (**targetTaxaPresent** = 'Y') downloaded. A record from from mos_sorting may have one or more child records in mos_identification depending on the number of species detected within the sample, if any. A record from from mos_identification may have zero or one child records in mos_archivepooling (if any material is archived) and zero or more child records in mos_pathogenpooling (if any material is pathogen tested). A record from from mos_pathogenpooling may have one or more child records in mos_pathogenresults depending on the number of tests run on a given testingVialID. 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.

mos_trapping.csv - > One record expected per sampleID for all time; one record per plotID per eventID

mos_sorting.csv - > One record expected per sampleID for all time, generates a single subsampleID if mosquitoes and/or bycatch is present in the trap

mos_identification.csv - > One record expected per subsampleID per scientificName per sex per identificationQualifier combination. The value in **individualCount** per subsampleID per scientificName per sex per identificationQualifier represents the total number of individuals of that type found within the subsample. Up to 10 individuals of a given scientificName/sex/identificationQualifier combination may be removed from a given subsampleID and pinned; the individualID of each pinned individual is given in a pipe-delimited list in the column labelled **individualIDList**. If the mosquitoes of a given scientificName/sex/identificationQualifier combination are pooled from a subsampleID into a pool of mosquitoes for archiving or a pool of mosquitoes for testing, then the archiveID or testingID that the subsampleID contributed to will be listed.

mos_archivepooling.csv - > One record expected per archiveID, which is a mixture of subsampleIDs (listed in archiveID column of the mos_indentification file). Not all subsampleID's contribute to mixtures; some are tested or pinned.

mos_pathogenpooling.csv - > One record expected per testingVialID, which is a mixture of subsampleIDs (listed in testingID column of the mos_indentification file). Not all subsampleID's contribute to mixtures; some are archived or pinned.

mos_pathogenresults.csv - > One or more records expected per testingVialID, which is a subsample from a given testingID (number of individuals within that subsample listed in the 'poolSize' column of the mos_pathogenpooling file).

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 Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[04]), 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. Also included in this file is a field named 'entryValidationRulesParser', which describes syntactically the validation rules for each field that is performed upon ingest of the data into the NEON Cyberinfrastructure, based on a standardized data validation language (Nicl) internal to NEON. Please see AD[13] 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 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[12]).

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. There are currently no dataQF codes in use in this data product.

4.5 Analytical Facility Data Quality

Three percent of all samples are quality checked for taxonomic difference by reprocessing at the external facility and percent difference in enumeration (PDE) and Percent Taxonomic Disagreement (PTD) are calculated (Stribling et al. 2008). Percent difference in enumeration (PDE) must not exceed 5%; PTD must not exceed 2% at the genus level and 5% at the species level. Where values exceed these thresholds, discrepancies are reconciled in the final datasheets. Notes on subsamples where QC was performed and results can be found in the 'remarks' column of the applicable tables.Details on the calculations of these fields can be found in the external lab SOP.

5 REFERENCES

Stribling, J. B., K. L. Pavlik, S. M. Holdsworth, and E. W. Leppo. 2008. Data quality, performance, and uncertainty in taxonomic identification for biological assessments. Journal of the North American Benthological Society. 27: 906-919.