NEON USER GUIDE TO TICK-BORNE PATHOGEN STATUS (DP1.10092.001)

PREPARED BY	ORGANIZATION
Sara Paull	SCI
Sarah Elmendorf	SCI
Kate Thibault	SCI

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	06/19/2017	Initial Release
B	05/15/2019	Revision
C	06/04/2020	Included general statement about usage of neonUtilities R package and statement about possible location changes.
C	11/01/2020	Added details about sampling design changes and dataQFs.
D	03/16/2022	Updated section 4.3 Data Revision with latest information regarding data release
E	08/23/2022	Added Haemaphysalis longicornis pathogen testing information

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 the L1 data product Tick-borne pathogen status - the presence/absence of select pathogens in collected ticks - 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 file, NEON Data Publication Workbook for Tick-borne pathogen status DP1.10092.001) (AD[05]), 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: Tick and Tick-Borne Pathogen Sampling (AD[07]). The raw data that are processed in this document are detailed in the file, NEON Data Ingest Workbook for Tick Sampling (DP0.10093.001) (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., '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	Validation csv
AD[05]	Available with data download	Variables csv
AD[06]	NEON.DOC.000911	TOS Science Design for Vectors and Pathogens
AD[07]	NEON.DOC.014045	TOS Protocol and Procedure: Tick and Tick-Borne Pathogen Sampling
AD[08]	NEON.DOC.000913	TOS Science Design for Spatial Sampling
AD[09]	NEON.DOC.000913	TOS Science Design for Plant Diversity
AD[10]	NEON.DOC.000008	NEON Acronym List
AD[11]	NEON.DOC.000243	NEON Glossary of Terms
AD[12]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[13]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[14]	Available on NEON data portal	NEON Ingest Conversion Language
AD[15]	Available with data download	Categorical Codes csv

3 DATA PRODUCT DESCRIPTION

Ticks are sampled in the field using drag and/or flag sampling (Milne 1943, Falco and Fish 1992, Ginsberg and Ewing 1989, Rulison et al. 2013). Collected ticks are enumerated by life-stage and taxonomic group. See data product Ticks Sampled Using Drag Cloths (DP1.10092.001) for details on collection and identification of ticks. Following identification, a target of 130 individual nymphs per site per year are selected for pathogen testing, with the total numbers tested in a given site-year dependent on availability. *Ixodes scapularis* and *Ixodes pacificus* (if available beyond what is used for archive) nymphs are targeted for testing of *Anaplasma phagocytophilum*, *Babesia microti*, *Ehrlichia muris*-like agent, *Borrelia burgdorferi* sensu lato, *Borrelia miyamotoi*, and *Borrelia mayonii*. Nymphs in the genera *Amblyomma* and *Dermacentor* (if available beyond what is used for archive) are targeted for testing of *Francisella tularensis*, *Rickettsia rickettsia*, *Anaplasma phagocytophilum*, *Ehrlichia chaffeensis*, *Ehrlichia ewingii*, and *Borrelia lonestari*. Nymphs of the species *Haemaphysalis longicornis* are tested for *Anaplasma phagocytophilum*, *Babesia microti*, *Borrelia burgdorferi*, *Borrelia lonestari*, *Borrelia mayonii*, *Borrelia miyamotoi*, *Ehrlichia chaffeensis*, *Ehrlichia ewingii*, *Ehrlichia muris*-like agent, *Francisella tularensis* and *Rickettsia rickettsii*.

Ticks transmit numerous pathogens of wildlife, livestock, and humans, including the etiological agent of Lyme disease (*Borrelia burgdorferi*), the most frequently reported vector-borne disease of humans in the United States. Among arthropod vectors, ticks are particularly sensitive to meteorological conditions and associated physiological constraints, making it highly likely that the demography and biogeography of many tick species, and the pathogens they transmit, will be affected by climate change. Further, the multi-host lifecycles of most tick species increase their ecological connectivity and sensitivity to community-level perturbations that may arise from changes in human land- and resource-use practices.

3.1 Spatial Sampling Design

Tick sampling is conducted at regular intervals by NEON field technicians at terrestrial sites. At each site, six plots are selected for sampling, distributed within the sites proportional to permitted sampling area contained within each National Land Cover Database (NLCD) class. Tick plots are situated such that the edge of each tick plot must be >150m from other NEON plots and infrastructure; tick plot centers are >500m from each other and no streams bisect tick plot perimeters. To reduce the probability that the sampling activities in other protocols impact tick diversity and distribution (e.g., technicians inadvertently attracting or redistributing ticks), tick plot centers are offset from distributed plot centers according to a specified distance (150m +/- 15m) in a randomly chosen direction selected prior to establishment of plots in the field. Sampling occurs by drag or flag sampling the 160m perimeter of each of the six 40m x 40m tick plots (Figure 1). See TOS Science Design for Vectors and Pathogens (AD[06]), TOS Protocol and Procedure: Tick and Tick-Borne Pathogen Sampling (AD[07]), TOS Science Design for Spatial Sampling (AD[08]) and TOS Science Design for Plant Diversity (AD[09]) for further details.

An exception to the design above occurs at NEON's Guanica Forest site (GUAN). At GUAN, the density of vegetation precluded sampling in any of the randomly selected plots using the above design. Instead, field staff located and marked six 'transect paths' where sampling is feasible. Each transect path contains between 80m^2 and 160m^2 of sampled area. Dragging and/or flagging then occurs over one or more 10m long segments within each plot, with each segment generally within 10m of another segment. Where possible, transect paths are located within 300m from the center of a mammal plot; 300m from the

center of a bird plot; and 300m from the center of a Distributed Base Plot, with preference given to co-location with mammal plots where all three co-location criteria cannot be met.

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/

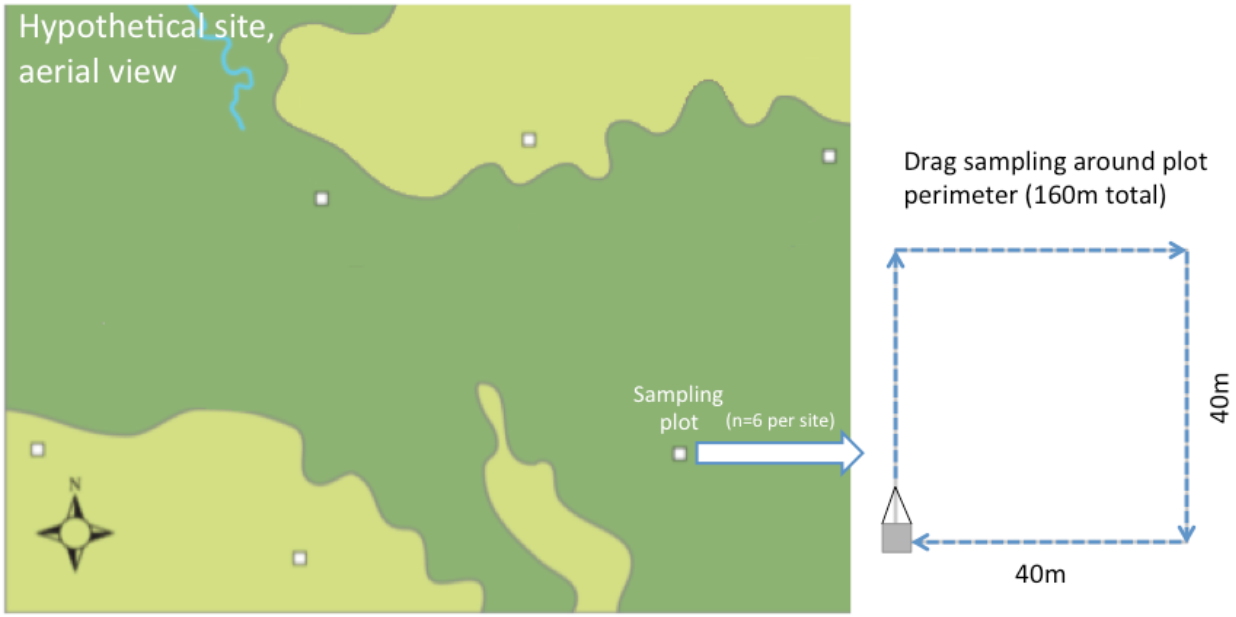

Figure 1: Drag and/or flag sampling occurs around the 160m perimeter of a 40mx40m plot

3.2 Temporal Sampling Design

Ticks are sampled during the growing season, which is determined by site-specific average greening period. During the growing season, sampling is conducted every three weeks at sites with high intensity sampling, and every six weeks at sites with low intensity sampling. Between 2013-2018, high intensity sampling occurred at any site where a single tick was detected within the last year (i.e., 365 calendar days). After 2018, high intensity sampling occurs at sites where more than five ticks have been detected in the last year. Sampling is conducted only when the ground is dry. Sampling also occurs only if the high temperature on two consecutive days prior to planned sampling exceeds 0°C. For additional details, see the TOS Protocol and Procedure: Tick and Tick-Borne Pathogen Sampling.

3.3 Sampling Design Changes

There have been several design changes that have been implemented over the course of data collection. Such changes arise due to continual evaluation of the sampling design for best practices in collaboration with technical working groups. They also occur when optimization of the design is necessary to ensure

that allocation of sampling effort is poised to maximize returns to the scientific community. A more complete list of design changes for tick data collection, including those less likely to influence interpretation of tick pathogen data, is reported in the NEON User Guide to Ticks Sampled Using Drag Cloths.

From 2015-2018, the taxonomy laboratory stopped identifying and counting ticks when invoice limits of 600 samples per domain were reached. One sample unit was counted as a single adult or nymph identification, or a full larval count from a plot. The uncounted samples are missing from the identification table.

Between 2016-2018 when more than 500 larval ticks were collected by the field team, a subset of approximately 500 was shipped to the taxonomy laboratory. In all cases where >500 larvae were collected in the field between 2016-2018, the associated identification record has 'ID lab count subsample of total field larvae' in the dataQF field.

Ticks collected during the 2016 field season were stored in RNA buffer solution. All other ticks are stored in 90-95% ethanol.

Starting with the 2019 field season, tick counts are no longer recorded by the field teams. Instead, ticks collected in the field are shipped to the taxonomy laboratory where all ticks are separated by life stage and counted. All nymph and adult ticks are identified up to the invoice limit. Ticks received in excess of invoice limits are entered as counts with the scientific name entered as *Ixodidae sp*. The taxonomy lab distributes identification effort evenly across bouts and life stages, randomly selecting ticks for identification within those criteria.

Beginning in the 2020 field season, the field **samplingImpractical** was added to the data to allow for the generation of a record when a plot could not be sampled for a particular bout. If field sampling was not possible **samplingImpractical** is populated with a value other than 'OK' (e.g., 'location flooded'). If sampling could not be conducted for all or part of the bout, the samplingImpractical field will communicate such missing records and the reason therefore.

Beginning in the 2021 field season *Haemaphysalis longicornis* was added to the list of tick species targeted for pathogen testing.

3.4 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the file, NEON Data Ingest Workbook for Tick Sampling (DP0.10093.001) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the file, NEON Data Publication Workbook for Tick-borne pathogen status DP1.10092.001) (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 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.

3.5 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a single plot.

plotID (ID of plot within site) → **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 centroid of the plot where sampling occurred + associated uncertainty due to GPS error and plot width. Shapefiles of all NEON Terrestrial Observation System sampling locations can be found in the Document Library: http://data.neonscience.org/documents.

3.6 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the range between the startDate and endDate of each tick collection event (the duration during which dragging and/or flagging a single plot occurred).

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. All queries, regardless of the date range specified, will include a copy of tck_pathogenqa which provides data from the pathogen lab about analytical quality control. Code to stack files across months is available here: https://github.com/NEONScience/NEON-utilities

3.7 Associated Data Streams

subsampleID is a linking variable that ties specific tick identifications and associated metadata to the Ticks sampled using drag cloths data product (DP1.10093). Visual assessments of the presence of ticks by life stage on captured mammals can be found in the Small mammal box trapping data product (DP1.10072). Beginning in the 2020 field season, a binned total count of ticks found attached to captured small mammals is also available.

Beginning in the 2020 field season, the rodent-borne pathogen status data product (DP1.10064) will include results of pathogen tests of small mammal ear tissues for *Borrelia burgdorferi*. Small mammal blood samples will also be tested for other tick-borne pathogens, including many of those currently analyzed in ticks.

3.8 Product Instances

A target of 130 individual ticks per site per year are selected for pathogen testing, with the total numbers tested in a given site-year dependent on availability. There is one record for every tick pathogen for each individual tick tested, yielding a variable number of tck_pathogen records per site per year. The specific set of pathogens for which testing is performed differs by tick genus; however each tick has a row of data reported for the full list of tick pathogens (rather than the subset of pathogens for which it was tested). The pathogens that are not tested for a particular tick genus have the testResult field left blank.

3.9 Data Relationships

The protocol dictates that each plot is sampled once per sampling bout (six expected records per eventID in the Ticks sampled using drag cloth product tck_fielddata). A record from Ticks sampled using drag cloth product tck_fielddata may have zero (if no ticks present) or more child records in Ticks sampled using drag cloth product tck_taxonomyProcessed. An equivalent number of records will occur in Ticks sampled using drag cloth product tck_taxonomyRaw. A record from tck_taxonomyProcessed may have zero (if the pertinent subsampleIDs are not sent for pathogen testing) or more records in the tck_pathogen table. 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.

tck_fielddata.csv (DP1.10093.001) - > One record expected per plotID for each eventID, generates a single sampleID if ticks are present

tck_taxonomyProcessed.csv (DP1.10093.001) - > Multiple records expected per sampleID, generates a unique subsampleID for each species*life stage combination found per sampleID. The taxonomic nomenclature in this file has been standardized and desynonymized according to NEON's master taxonomy for ticks.

tck_taxonomyRaw.csv (DP1.10093.001 - expanded package) - > Multiple records expected per sampleID, generates a unique subsampleID for each species*life stage combination found per sampleID. The taxonomic nomenclature in this file reflects the verbatim identifications provided by the external taxonomist and may contain synonyms.

tck_pathogen.csv (DP1.10092.001) - > Multiple records expected per subsampleID, generates a unique testingID for each tick selected for pathogen testing. Note not all subsampleIDs contribute to testing; some are simply archived in 95% EtOH (or an RNA buffer solution for 2016 ticks). Note that each subsampleID present in the pathogen data has a row for every possible pathogen type tested for across all genera of ticks. If an individual was not tested for a particular pathogen type (because it is not in the tick genus for which that pathogen test is performed), the testResult field is blank for that subsampleID-testPathogenName combination.

tck_pathogenqa.csv (DP1.10092.001 - expanded) - > One record expected per batchID, contains relevant batch-level quality assurance procedures for each batch tested over all dates testingID for each tick selected for pathogen testing. Note not all subsampleIDs contribute to testing; some are simply archived in 95% EtOH (or an RNA buffer solution for 2016 ticks).

Data downloaded from the NEON Data Portal are provided in separate data files for each site and month requested. The neonUtilities R package contains functions to merge these files across sites and months into a single file for each table described above. The neonUtilities 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. For instructions on using neonUtilities 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

4 DATA QUALITY

4.1 Data Entry Constraint and Validation

Constraints and data validation 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 Data Ingest Workbook for Tick Sampling (DP0.10093.001), 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 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[13]).

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. 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 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. Note that this applies to all data collected prior to 2017; however only a subset of those data are marked with this dataQF.
dataQF	ID lab count sub-sample of total field larvae	More than 500 larvae were collected by the field team, but only a subset of approximately 500 were shipped to the taxonomy laboratory. The resultant larval counts in the identification table reflect this sub-setted number and the field collection larvae counts represent a more accurate number. This occurred in the 2016-2018 field seasons.
dataQF	field/ID lab larva/nymph/adult count higher than field/ID lab (PDE > 25 percent)	The percent difference in enumeration, or PDE, between counts reported by the field and taxonomy laboratory is more than 25 percent. PDE is calculated as $\frac{\lvert fieldcount - labcount \rvert}{fieldcount + labcount} * 100$. There will be a separate dataQF indicating which count is higher for each instance of a life stage where field and lab counts differ by more than 25 percent

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)

The quality flags for discrepancies of more than 25% between field and taxonomy laboratory counts within a given life stage are intended to help with the determination of whether the field or the laboratory count is the more reliable datapoint. Counts from the identification lab for a given life stage are sometimes higher than field counts, and this could occur if the field team incorrectly identified the life stage of the ticks. This would be more likely if there is another quality flag for the same sample indicating that the field count was high for a different life stage. Counts from the identification lab for a given life stage are sometimes lower than field counts, which could indicate that the identification lab was subsampling and did not count all of the ticks received due to invoice limitations. These excess, uncounted ticks were not recorded in the data. The counts from the identification laboratory, therefore would be more likely to be low for this reason if the counts are from a domain and year where subsampling occurred (e.g., more than 600 samples from a given domain were generated in a single year). A remark stating: 'site-year possibly subsampled in ID lab counts' has been added to to the collection and identification tables for those samples with count-related quality flags.

4.5 Analytical Facility Data Quality

Batches of samples analyzed for pathogens are checked for contamination via a negative control on nucleic acid extraction (tck_pathogenqa_pub.controlType='extractionNegative'), a negative control for master mix (tck_pathogenqa_pub.controlType='mixedSampleNegative'), a negative control for master mix (tck_pathogenqa_pub.controlType='masterMixNegative'), and a known positive control (tck_pathogenqa_pub.controlType='positive'). Details on these control measures can be found in the external lab SOP.

5 REFERENCES

Falco, R. C., and D. Fish. 1992. A comparison of methods for sampling the deer tick, Ixodes dammini, in a Lyme disease endemic area. Experimental & Applied Acarology 14:165–173.

Ginsberg, H. S., and C. P. Ewing. 1989. Comparison of flagging, walking, trapping, and collecting from hosts as sampling methods for northern deer ticks, Ixodes dammini, and lone star ticks, Amblyomma americanum (Acari, Ixodidae). Experimental and Applied Acarology 7:313–322.

Milne, A. 1943. The comparison of sheep tick populations (Ixodes ricinus L). Annals of Applied Biology 30:240–250.

Rulison, E. L., I. Kuczaj, G. Pang, G. J. Hickling, J. I. Tsao, and H. S. Ginsberg. 2013. Flagging versus dragging as sampling methods for nymphal Ixodes scapularis (Acari: Ixodidae). Journal of Vector Ecology 38:163–167.