NEON USER GUIDE TO PLANT FOLIAR TRAITS (DP1.10026.001)

PREPARED BY	ORGANIZATION
Samantha Weintraub-Leff	TOS

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	1/22/2018	Initial Release
B	10/01/2020	Updates to reflect changes to sampling design, including shift from plot-based to site-based approach starting in 2019 and collection of crown polygon shapefiles starting in 2020. Also reflects integration of stable isotope data into this data product.
C	11/08/2021	Added information about aopCollectDate field, recording the mapping layer for crown polygon shapefiles, and a new dataQF option for chlorophyll-only samples.
D	04/11/2022	Updated language in section 4 Taxonomy addressing RTE species obfuscation in the data. Updated section 5.3 Data Revision with latest information regarding data release.
E	12/05/2022	Section 3.3.2, added more detail on remote sensing layers used to delineate crown polygons. Section 3.7, clarified start and end dates in shapefile table. Section 5.5, added thresholds used to flag macro and micro nutrient data. Figure 3, updated graphic to include all data entry applications. Minor text clarifications throughout
E.1	3/10/2023	Section 3.1, added reference to the new samplePosition field. Section 3.2, explained how sampling outside of peak green condition is captured in plantStatus. Section 5.1 Figure 3, updated the schematic for data entry applications
E.2	5/11/2023	Section 3.3.3, details the addition of a leaf cleaning step to the collection protocol starting in 2020. Section 5.5, explains that beacuse of this, the 'soil contamination' part of the outlier flag for Fe and Mn data is no longer used for samples collected from 2020 onward.

TABLE OF CONTENTS

LIST OF 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 measured height of a sunlit foliage sample, are considered the lowest level (Level 0). Raw data that have been quality checked via the steps detailed herein, as well as simple metrics that emerge from the raw data, are considered Level 1 data products.

The text herein provides a discussion of measurement theory and implementation, data product provenance, quality assurance and control methods used, and approximations and/or assumptions made during L1 data creation.

1.2 Scope

This document describes the steps needed to generate the Plant foliar traits (DP1.10026.001) data product, which includes field metadata from collection of sunlit foliage samples, leaf mass per area calculations, and a host of chemical measurements including carbon (C) and nitrogen (N) concentrations and stable isotopes, lignin, chlorophyll, and major/minor/trace element concentrations. As of October 2020, the Plant foliar stable isotopes (DP1.10053.001) data product is being published as part of Plant foliar traits (DP1.10026.001). This user guide also provides details relevant to the publication of the Plant foliar traits data product via the NEON data portal, with additional detail available in the files NEON Data Variables for Plant foliar traits (DP1.10026.001) (AD[07]) and NEON Categorical Codes for Plant foliar traits (DP1.10026.001) (AD[08]), 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: Canopy Foliage Sampling (AD[10]). The raw data that are processed in this document are detailed in the files NEON Raw Data Validation for Plant foliar physical and chemical properties, Level 0 (DP0.10026.001) (AD[04]), NEON Raw Data Validation for Carbon and nitrogen concentrations and stable isotopes in plants and soil (DP0.10103.001) (AD[05]), and NEON Raw Data Validation for Plant lignin concentrations (DP0.10031.001) (AD[06]), 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	NEON Raw Data Validation for Plant foliar physical and chemical properties, Level 0 (DP0.10026.001)
AD[05]	Available with data download	NEON Raw Data Validation for Carbon and nitrogen concentrations and stable isotopes in plants and soil (DP0.10103.001)
AD[06]	Available with data download	NEON Raw Data Validation for Plant lignin concentrations (DP0.10031.001)
AD[07]	Available with data download	NEON Data Variables for Plant foliar traits (DP1.10026.001)
AD[08]	Available with data download	NEON Categorical Codes for Plant foliar traits (DP1.10026.001)
AD[09]	NEON.DOC.000906	TOS Science Design for Terrestrial Biogeochemistry
AD[10]	NEON.DOC.001024	TOS Protocol and Procedure: Canopy Foliage Sampling
AD[11]	NEON.DOC.000987	TOS Protocol and Procedure: Measurement of Vegetation Structure
AD[12]	NEON.DOC.000008	NEON Acronym List
AD[13]	NEON.DOC.000243	NEON Glossary of Terms
AD[14]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[15]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[16]	Available on NEON data portal	NEON Ingest Conversion Language

2.2 Acronyms

Acronym	Definition
δ13C	delta 13C, the stable carbon isotope ratio (13C:12C) in a sample compared to a reference material, reported in parts per thousand
δ15N	delta 15N, the stable nitrogen isotope ratio (15N:14N) in a sample compared to a reference material, reported in parts per thousand
AOP	Airborne Observation Platform
C	Carbon
LMA	Leaf mass per area
MODIS	Moderate Resolution Imaging Spectroradiometer
N	Nitrogen

3 DATA PRODUCT DESCRIPTION

The Plant foliar traits (DP1.10026.001) data product provides physical (leaf mass per area, leaf water content), chemical (C, N, lignin, chlorophyll, and major/minor/trace element), and stable isotope ($\delta^{13}C$ and $\delta^{15}N$) data for plant foliage samples collected using TOS Protocol and Procedure: Canopy Foliage Sampling (AD[10]). Plant foliar sampling and analyses implement the guidelines and requirements described in TOS Science Design for Terrestrial Biogeochemistry (AD[09]). All physical, chemical and isotopic data are reported at the spatial resolution of a woody individual or a bulk herbaceous clip strip located within a NEON plot. The temporal resolution is that of a single collection date. Site-date availability for these data are provided on the NEON Data Portal landing page: https://data.neonscience.org/data-products/DP1.10026.001.

Whenever possible, foliar data are collected in conjunction with overflights of the NEON Airborne Observation Platform (AOP), which conducts remote sensing of ecosystem chemical and physical characteristics using hyperspectral, LiDAR, and high-resolution RGB measurements. Ground-based foliar data can be used to ground-reference and validate AOP measurements, and may be of utility to the community in refining algorithms to map canopy foliar traits. Additionally, foliar data may inform species and site-level estimates of foliar physio-chemical properties and how those change over time, which will have value independent of remote sensing observations.

During canopy foliage sampling, only sunlit vegetation is targeted for collection and analysis. A variety of techniques are used to accomplish this, from hand clippers and pole pruners in short-stature sites to line launchers, slingshots and unoccupied aerial systems in tall-stature sites. The sampling method is recorded in the field data.

Plant foliar physical, chemical and isotopic data can shed light onto key ecological processes including net primary productivity, nutrient uptake and allocation, rates of decomposition and herbivory, and responses to environment stress such as drought at the plot, site, and continental scales. They also provide essential data for understanding change in vegetation and canopy dynamics over time.

3.1 Spatial Sampling Design

Sunlit plant foliage is sampled at all terrestrial NEON sites, from a combination of plots found within the NEON tower airshed (Tower plots) as well as those distributed across the landscape (Distributed plots, Figure 1).

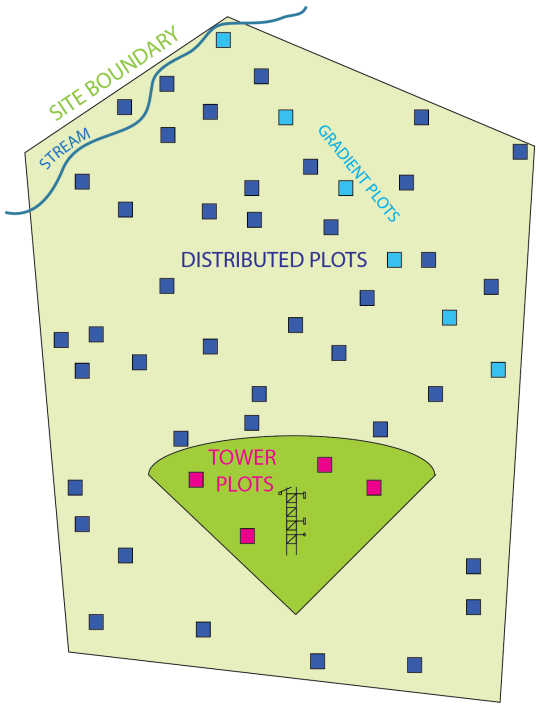

Figure 1: Representation of a NEON site with select Tower and Distributed plots shown

For forest and shrubland sites where the canopy is dominated by woody individuals, sampling is conducted according to a target taxa list containing all possible canopy species. Total sample numbers for the site are allocated according to canopy diversity, and replicates are assigned based on canopy rank abundance as measured in TOS Protocol and Procedure: Measurement of Vegetation Structure (AD[11]). Rare species are sampled but only if sunlit members can be found in NEON baseplots. Most sites have 30-40 plots in which to find high-quality sunlit samples, and TOS Protocol and Procedure: Canopy Foliage Sampling (AD[10]) gives guidance on spreading out samples and replicates to span plot types, stand age, and other site gradients.

As of the 2020 field season, in order to georeference sampled crowns for larger individuals, polygon shapefiles are mapped on top of recent AOP remote sensing data in the field while sampling, following Graves et al (2018). These shapefile polygons are delivered along with the trait data. Three data variables called **canopyStatus**, **canopyPosition** and **samplePosition** are also recorded in order to communicate the physical environment and vertical positions of sampled individuals. Additionally, georeferencing of stems is accomplished using the Mapping and Tagging procedure outlined in TOS Protocol and Procedure: Measurement of Vegetation Structure (AD[11]), which is important for woody individuals that are too small to create a polygon or fall in shadows in the imagery. Moreover, many trees sampled for foliar traits also have structure data available from that protocol.

For sites that are predominantly grassland, foliage is sampled from 20 assigned plots, four Tower and 10-16 Distributed. In each plot, all aboveground material in one or two (depending on plot size) *randomly selected* clip strips, which are nested within 0.5 x 3 m clip harvest areas or cells (Figure 2), are harvested, homogenized, and subsampled for physical, chemical, and isotopic analyses. Clip strips are usually 0.1 X

2 m, unless the site has patchy or scarce vegetation in which case a larger area can be sampled. Clip strip dimensions are recorded in the data, and precise geolocations can easily be calculated because the strips are oriented in a grid across the plot. Clip strip material is not sorted by species or functional group prior to subsampling and analysis, such that all traits (including LMA) are a community average of the plants present in that strip.

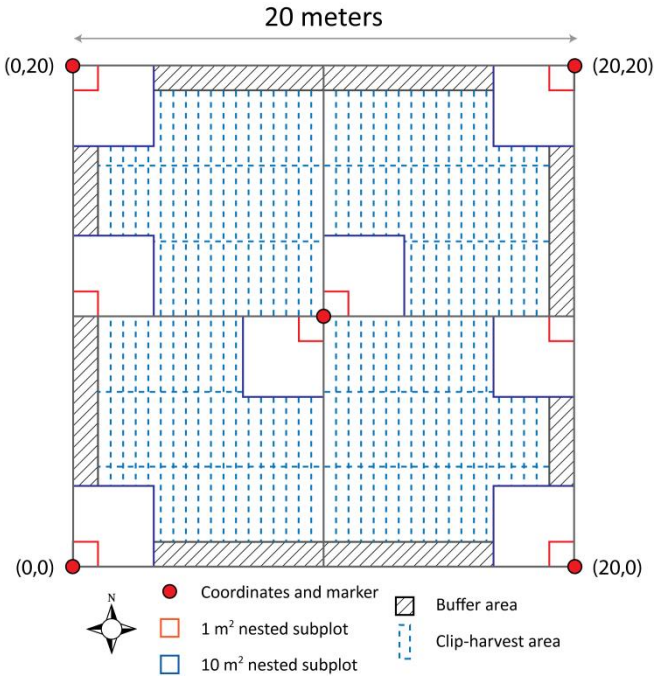

Figure 2: Representation of a NEON base plot showing the layout of clip harvest cells

In sites where a mix of woody individuals and herbaceous plants dominate the aerial cover, a combination of the two sampling approaches described above is employed. This equates to both sample collection of sunlit trees or shrubs based on a target taxa list as well as collecting up to 8 sunlit clip strips that span site bioclimatic gradients. In plots where there is a mix of cover types, clip strip selection is not entirely random as only those strips in sunlit areas are considered for harvest.

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/

3.2 Temporal Sampling Design

Foliar properties are measured at each site once every five years. During 'on' years, samples are collected during the period of historic peak greenness at the site, according to Moderate Resolution Imaging Spectroradiometer (MODIS) data. As much as possible, sampling occurs in coordination with the AOP over-

flight, which is also timed to coincide with peak greenness. The field **aopCollectDate** is included in the data to communicate the temporal links between ground and airborne data for each sample, so users can decide how to use trait data for modeling or validation depending on requirements of the research. Additionally, if the vegetation is in fact sampled outside of peak green condition (e.g., leaves not fully expanded, or leaves starting to senescence), this is recorded in the **plantStatus** field.

Upon collection of sunlit foliage, chlorophyll subsamples are immediately packaged in tinfoil, flash-frozen using dry ice, then maintained in the dark and shipped frozen to an external facility for pigment extraction within 14 days of collection. Additional processing of samples, including scans for leaf mass per area (LMA) measurement and initiation of sample drying for chemical and isotopic analyses, occurs within 1-5 days of collection. During this holding time, samples are held at 4°C in sealed plastic bags.

3.3 Sampling Design Changes

There have been two major design changes to canopy foliage sampling over time, related to the selection of woody individuals and how they are georeferenced. The rational for these changes, as well addition of a tissue cleaning step following foliage collection, are outlined below.

3.3.1 Choosing woody individuals

For sites with predominantly forest/shrubland cover, spatial collection methods from 2016-2018 involved sampling the top three dominant species present in each of 14 pre-assigned plots, 4 Tower and 10 Distributed. However, with input from the community including the Foliar Sampling Technical Working Group, we decided that this was not sufficient to capture the full spectrum of canopy biodiverity present at the site, which is needed if using foliage data in conjunction with the AOP. As such, starting in 2019 the design shifted away from plots and towards a site-based approach focused on sampling woody individuals as described above.

3.3.2 Georeferencing

For sites with predominantly forest/shrubland cover, georeferencing of sampled individuals collected from 2016-2019 was achieved only using the Mapping and Tagging procedure outlined in TOS Protocol and Procedure: Measurement of Vegetation Structure (AD[11]). However, feedback from the community suggested that this method often lead to ambiguity as to which crowns were sampled when trying to align with remote sensing data. As such, starting in 2020, in addition to the mapping and tagging procedure, polygon shapefiles of sampled crowns are created on top of recent AOP remote sensing data while in the field, as described above.

Starting in 2021, the type of remote sensing data used to create each crown polygon is indicated in the attribute table of the crown polygon shapefiles. This information is shared because there are small spatial offsets between the different layers. The choices are: CHM (canopy height model, derived from LiDAR); FCOL (false-color spectrometer), or CAM (high-resolution camera). In 2020 all crown mapping was done in reference to FCOL layers; in 2021 and onward, the CHM layer is used whenever possible since the LiDAR system provides the highest absolute geospatial accuracy. However, CHM is not useful in mapping crowns in all cases and sometimes other layers are used.

3.3.3 Foliage cleaning

Prior to 2020, leaf and needles samples were not cleaned following field collection, instead they were dried, ground and sent for processing 'as is'. After conferring with the Foliar Sampling TWG and external labs, NEON decided it would be prudent to clean foliar tissues prior to further processing and analyses. This step was added to the collection protocol for the 2020 field season and has been used ever since.

3.4 Theory of Measurements

LMA, a measure of plant investment in storage versus light capture, is estimated using the standard flatbed scanning method (Pérez-Harguindeguy et al. 2013). Briefly, a representative subsample of foliage (one to many leaves or needles) is scanned, then the scanned area is calculated using ImageJ (Schneider et al. 2012) and normalized by the dry mass of foliage scanned. In herbaceous samples, LMA is a community average of the dominant plants in a given clip strip. See AD[10] for more details on the procedure.

Chlorophyll a, b, and carotenoids are measured using 100 % methanol extractions, with absorbance read on a spectrophotometer and concentrations estimated using the coefficients reported in Lichtenthaler (1987). Chlorophyll data are reported in ug/ml but should be converted to more biologically relevant units prior to analyses, such as mg/g dry leaf tissue or mg/m2 leaf area using the equations below. Note that foliage dry mass fraction and LMA values are reported in the LMA table.

$$chlorophyll, mg/g\ dry\ leaf = extractChlAConc\ (ug/mL) * solventVolume\ (mL)\ /\ freshMass\ (g)\ /\ dryMassFraction\ (g\ dry\ leaf\ /\ g\ fresh\ leaf) * .001\ mg/ug$$

$$chlorophyll,\ mg/m2\ leaf\ area = chlorophyll,\ mg\ /\ g\ dry\ leaf * leafMassPerArea\ (g/m2)$$

Concentrations of lignin are determined using the acid detergent lignin method and defined operationally as the acid-insoluble foliar residue. Concentration estimates of both lignin and cellulose are provided from this procedure. Major, minor and trace elements are measured using nitric acid-peroxide digestions and inductively coupled plasma mass spectrometry (ICP-MS) analyses.

Concentrations and stable isotope ratios of carbon and nitrogen are measured simultaneously using elemental analysis coupled to isotope ratio mass spectrometry (EA-IRMS). In some cases, in order to get good N data, the CO_2 peak must be trapped, requiring two analytical runs to get both C and N values. This is reported in the data. Isotopes are measured as the abundance ratio of a heavy, rare isotope (H) to a light, more common isotope (L), normalized by those same ratios in a standard reference material.

$$\delta = [(R_{sample}/R_{standard} - 1)] \times 1000$$

where R = H/L. For all NEON stable isotopic data, $\delta^{15}N$ values are expressed on the atmospheric N_2 scale and $\delta^{13}C$ values are expressed on the Vienna Pee Dee Belemite scale.

NEON partners with external laboratories to conduct these chemical and stable isotope analyses. Standard operating procedures (SOPs) from these labs can be found in the NEON Data Portal document library

(http://data.neonscience.org/documents), in the External Lab Protocols > Terrestrial Biogeochemistry section. Some labs analyze a percentage of samples in duplicate in order to monitor internal consistency and repeatability. In these cases, NEON passes along replicate analyses in case the uncertainty information is of interest.

3.5 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the files NEON Raw Data Validation for Plant foliar physical and chemical properties, Level 0 (DP0.10026.001) (AD[04]), NEON Raw Data Validation for Carbon and nitrogen concentrations and stable isotopes in plants and soil (DP0.10103.001) (AD[05]), and NEON Raw Data Validation for Plant lignin concentrations (DP0.10031.001) (AD[06]). All variables reported in the published data (L1 data) are also provided separately in the file NEON Data Variables for Plant foliar traits (DP1.10026.001) (AD[07]).

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.

3.6 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a sample, taken from a woody individual or herbaceous clip strip, within a NEON plot.

sampleID (unique ID given to the sample) → **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 download 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.

In order to link ground-based foliar data to remote sensing observations, users will most likely wish to derive more precise estimates of sample locations. The method to achieve this depends on the the sample type.

When **sampleType** = *woody individual*, there are a couple of options depending on the year that samples were collected. For data collected in 2020 and onward, polygons are created in the field for each crown sampled (unless samples are too small or fall in 'dark pixels'/shadows). These polygons can be downloaded using the 'expanded' data package option, which provides a single shapefile containing all crown polygons per site-year bout. The shapefile information comes as a .zip file; all of the files contained in the

zip are likely needed to load crown polygons and extract their spatial information using a GIS. The year of AOP spectrometer data onto which the shapefiles were mapped is recorded in the cfc_shapefile table.

For data collected prior to 2020, or any case where crown polygons are not available or stem locations are needed, the mapping procedure using data from Woody vegetation structure (DP1.10098.001) should be used. Note that this procedure maps the stems, but many users may wish to delineate aerially visible crown areas. This can be done leveraging AOP imagery plus crown diameter and height measurements from DP1.10098.001 as available. There is at least one community code contribution on this theme and more are welcome, see the NEON code resources page for upload instructions and prior submissions.

If users select the 'expanded' package, the necessary data to map stems will come with the download package for Plant foliar traits. From there, the options are to:

- Filter the vst_mappingandtagging table to only the **individualIDs** that appear in the foliar data, then feed it into the getLocTOS() function from the geoNEON package, available here: https://github.com/NEONScience/NEON-geolocation/tree/master/geoNEON, or
- Follow these steps to perform the same calculation:

1. Add the variables **pointID**, **stemDistance**, and **stemAzimuth** from vst_mappingandtagging to each **individualID** that appears in cfc_fieldData using a match or join function.
2. Construct a pointID named location for each record in cfc_fieldData by concatenating the fields for namedLocation and pointID as follows: namedLocation + '.' + pointID, e.g. pointID '41' of namedLocation 'HARV_001.basePlot.cfc' has a pointID named location of 'HARV_001.basePlot.cfc.41'.
3. Use the API (http://data.neonscience.org/data-api; e.g. http://data.neonscience.org/api/v0/locations/HARV_001.basePlot.cfc.41) to query for easting ("locationUtmEasting"), northing ("locationUtmNorthing"), coordinateUncertainty ("Value for Coordinate uncertainty"), and utmZone ("locationUtmZone") for each pointID named location and use as inputs to the next step.
4. Calculate absolute position in UTMs of each woody individual stem using **stemAzimuth**, **stemDistance** and the easting and northing values derived in step 3 above, using equations (1) and (2):

$$Easting = easting.pointID + d * \sin\theta \quad (1)$$

and

$$Northing = northing.pointID + d * \cos\theta \quad (2)$$

where,

$$\theta = \frac{stemAzimuth * \pi}{180} \quad (3)$$

$$easting.pointID = \tag{4}$$

the easting value of the pointID named location,

$$northing.pointID = \tag{5}$$

the northing value of the pointID named location,

$$d = \tag{6}$$

stemDistance

5. Increase **coordinateUncertainty** by an appropriate amount (suggested 1 m) to account for error introduced by navigating around the plot. Keep in mind that geolocations are of the stem.

In rare cases samples may be from vines or lianas; this is allowed when vines are important in the canopy and atop dead trees. IndividualIDs from vine/liana samples do not have pointID, distance, or azimuth. Instead, users should identify the **supportingStemIndividualID** in the vst_mappingandtagging table, then use the mapping data of that supporting stem to geolocate the sample.

When **sampleType** = *herbaceous clip strip*, there are two options to derive a more precise location:

- Use the getLocTOS() function from the geoNEON package, available here: https://github.com/NEONScience/NEON-geolocation/tree/master/geoNEON, or
- Follow these steps to perform the same calculation:
    1. Construct the named location of the subplot of each record in cfc_fieldData by concatenating the fields for namedLocation and subplotID as: namedLocation + '.' + subplotID, e.g. subplotID '41' of namedLocation 'WOOD_002.basePlot.cfc' has a subplotID named location of 'WOOD_002.basePlot.cfc.41'.
    2. Use the API (http://data.neonscience.org/data-api; e.g. http://data.neonscience.org/api/v0/locations/WOOD_002.basePlot.cfc.41) to query for easting ("locationUtmEasting"), northing ("locationUtmNorthing"), coordinateUncertainty ("Value for Coordinate uncertainty"), and utmZone ("locationUtmZone") for each subplot named location as inputs to the next step.
    3. Use the clip cell lookup table, available here: http://data.neonscience.org/api/v0/documents/clipCellNumber_lookup (clicking on link will initiate download), to find the offsets for each clipCellNumber and subplot (note that **subplot** in cfc_fieldData is the same as pointID in the offset table).
    4. Use these offsets to adjust the easting and northing values downloaded in step 2, using equations (7) and (8):

$$Easting = easting.subplotID + eastingOffset \quad (7)$$

and

$$Northing = northing.subplotID + northingOffset \quad (8)$$

4. Increase coordinateUncertainty by an appropriate amount to account for error introduced by navigating within plots (suggested 0.5 m). Keep in mind that the calculated value is the center of the clip strip.

3.7 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the **collectDate**.

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. For code resources to work with these files, see Data Relationships (3.9).

StartDate and endDate in the cfc_shapefile table reflect the earliest and latest collect dates for foliar samples, not necessarily the earliest and latest create dates for crown polygons. While the two are often the same, they may differ for some bouts when polygons are created before or after actual foliage collection. The attribute table for each shapefile lists polygon create dates.

Field collection metadata and LMA measurements are scheduled to appear on the NEON data portal 90 days after sample collection, whereas chemistry measurements will appear 4-9 months following sample collection depending on external lab turnaround times. As such, site/month combinations will show 'available' well before the full suite of expected data have been returned.

3.8 Associated Data Streams

The vst_mappingandtagging table from the Woody plant vegetation structure (DP1.10098.001) data product comes in the expanded data download package in order to allow for calculation of precise geolocations of woody individual samples. This is especially relevant to pre-2020 data. The other tables from DP1.10098.001 can be downloaded separately in order to access relevant vegetation structure data for many of the foliar samples, including tree height, dbh, etc. Tables can be linked using the **individualID** variable.

In addition, the Plant presence and percent cover (DP1.10058.001) data product may be useful for understanding approximate species composition of the clip strip samples. It is possible to make plot-level estimates leveraging annual measurements of taxonID and cover in 1 m2 subplots nested within many of the same plots used for clip strip foliage sampling. Data can be aligned using the **plotID** variable. In arid sites, DP1.10058.001 may also be useful because much of the cover is bare ground and/or biological soil

crust. Cover estimates for these non-plant categories are recorded and may be helpful for pairing with foliage data and airborne observations.

Each field season, there is typically at least one site where leaf-level field spectra measurements are taken in conjunction with sunlit foliage sampling. These data are available as part of the Field spectral data (DP1.30012.001) data product.

Users may wish to link foliar trait measurements to other biogeochemical pools and fluxes measured during the same year (soil chemistry, root biomass and chemistry, etc). In some cases, measurements will be coincident at the plot level, and different measurement streams can be joined using the **plotID** variable. In other cases, data will not overlap in space and users may need to take site-level means and join using **siteID**, or use other spatially explicit approaches to align the data.

3.9 Product Instances

There is one foliar sampling bout per year at 8-11 sites, with 20-60 samples collected per site depending on cover type and site biodiversity. This will yield 160-660 unique foliar samples per year observatory-wide. Each sample is then subsampled for 6 unique physico-chemical trait measurements, yielding 960-3,960 unique data product instances in a given calendar year.

3.10 Data Relationships

TOS Protocol and Procedure: Canopy Foliage Sampling dictates that each woody individual or herbaceous clip strip collected yields a unique **sampleID** in the cfc_fieldData table. A record from cfc_fieldData is expected to have one child record in cfc_LMA and several child records in cfc_chemistrySubsampling. Each site-year combination where woody individuals are sampled starting in 2020 is expected to have one record in cfc_shapefile, and the shapefiles themselves are available as a zip file in the expanded package. Each record from cfc_chemistrySubsampling is expected to have one to several (if analytical replicates were conducted) child records in cfc_chlorophyll, cfc_carbonNitrogen, cfc_elements and cfc_lignin. 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.*

cfc_fieldData.csv - > One record expected per **sampleID**, generates a single **chlorophyllSampleID** used to measure chlorophyll and carotenoids

cfc_LMA.csv - > One record expected per **sampleID**, generates a single **lmaSampleID** used to measure LMA

cfc_chemistrySubsampling.csv - > One record expected per **sampleID**, generates a **cnSampleID** used to measure carbon and nitrogen concentrations and stable isotopes, a **ligninSampleID** used to measure lignin and elements, and (sample size permitting) an **archiveSampleID** for the NEON Biorepository.

cfc_shapefile.csv - > One record expected per **downloadFileName**, links to **shapefileID** in the field table and a zip file downloaded as part of the expanded package

cfc_chlorophyll.csv - > One record expected per **chlorophyllSampleID** x **analyticalRepNumber** combination, associated with chlorophyll and carotenoid data

cfc_carbonNitrogen.csv - > One record expected per **cnSampleID** x **analyticalRepNumber** x **co2Trapped** combination, associated with carbon and nitrogen concentration and stable isotope data

cfc_elements.csv - > One record expected per **ligninSampleID** x **analyticalRepNumber** combination, associated with major, minor and trace element data

cfc_lignin.csv - > One record expected per **ligninSampleID** x **analyticalRepNumber** combination, associated with lignin data

cfc_chlorophyllParameters.csv - > One record expected per **analyte** x **chlCarotEquationInput** x **chlCarotExtinctionCoefficient** x **laboratoryName** combination, parameters used to calculate concentrations from absorbance

cfc_chlorophyllSummary.csv - > One record expected per **analyte** x **sampleType** x **laboratoryName** x **qaReportingStartDate** combination, used to associate sample data with relevant uncertainty values.

bgc_CNiso_externalSummary.csv - > One record expected per **analyte** x **sampleType** x **laboratoryName** x **qaReportingStartDate** combination, used to associate sample data with relevant uncertainty values.

cfc_elementsSummary.csv - > One record expected per **analyte** x **sampleType** x **laboratoryName** x **qaReportingStartDate** combination, used to associate sample data with relevant uncertainty values.

lig_externalSummary.csv - > One record expected per **analyte** x **sampleType** x **laboratoryName** x **qaReportingStartDate** combination, used to associate sample data with relevant uncertainty values.

Sample identifiers and barcodes will be generated for each collection event and subsample type. Only samples collected for archive (**archiveSampleID** and **archiveSampleCode**) will be retained; in all other cases the physical sample will be discarded following measurement or analysis.

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

3.11 Special Considerations

3.11.1 Anytical Replicates

Before using the chemistry data, end users will likely wish to remove or average the information from analytical replicates, by taking 'mean' or 'first' of all instances that have the same sampleID. In addition, several variables used to track external laboratory procedures, such as raw wavelengths from pigment extractions and digest concentrations from elemental analyses, are available in the expanded package. These variables can be used to reconstruct the final, reported values for those users who are interested.

3.11.2 Sampling Impractical

For records collected in 2020 and beyond, the field **samplingImpractical** in cfc_fieldData is used to track planned but missed collection of herbaceous clip strip samples. Records that have anything other than 'OK' for the sampling impractial value will not have a sampleID but will show the plot and subplot assigned for sampling and why it was not collected. Woody individual sampling is not tracked in this way since there are no pre-assigned plots.

3.11.3 Sample drying issue, C-N data

Due to a miscommunication, prior to 2020-09-08 foliage samples analyzed for carbon (C) and nitrogen (N) concentrations and stable isotopes were not re-dried prior to weighing and analysis at the external lab. All NEON foliage samples are dried at 65C in the domain labs, but they are sometimes then stored in paper bags or coin envelopes for weeks to months before being ground, transferred to vials, and shipped. During this time they may accumulate moisture, especially in humid areas.

Subsequent testing revealed that %C data measured prior to 2020-09-08 are likely underestimated by 1.5-2.5% due to this lack of re-drying prior to analysis. As foliage samples tend to have high %C (30% - 55%), this bias may have only minor impacts on many analyses, but is something for users to keep in mind. For the other parameters (%N, C:N, d15N, d13C), testing suggests there were no detectable differences between re-dried samples and originals. All affected records have been flagged, see section 5.4 for more details. All samples collected after 2020-09-08 are re-dried prior to external analysis.

4 TAXONOMY

NEON manages taxonomic entries by maintaining a master taxonomy list based on a community standard, if one exists. Through the master taxonomy list, synonyms submitted in the data are converted to the appropriate name in use by the standard. The master taxonomy for plants is the USDA PLANTS Database (USDA, NRCS. 2014. https://plants.usda.gov). Taxon ID codes used to identify taxonomic concepts in the NEON master taxonomy list are alpha-numeric codes, 4-6 characters in length based on the accepted scientific name. Each code is composed of the first two letters of the genus, followed by the first two letters of the species and first letter of the terminal infraspecific name (if applicable), then if needed, a tiebreaking number to address duplicate codes. Genus and family symbols are the first five (genus) or six (family) letters of the name, plus tiebreaking number (if needed). Symbols were first used in the Soil Conservation Service's National List of Scientific Plant Names (NLSPN) and have been perpetuated in the PLANTS system. The portions of the PLANTS Database included in the NEON plant master taxonomy list includes native and naturalized plants present in NEON observatory sampling area including the Lower 48 U.S. States, Alaska, Hawaii, and Puerto Rico. NEON plans to keep the taxonomy updated in accordance with USDA PLANTS Database starting in 2020 and annually thereafter.

In the course of data collection for Plant foliar traits, threatened and endangered (T&E) species may occasionally be sampled. To avoid publishing locations of sensitive species, identifications of T&E species are 'fuzzed', i.e., reported at a higher taxonomic rank than the raw data. Moreover, as sample identifiers contain taxonomic information (to aid technicians with field and laboratory workflows), all sample identifiers are obfuscated upon data product publication to the NEON data portal. Sample identifiers should thus be

treated like barcodes, containing no human-readable information about the samples but being unique tags that allow for consistent sample linkage across tables.

Prior to the 2022 data release, publication of species identifications were obfuscated to a higher taxonomic rank when the taxon was found to be listed as threatened, endangered, or sensitive at the state level where the observation was recorded. The state-level obfuscation routine was removed from the data publication process at all locations excluding sites located in D01 and D20, and data have been reprocessed to remove the obfuscation of state-listed taxa for all years. Federally listed threatened and endangered or sensitive species remain obfuscated at all sites and sensitive species remain redacted at National Park sites.

The full master taxonomy lists are available on the NEON Data Portal for browsing and download: http://data.neonscience.org/static/taxon.html.

5 DATA QUALITY

5.1 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. A schematic of the data entry application design for field collection of foilar samples is depicted in Figure 3. An additional set of constraints are implemented during the process of ingest into the NEON database.

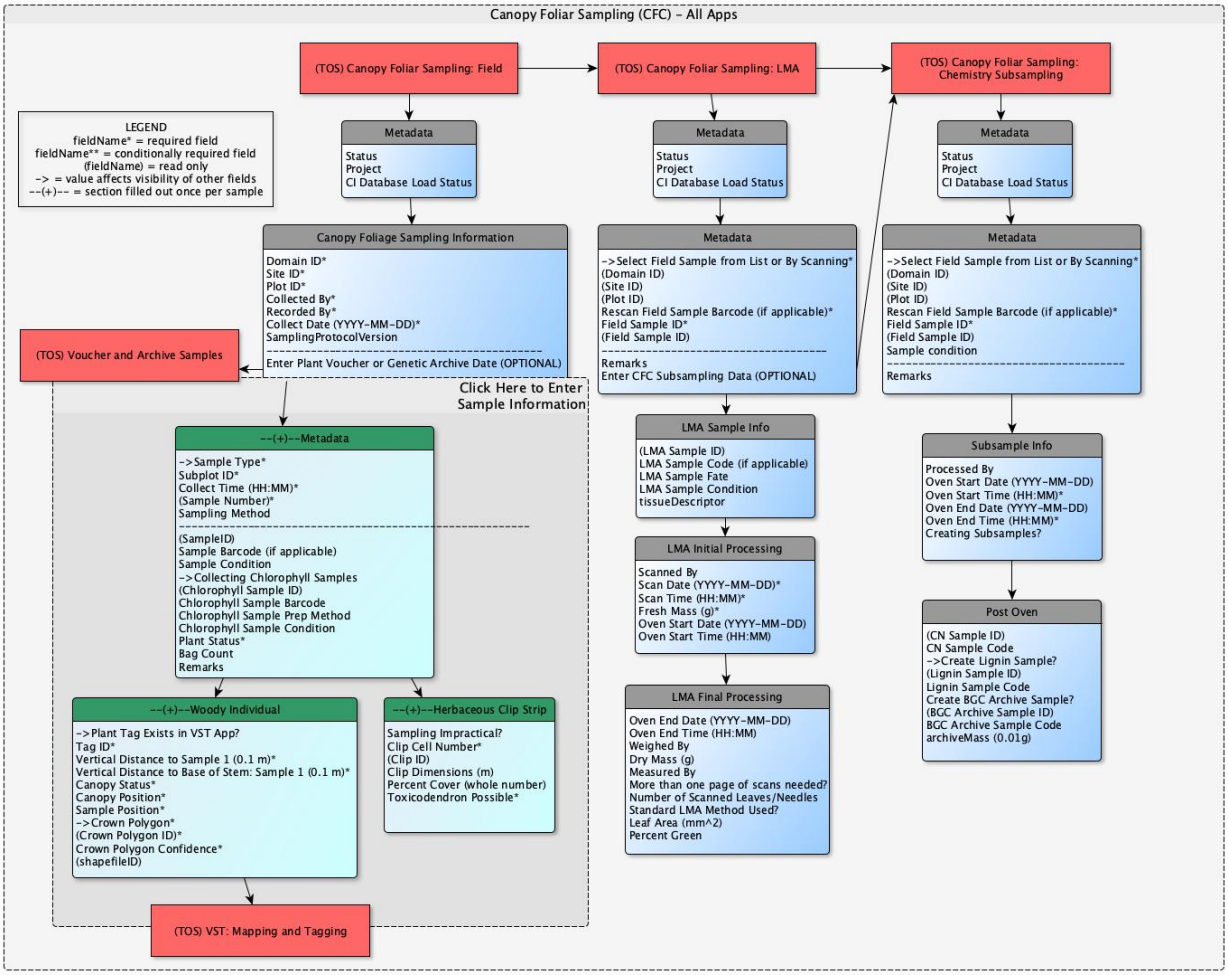

Figure 3: Schematic of the applications used by field technicians to enter foliar sampling data

The product-specific data constraint and validation requirements built into data entry applications and database ingest are described in the documents NEON Raw Data Validation for Plant foliar physical and chemical properties, Level 0 (DP0.10026.001) (RD[04]), NEON Raw Data Validation for Carbon and nitrogen concentrations and stable isotopes in plants and soil (DP0.10103.001) (RD[05]), and NEON Raw Data Validation for Plant lignin concentrations (DP0.10031.001) (RD[06]) provided with every download of this data product. Contained within these files are fields named 'entryValidationRulesForm' and 'entryValidationRulesParser', which describe syntactically the validation rules for each field built into the data entry application and spreadsheet uploader, respectively. 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[15]).

Note that field 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. Moreover, external laboratory data returned during this time were also not subject to same full suite of quality

controls.

5.2 Automated Data Processing Steps

Following data entry into a mobile application or web user interface, the steps used to process the data through to publication on the NEON Data Portal are detailed in the NEON Algorithm Theoretical Basis Document: OS Generic Transitions (AD[14]).

5.3 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.

5.4 Quality Flagging

The **dataQF** field in each record is a catch-all quality flag for known issues that apply to a record. Values are added by NEON Science upon data review. For Plant foliar traits, the list of dataQF values is given in the table below.

The entry 'deprecatedMethod' refers to chlorophyll and carotenoid data from 2016 that was analyzed after a very long sample hold time (several months). From 2017 and beyond, samples are analyzed within two weeks of collection. Accordingly, pigment data from 2016 have been flagged. The entry 'dryingProtocolError' refers to the drying issue for C-N analyses described in the *Special Considerations* section above. The entry 'chlorophyllOnlySample' refers to field samples that were collected for pigment analyses only because the pigment sub-sample of the initial foliar sample was destroyed prior to analysis. In order to relate the chlorophyll data for those individuals to other trait measurements, users will need to join on tagID or individualID plus year, and should beware that the collect dates and exact sampled branch differ.

tableName	fieldName	value	definition
chlorophyll	dataQF	deprecatedMethod	Data generated using deprecated/legacy workflows; overly long hold time for pigment samples
carbonNitrogen	dataQF	dryingProtocolError	Samples were not re-dried prior to external lab analysis, weight percent C values are likely underestimated by 1.5-2.5 percent
fieldData	dataQF	chlorophyllOnlySample	Material collected for pigment analysis only

Additionally, several 'condition' type fields have been added to the field and lab processing tables over time in order to communicate anomalous sample conditions or method deviations. Definitions for the categorical codes used for these fields are included in the file NEON Categorical Codes for Plant foliar

traits (DP1.10026.001) (AD[08]), provided in the download package for this data product. Fields have been added over time and entries may be missing in older data.

Records of land management activities, disturbances, and other incidents of ecological note that may have a potential impact are found in the Site Management and Event Reporting data product (DP1.10111.001).

5.5 Analytical Facility Data Quality

Analytical labs that generate plant foliar trait data include standards or secondary reference materials run as unknowns alongside NEON samples to gauge run acceptability. Labs communicate batch-level issues with accuracy of check-standards or secondary reference materials, as well as record-level issues with samples or measurements, using a suite of quality flags. Definitons for the categorical codes used for these QF fields are included in the file NEON Categorical Codes for Plant foliar traits (DP1.10026.001) (AD[08]). Fields have been added over time and entries may be missing in older data.

In addition, for foliar macro and micro nutrients, which are reported in cfc_elements, all data are checked against thresholds established using the TRY plant trait database (https://www.try-db.org/TryWeb/Home.php). TRY data were downloaded in 2020 and are visualized in Figure 4.

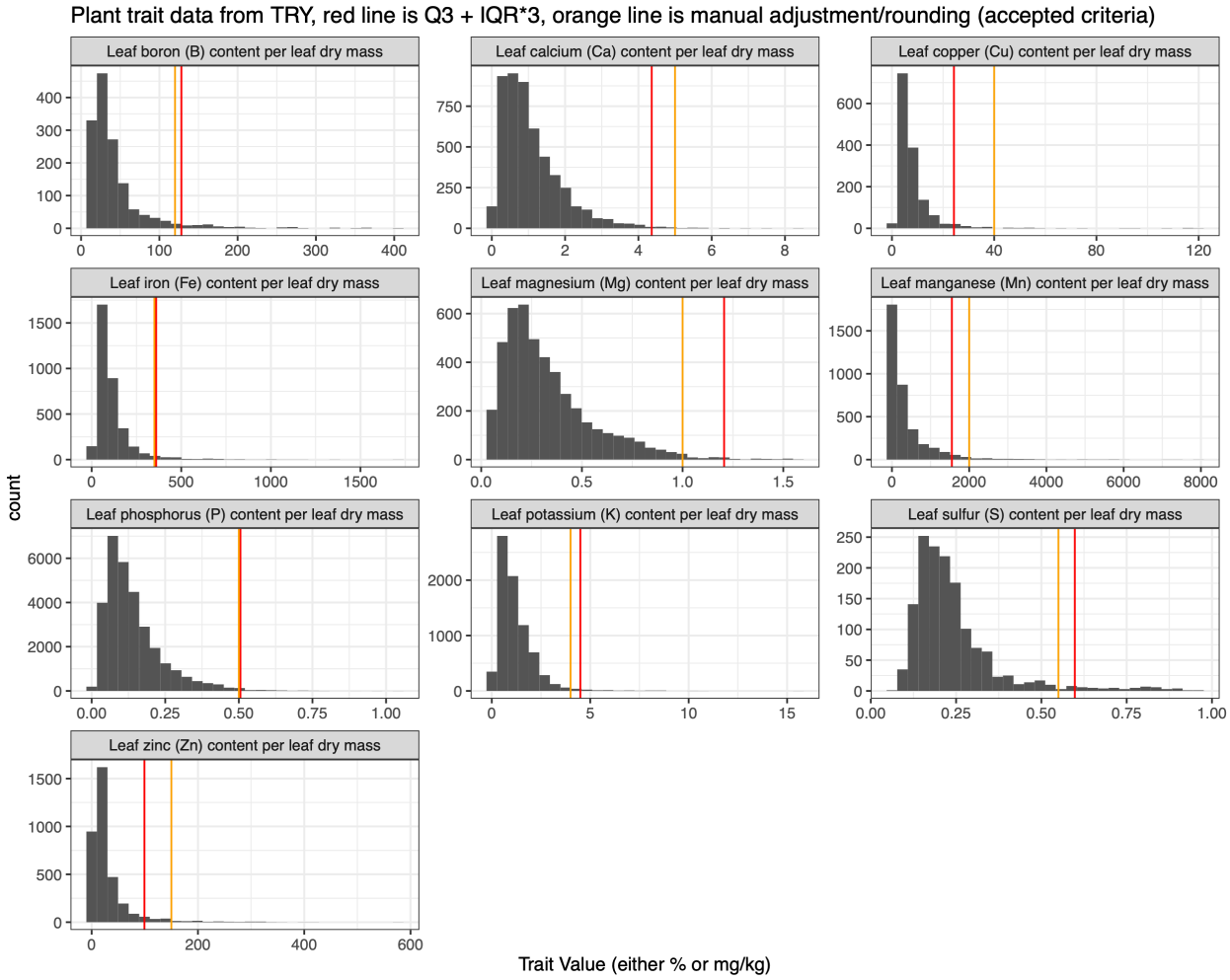

Figure 4: Plant trait data from TRY. Q3 = third quartile, IQR = interquartile range.

The accepted threshold values are listed in the table below along with the number of observations for each element in the TRY data. If a NEON sample values exceed these thresholds, the analytical lab must carefully review the data, and either re-run the digest solution and/or re-digest the plant tissue, depending on what the investigation reveals and how much material is available. If values are confirmed, data will be reported with a flag indicating 'sample value outlier' in the element-specific QF field.

For Fe and Mn, NEON had initially included an additional note about 'soil contamination' in the outlier flag. Those elements can be high in plant tissue samples that are dirty or dusty, and as detailed in Section 3.3, leaf cleaning was not included in the NEON protocol prior to 2020. However, once the tissue cleaning step was added, high Fe and Mn values are much less likely due to soil contamination. Accordingly, this component of the flag is no longer used, and has been removed from 2020 and onward records where it formerly existed. The 'soil contamination' flag appears in 2020 and later data in RELEASE-2021, RELEASE-2022, and RELEASE-2023; it is not present in these data in subsequent releases and provisional data, though it is still present in 2019 and earlier records.

Element	Criteria	N observations from TRY
P	0.5%	30,703
Ca	5%	5,040
Mg	1%	4,551
K	4%	7,692
S	0.55%	1,455
Mn	2,000 mg per kg	3,610
Fe	350 mg per kg	3,441
Zn	150 mg per kg	3,529
B	120 mg per kg	1,445
Cu	40 mg per kg	1,440

Data users may choose to keep or ignore the flagged macro and micro nutrient values, depending on goals of their research and any independent validation or outlier tests they may wish to perform. At present, the cfc_elements table is the only one that uses this routine for data flagging, but it is possible that NEON will adopt this approach for additional foliar trait data tables or analytes in the future.

Long-term analytical precision and accuracy of analytical laboratory check-standards or secondary reference material analyses are reported to allow users to interpret and analyze foliar chemistry and stable isotope data in the context of their uncertainty ranges. The data tables cfc_chlorophyllSummary, bgc_CNiso_externalSummary, cfc_elementsSummary, and lig_externalSummary, which are available in the data product expanded package, contain the long-term precision and accuracy of lab analyses.

For further information about individual laboratory QA procedures, refer to the lab-specific SOPs found in the NEON Data Portal document library (http://data.neonscience.org/documents), External Lab Protocols > Terrestrial Biogeochemistry section.

6 REFERENCES

Graves, S., J. Gearhart, T.T. Caughlin, and S. Bohlman (2018). A digital mapping method for linking high-resolution remote sensing images to individual tree crowns. PeerJ Preprints 6:e27182v1 https://doi.org/10.7287/peerj.preprints.27182v1.

Lichtenthaler, H.K. (1987) Chlorophylls and Carotenoids: Pigments of Photosynthetic Biomembranes. Methods in Enzymology 148: 350-382.

Pérez-Harguindeguy, N., S. Díaz , E. Garnier, S. Lavorel, H. Poorter, et al. (2013). New handbook for standardised measurement of plant functional traits worldwide. Australian Journal of Botany 61: 167-234.

Schneider, C, W. Rasband, and K. Eliceiri. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature methods 9: 671-675.

USDA, NRCS. 2014. The PLANTS Database (http://plants.usda.gov, 25 August 2014). National Plant Data Team, Greensboro, NC 27401-4901 USA.