Biodiversity Informatics, 19, 2025, pp. 1-6

DERIVING BEST USE DATA FROM NEON FOR MOSQUITO RESEARCH APPLICATIONS: A PRACTICAL GUIDE WITH CODE

Amely M. Bauer[1*], Sara Paull[2], Robert P. Guralnick[3], Lindsay P. Campbell[1]

[1]*Florida Medical Entomology Laboratory, Department of Entomology and Nematology, IFAS, University of Florida, Vero Beach, FL, USA*
[2]*Battelle, National Ecological Observatory Network, Boulder, CO, USA*
[3]*Florida Museum of Natural History, Department of Natural History, University of Florida, Gainesville, FL, USA*

Abstract.—The National Ecological Observatory Network (NEON) is a long-term monitoring program at the continental scale designed to understand and forecast ecological responses to environmental change at local to broad scales. However, despite robust and nearly continuous collections, there are several challenges to deriving analysis-ready data sets from the available raw data that must be overcome to maximize the use of NEON mosquito collections. Here, we provide species-level estimated abundances for nighttime collected female mosquitoes derived from the Mosquitoes sampled from CO_2 traps. By including zero counts, our derived data complement existing data sets and provide an analysis-ready time series useful for investigating mosquito phenology, abundances, and diversity at the species or community level. We also outline a set of considerations specific to filtering NEON mosquito data by sex and for day or nighttime collections, highlighting factors that could introduce uncertainty to abundance estimates. Along with the data set, we provide an R Markdown file that includes annotated code and documents our data filtering and quality assurance and quality control (QA/QC) steps, as well as data files used to filter the mosquito data based on QA/QC criteria. All files are freely available for download through the Environmental Data Initiative data portal. Our reproducible and fully documented workflow can be easily adapted for specific needs or other NEON surveillance data. Our work aims to enhance the accessibility and use of NEON's rich, long-term monitoring data.

Key words.—Culicidae, vector abundance, mosquito diversity, population, spatiotemporal data, ecology of vector-borne disease, modeling

Mosquitoes are a highly diverse group, with >3700 species globally (Harbach, 2018). Multiple species are a substantial public and veterinary health concern because they are capable of transmitting disease-causing pathogens to humans and animals (Mullen & Durden, 2019). As ectotherms, mosquito abundances and phenology link closely to abiotic environmental conditions and can serve as a sentinel to global environmental change. In addition, mosquito species distributions are changing, with movements facilitated by greater global connectivity.

Mosquito monitoring is often conducted at local or regional levels to support understanding mosquito populations and potential disease risk. The National Ecological Observatory Network (NEON) is a long-term monitoring program at the continental scale and is designed to understand and forecast ecological responses to environmental change at local to broad scales (Kao et al., 2012). NEON began routine mosquito collections at a subset of sites in 2014, with all 47 core and terrestrial gradient sites collecting samples by 2019 (Hoekman et al., 2016). Sampling follows a standardized protocol to facilitate broad-scale analyses of e.g. mosquito abundances, diversity, and phenology (LeVan et al., 2022; Paull et al., 2024). NEON mosquito collections provide a unique opportunity to understand and predict drivers of mosquito population dynamics under a range of environmental conditions. Despite the raw data reflecting robust and nearly continuous collections, producing analysis-ready data sets is a multi-step process that requires careful review of the NEON documentation and considerable effort (Atkins et al., 2025). Providing a derived mosquito abundance data set will eliminate these steps, while substantially broadening the scope of analyses that can be performed with the data (Atkins et al., 2025; Nagy et al., 2021).

* Corresponding author: amelybauer92@gmail.com.

One challenge to maximizing the use of NEON mosquito collections is that the data are provided in a raw format after initial quality assurance and quality control (QA/QC) by NEON staff, so appropriate filtering and processing steps must be taken prior to conducting analyses. NEON and the broader community provide tutorials and two R packages to facilitate assembly of raw data files (Lunch, Laney, et al., 2024; Lunch, Sokol, et al., 2024; NEON 2025). However, implementing assembly steps still requires data wrangling skills. A second challenge is that the NEON raw data includes a free text field that contains comments about factors that can introduce uncertainty into abundance estimates during trapping, sorting, and taxonomic identification. These comments provide additional insights beyond the initial QA/QC check in the raw data tables, but require comprehensive assessment. The third challenge is that zero counts and species-level absences are not explicitly entered into the identification data tables and must be inferred from trap collection records. An absence of mosquito counts can reflect a lack of trapping effort or indicate that no mosquitoes were captured for an individual species during a trapping event. Inferring these values is essential for many downstream analyses, including occupancy models or other approaches that require either presence/absence or abundance-based counts. Producing a derived data set from the robust, continental-scale raw time-series data that NEON publishes facilitates increased use of the NEON mosquito collections data.

Here, we provide species-level estimated abundances and estimated mean number of mosquitoes per trap hour for nighttime collected female mosquitoes derived from the mosquitoes sampled from CO_2 traps (DP1.10043.001) 2024 data release (NEON, 2024). We focus on nighttime host-seeking female mosquitoes because they are the primary focus of mosquito control programs and offer the opportunity for broader data integration with monitoring and surveillance efforts outside of the NEON (Nagy et al., 2021), including with multiple digitized control program collections accessible through the VectorBase repository (VectorBase 2025). This data set complements an existing data set available through the `neonDivData` R package (Li et al., 2022), which provides standardized species counts adjusted for trapping effort for multiple NEON biological collections but does not include species-level zero counts for trap events (Li et al., 2022; O'Brien et al., 2021). We also outline a set of considerations or "best practices" specific to filtering NEON mosquito data by sex and for day or nighttime collections and highlight additional factors that could introduce uncertainty to abundance estimates during trapping, sorting, and taxonomic identification. Along with the data set, we provide a corresponding R Markdown file with annotated code and two .csv files that we used to exclude records that did not meet our QA/QC criteria for full reproducibility. The aim of this work is to provide an analysis-ready data set that can be used for a variety of different studies of nighttime collected female mosquitoes, as well as an adaptable workflow to process NEON biodiversity surveillance data.

NEON Mosquito Collections

NEON mosquito collections are distributed across 20 terrestrial core and 27 terrestrial gradient sites and trapped using CO_2-baited Centers for Disease Control (CDC) light traps with no light attractant. Typically, mosquitoes are collected at ten plots within each NEON site, and traps are set at a minimum distance of 310 m from one another. To account for seasonal activity patterns at NEON sites where mosquitoes are absent during winter months when temperatures are low, trap collections follow a "field season" and "off season" sampling protocol. During the field season, terrestrial core sites collect mosquitoes every two weeks and terrestrial gradient sites collect mosquitoes every four weeks. Following three consecutive trapping events with zero mosquitoes collected at terrestrial core sites, regular field season sampling within the NEON domain ends and the sampling protocol shifts to the off-season schedule. During the off-season, three traps instead of ten traps are sampled weekly at terrestrial core sites and no collections are made at the terrestrial gradient sites within the domain. The increase in trapping frequency during the off-season at fewer trap locations is designed to more precisely capture the return of mosquito flight activity following winter dormancy. Once mosquito activity returns at a terrestrial core site, the regular field season protocol resumes for both terrestrial core and gradient sites within the domain (LeVan et al., 2022). In line with NEON's mission as long-term monitoring program, sampling protocols for sample collection, sorting, and taxonomic identification steps have largely remained constant over time. Minor adjustments to protocols that can lead to small changes in the records between version releases are outlined in detail in the product release details[1] and documentation (Paull et al., 2024).

The mosquitoes sampled from CO_2-baited CDC light traps (DP1.10043.001) data product is available for download from the NEON data portal[2] (NEON, 2024). Functions in the `neonUtilties` and `neonOS` R packages

[1] https://www.neonscience.org/data-samples/data-management/data-revisions-releases/release-2024.

[2] https://data.neonscience.org/data-products/DP1.10043.001/RELEASE-2024.

can then be used to stack and join the downloaded files in preparation for data assembly (Lunch, Laney, et al., 2024; Lunch, Sokol, et al., 2024). The mosquito collection data are contained in comma-delimited files that must be joined, then filtered, and processed to estimate abundances of nighttime collected female mosquitoes (Paull & LeVan 2024). The files are organized in long format, with each row in the trapping and sorting files corresponding to a single event, and each row in the taxonomic file corresponding to species, genus, or family-level counts by sex. The files contain information about the location, sampling dates, trapping hours, day or nighttime collection, collection time, sex, species, counts, and proportion of the trap sample identified needed for data assembly. In addition, each file contains a `sampleCondition` field and a `remarks` field where NEON staff report whether the collections were compromised, damaged, or lost (Paull et al., 2024). The `remarks` field is a free text field that contains important information that can be used to evaluate uncertainty in the mosquito collection data, and we examined these remarks carefully to filter records to assemble the nighttime collected female mosquito data set.

Estimating Species-level Female Abundances

NEON has a specific approach for assessing mosquito abundances. For each trap event, expert taxonomists identify a representative subsample of up to ~200 mosquitoes. The identified subsample can then be used to estimate species-level mosquito abundances of the trap collection. The `proportionIdentified` field indicates the proportion of the total trap collection (based on sample weights) identified by the taxonomic laboratory (LeVan et al., 2022; Paull & LeVan 2024). We use this information to estimate the total number of mosquitoes for each species and sex, calculated as `individualCount` divided by `proportionIdentified`. This approach assumes that the counts of individual species by sex in the subsample scale proportionally to the full trap sample.

When assembling the data set for species-level nighttime collected female mosquitoes, an important consideration is the proportion of mosquitoes for which sex was not determined to ensure accurate representation of the female mosquito population in the subsample. As part of the taxonomic identification process, NEON taxonomists indicate whether a mosquito is a female, male, or is unidentified to sex (LeVan et al.; 2022). Here, we calculated the proportion of mosquitoes with sex identification for each trapping event, and kept only trap events where the sex of at least 90% of the mosquito subsample was determined. Similarly, the proportion of mosquitoes identified to the species-level is important for estimating individual species abundances in trap events. We decided to include only trapping events where at least 90% of the mosquito subsample was identified to the species level. We then aggregated subspecies records at the species level and removed all records that identified mosquitoes to the family or genus level. To filter the data, we implemented two specific variables in the provided R code. Changing the values of these variables allows users to easily adjust the resulting data set to their specific needs.

Filtering to Nighttime Collections

The NEON sampling protocol includes daytime and nighttime collections, generally with each trap event occurring at an individual plot during a 24-hour sampling bout (or 40 hours prior to 2018: Paull et al., 2024). Within the data set, the `nightOrDay` field indicates whether the trap was set for daytime or nighttime collection. Nighttime traps are set approximately at dusk and collected the following morning, near dawn. A four-hour time window is allowed surrounding trap setting and collection times to accommodate unexpected logistical constraints.

Nighttime collections accounted for approximately half of all trap events included in the NEON 2024 data release. However, we found that filtering trap events to nighttime collections identified by the `nightOrDay` field alone can exclude some trap events that are important for including zero counts. In order to retain the maximum number of trap events, we first filtered records that indicated nighttime collection events. Then, we filtered records with no indicator in this field but with zero trap hours, which accounted for 15% of all trap events in the 2024 data release. These records occur when sampling is considered “impractical” because the weather was too cold for mosquitoes to be active (i.e., a zero for mosquito abundance is appropriate), the site could not be reached, or some other factor precluded sampling at the scheduled time. We evaluated the records in these events in a later step to determine whether to include a zero count because sampling was impractical due to cold weather. These steps can be easily updated using the code provided in the corresponding R markdown file if users are interested in filtering for daytime collections.

Considerations for Including or Excluding a Trap Event

The NEON mosquito collection data contain multiple QA/QC fields that provide useful information about the trap event, sample sorting, and taxonomic identification. Multiple factors can introduce uncertainty into abundance estimates across these categories. We found that considering information within designated QA/QC fields, as well

as free text comments from NEON staff members included in the `remarks` columns, was critical for determining whether to include or exclude a trap event. For example, in addition to mosquito sampling, a subset of NEON mosquito collections are tested for pathogens, which requires that a strict cold chain is maintained during sample transport, sorting, and taxonomic identification. Here, we found that filtering records based on `sampleCondition` alone can eliminate trap events that were compromised for pathogen testing but not necessarily compromised for abundance estimation. In addition, in some cases, the `sampleCondition` field does not indicate a compromised trap, but the free-text field providing comments from NEON staff in the `remarks` columns contained information showing that uncertainty could be introduced to abundance estimates.

When considering factors that could introduce uncertainty to abundance estimates, we focused on general conditions. First, we checked QA/QC fields containing information about the status of the trap fan, CO_2 sublimation, and catch cup condition, and we eliminated trap events where these factors were compromised. Next, we examined the `remarks` fields for trapping, sorting, and taxonomic identification. General criteria for excluding a trap event included traps that were tipped over and on the ground, the presence of ants or other predators in the sample, mosquitoes frozen to the sides of catch cups, sample spills where it was unclear if all mosquitoes were recovered, or issues affecting taxonomic identification. Because these remarks are free text comments entered by NEON staff across three categories of the sample collection process, we exported unique remarks to a separate .csv file and then examined them using the open-source software program OpenRefine. We used faceting, clustering, and text search functions within OpenRefine v3.7.9 (Delpeuch et al., 2024) to facilitate examination of unique remarks, and we indicated whether the trap event should be kept or excluded in a `keepEvent` field. We then joined the `keepEvent` field back to the NEON data to narrow collection records based on our QA/QC criteria for estimating abundances.

Using this approach, we removed 7% of trap events from the active nighttime mosquito collections data set. For ~40% of the excluded events, `sampleCondition` indicated no known compromise during sample collection, sorting, and identification, but our comprehensive evaluation of the free text and trap status fields provided more detailed information that we determined could affect confidence in estimated abundances. We also identified that ~40% of the active trap events that we retained during this step did not include a compromise status, meaning that these records would be eliminated during a filtering step based on the `sampleCondition` fields alone. The result is a QA/QC data set that is comprehensive to estimating species level nighttime collected female mosquito abundances.

Including Zero Counts

One contribution of the NEON mosquito collection data is relatively consistent and standardized sampling, which allows for the inclusion of zero counts when no mosquitoes are collected or when an individual species is not captured in a trap. The raw data downloaded from NEON does not explicitly incorporate zero counts in the taxonomic data tables, but these values can be inferred from the trap collections data (e.g., by checking the trap hours and values of the `targetTaxaPresent` columns). Here, we added zero counts for individual trap events and for individual species within each trap using a set of criteria. We considered two factors to determine whether to include a zero for a trap event. First, we identified whether a trap was active, indicated by the `trapHours` field containing a value greater than zero. If the trap was active but no female mosquitoes were collected, we included a zero for the count for the trap event. Then, we examined trap events where the trap was inactive, indicated by the `trapHours` field equaling zero. For these events, we observed the `samplingImpractical` and `remarks` columns and added zero counts for traps that were inactive because of low temperature conditions or snow cover preventing access to trap locations. The `samplingImpractical` field was introduced to the NEON data in September of 2019 to better indicate when planned sampling did not take place (Paull et al., 2024). For records prior to 2020, we relied on information provided in the remarks field alone. Combined, these considerations allowed us to include almost 18,000 zero count events that make up ~40% of all trap events in the final data set. We also included a zero count for individual species when a trap was set but the species was not collected.

The NEON mosquito data includes a field `nativeStatusCode` indicating native and non-native species, which can be used as an indicator of whether a species may have moved into a region after a specific date. Rather than trying to determine the timing of the distribution of non-native species across NEON sites, we opted to include a zero count for a species in a trap event if the species was collected at the trap site at least once within the given sampling year. One limitation of this approach is that interannual variation in a species' presence may be incomplete because no record will be included if the species was not collected at least once during a trapping year. This approach could be updated in the future, for example using more specific information

about arrival dates of non-native species to a geographic area.

R Markdown File and "remarks_OR.csv" Files

Along with the final filtered and QA/QC-processed data set of species-level, nighttime collected female mosquitoes with corresponding zero counts, we provide an R Markdown file containing annotated code used to filter and conduct QA/QC steps specific to the inclusion/exclusion criteria outlined above. This file provides information for data access and download, as well as code required to reproduce the final data set made available here. The file also documents the general steps used to process the remarks field in OpenRefine, with screenshots illustrating where the facet, clustering, and text search functions are located within the software. In addition to this information, further minor QA/QC steps are outlined in the R Markdown documentation, including checks for duplicate trap events and collection records, including events where counts of a mosquito species occur more than once for a given event. The documentation includes a list of all used R packages, with package versions reported in the markdown file reference section. We also provide the two .csv files exported from OpenRefine that indicate whether a trap event should be included or excluded in the `keepEvent` field. The purpose of the markdown and .csv files is to provide a fully reproducible workflow for the final data set. Further, the code and steps used may be altered in future iterations to filter NEON data under different user defined criteria and these steps are easily adapted to derive abundance data sets from similar raw NEON data. The derived data set, the two .csv files containing processed remarks, as well as the R Markdown file in form of its executable .Rmd script and rendered .hmtl version are available for download through the Environmental Data Initiative data portal[3] (Bauer et al.,2025).

Conclusions

The derived species-level nighttime female abundance data set described here provides a comprehensive resource for examining broadscale mosquito population dynamics at NEON sites across the U.S. The inclusion of zero counts for trap events for which no mosquitoes are collected provides an analysis-ready time series useful for examining mosquito phenology, abundances, and diversity at the species or community level, and the reproducible workflow, including annotated coding steps, provides a blueprint for users to alter filtering steps for specific needs. We hope the derived data set and corresponding code will help to maximize accessibility and use of this rich, long-term monitoring data set.

[3] https://doi.org/10.6073/pasta/5bd343a4d15fe7ac70107b0c4a0171b7

Acknowledgments

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

Competing Interests

The authors have declared that no competing interests exist.

Literature Cited

Atkins, J. W., Aho, K. S., Chen, X., Elmore, A. J., Fiorella, R., Luo, W., Lombardozzi, D., Lunch, C., Manak, L., Pablo, L. X. de, Myers-Pigg, A. N., Record, S., Qiu, T., Reed, S., Ruddell, B., Strange, B., Torrens, C. L., Yule, K., & Richardson, A. D. (2025). Recommendations for developing, documenting, and distributing data products derived from NEON data. *Ecosphere*, 16(1), Article e70159. https://doi.org/10.1002/ecs2.70159

Bauer, A. M., Paull, S., Guralnick, R. P., & Campbell, L. P. (2025). Species-level estimated abundances and zero counts of nighttime collected female mosquitoes 2014 - 2022 (Derived from NEON Mosquitoes sampled from CO_2 traps (DP1.10043.001, RELEASE-2024)) (Version 2). Environmental Data Initiative.

Delpeuch, A., Morris, T., Huynh, D., Weblate, Mazzocchi, S., Jacky, Guidry, T., elebitzero, Stephens, O., Matsunami, I., Sproat, I., Santos, S., Larsson, A., allanaaa, kushthedude, Fauconnier, S., Mishra, E., Beaubien, A., Magdinier, M., . . . Chandra, L. (2024). *OpenRefine* (version 3.7.9). Zenodo. https://doi.org/10.5281/zenodo.10644943

Harbach, R. E. (2018). *Culicipedia: Species-group, genus-group and family-group names in Culicidae (Diptera)*. CABI Publishing.

Hoekman, D., Springer, Y. P., Gibson, C., Barker, C. M., Barrera, R., Blackmore, M. S., Bradshaw, W. E., Foley, D. H., Ginsberg, H. S., Hayden, M. H., Holzapfel, C. M., Juliano, S. A., Kramer, L. D., LaDeau, S. L., Livdahl, T. P., Moore, C. G., Nasci, R. S., Reisen, W. K., & Savage, H. M. (2016). Design for mosquito abundance, diversity, and phenology sampling within the National Ecological Observatory Network. *Ecosphere*, *7*(5), e01320. https://doi.org/10.1002/ecs2.1320

Kao, R. H., Gibson, C. M., Gallery, R. E., Meier, C. L., Barnett, D. T., Docherty, K. M., Blevins, K. K., Travers, P. D., Azuaje, E., Springer, Y. P., Thibault, K. M.,

McKenzie, V. J., Keller, M., Alves, L. F., Hinckley, E.-L. S., Parnell, J., & Schimel, D. (2012). NEON terrestrial field observations: designing continental-scale, standardized sampling. *Ecosphere*, *3*(12), 1–17. https://doi.org/10.1890/ES12-00196.1

Li, D., Record, S., Sokol, E. R., Bitters, M. E., Chen, M. Y., Chung, Y. A., Helmus, M. R., Jaimes, R., Jansen, L., Jarzyna, M. A., Just, M. G., LaMontagne, J. M., Melbourne, B. A., Moss, W., Norman, K. E. A., Parker, S. M., Robinson, N., Seyednasrollah, B., Smith, C., . . . Zarnetske, P. L. (2022). Standardized NEON organismal data for biodiversity research. *Ecosphere*, *13*(7), e4141. https://doi.org/10.1002/ecs2.4141

LeVan, K., Hoekman, D., & Gibson, C. M. (2022). *TOS Science Design for Mosquito Abundance, Diversity, and Phenology: NEON.DOC.000910vC*. National Ecological Observatory Network (NEON).

Lunch, C., Laney, C., Mietkiewicz, N., Sokol, E., Cawley, K., & National Ecological Observatory Network. (2024). *neonUtilities: Utilities for working with NEON data* (Version 2.4.2). https://CRAN.R-project.org/package=neonUtilities

Lunch, C., Sokol, E., Robinson, N., & National Ecological Observatory Network. (2024). *neonOS: Basic data wrangling for NEON observational data* (Version 1.1.0). https://CRAN.R-project.org/package=neonOS

Mullen, G. R., & Durden, L. A. (2019). *Medical and veterinary entomology* (Third edition). Academic Press.

Nagy, R. C., Balch, J. K., Bissell, E. K., Cattau, M. E., Glenn, N. F., Halpern, B. S., Ilangakoon, N., Johnson, B., Joseph, M. B., Marconi, S., O'Riordan, C., Sanovia, J., Swetnam, T. L., Travis, W. R., Wasser, L. A., Woolner, E., Zarnetske, P., Abdulrahim, M., Adler, J., . . . Zhu, K. (2021). Harnessing the NEON data revolution to advance open environmental science with a diverse and data-capable community. *Ecosphere*, 12(12), Article e03833. https://doi.org/10.1002/ecs2.3833

NEON (National Ecological Observatory Network). (2024). *Mosquitoes sampled from CO_2 Traps (DP1.10043.001)*. https://data.neonscience.org/data-products/DP1.10043.001/RELEASE-2024

NEON (National Ecological Observatory Network). (2025). *Resources: Educational resources*. https://www.neonscience.org/resources

O'Brien, M., Smith, C. A., Sokol, E. R., Gries, C., Lany, N., Record, S., & Castorani, M. C. (2021). ecocomDP: A flexible data design pattern for ecological community survey data. *Ecological Informatics*, *64*, 101374. https://doi.org/10.1016/j.ecoinf.2021.101374

Paull, S., & LeVan, K. (2024). *NEON User Guide to Mosquitoes sampled from CO_2 traps (DP1.10043.001) and Mosquito-borne pathogen status (DP1.10041.001)*, Version F. National Ecological Observatory Network (NEON).

Paull, S., LeVan, K., Tsao, K., Hoekman, D., & Springer, Y. P. (2024). *TOS Protocol and Procedure: MOS – Mosquito Sampling: NEON.DOC.014049vN*. NEON (National Ecological Observatory Network).