NEON ALGORITHM THEORETICAL BASIS DOCUMENT (ATBD): TOS PLANT PHENOLOGY - QA/QC OF RAW FIELD AND LAB DATA

PREPARED BY	ORGANIZATION	DATE
Sarah Elmendorf	FSU	08/21/2014
Katie Jones	FSU	08/21/2014

APPROVALS (Name)	ORGANIZATION	APPROVAL DATE
Dave Tazik	SCI	08/06/2015
Hanne Buur	PSE	07/24/2015

RELEASED BY (Name)	ORGANIZATION	RELEASE DATE
Jennifer DeNicholas	CM	08/12/2015

See configuration management system for approval history.

The National Ecological Observatory Network is a project solely funded by the National Science Foundation and managed under cooperative agreement by NEON, Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Change Record

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
A	08/12/2015	ECO-02342	Initial release

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCIPTION

1.1 Purpose

This document details the algorithms used for creating a subset of NEON Level 1 data products that are the quality controlled products generated from raw Level 0 data, and associated metadata.

In the NEON data products framework, the raw data collected in the field are considered the lowest level (Level 0). Raw data that have been quality checked via the algorithms detailed herein, as well as simple metrics that emerge from the raw data, such as total species richness, are considered Level 1 data products. This document relates only to the former group of L1 data products, the quality controlled pass-through products from the Level 0 data products.

Plant phenology status and intensity observations are taken at all NEON sites, using a protocol derived from the USA National Phenology Network. This document details the processes necessary to QA/QC Level 0 plant phenology observations. It includes a detailed discussion of measurement theory and implementation, appropriate theoretical background, data product provenance, quality assurance and control methods used, approximations and/or assumptions made, and a detailed exposition of uncertainty resulting in a cumulative reported uncertainty for this product.

1.2 Scope

This document describes the theoretical background and entire algorithmic process for creating a subset of quality controlled and calibrated L1 data products from input data. These data products include plant phenology observations (NEON.DOM.SIT.DP1.10055), and later revisions will add derived products such as plant phenophase transition dates and plant phenometric summary statistics. It does not provide computational implementation details, except in cases where these stem directly from algorithmic choices explained here. This document also provides details relevant to the publication of the data products for NEON Data Publication Workbook for TOS Plant phenology observations [AD[12]) via the NEON data portal.

This document describes the algorithms for ingesting and performing automated quality assurance and control procedures on the plant phenology data collected in the field as described in Field and Lab Protocol: Plant phenology observations (AD[10]). The raw data that are processed in this document are detailed in the NEON Raw Data Ingest Workbook for TOS Plant Phenology observations (AD[11]).

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.005003	NEON Scientific Data Products Catalog
AD[03]	NEON.DOC.002652	NEON Level 1-3 Data Products Catalog
AD[04]	NEON.DOC.005005	NEON Level 0 Data Product Catalog
AD[05]	NEON.DOC.005011	NEON Coordinate Systems Specification
AD[06]	NEON.DOC.001928	NEON Raw Data Ingest Workbook for TOS Vegetation Structure
AD[07]	NEON.DOC.004309	NEON Field Site Information
AD[08]	NEON.DOC.00XXXX	NEON ATBD QA/QC Taxonomic Identification of Organisms – in prep
AD[09]	NEON.DOC.000907	TOS Science Design for Plant Phenology
AD[10]	NEON.DOC.014040	Field and Lab Protocol: Plant Phenology
AD[11]	NEON.DOC.001408	NEON Raw Data Ingest Workbook for TOS Plant Phenology observations
AD[12]	NEON.DOC. 001420	NEON Data Publication Workbook for TOS Plant phenology observations
AD[13]	NEON.DOC.002257	NEON Taxonomic Name List for Plants
AD[14]	NEON.DOC.002261	TOS Spatial Data
AD[15]	NEON.DOC.001403	NEON Raw Data Ingest Workbook for TOS Terrestrial Biogeochemistry: Chemistry of Soils and Plants
AD[16]	NEON.DOC.00XXX	List of Valid Identification Qualifiers

2.2 Reference Documents

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.000243	NEON Glossary of Terms

3 DATA PRODUCT DESCRIPTION

Plant phenology will be monitored at site- and season-specific sampling intervals by NEON field technicians at core and relocatable sites. The particular species and sampling schedules for plant phenological observations will vary between sites. For an overview of the data collection strategy, see the NEON TOS Science Design - Plant Phenology [AD 09]) as well as plant phenology protocols NEON Plant Phenology protocol [AD10]). Products derived from this sampling include the observed phenophase status (is the phenophase occurring?) for each individual plant monitored during each sampling bout, as well as the intensity of phenophases. According to the USA National Phenology Network, a phenophase is defined as 'an observable stage or phase in the annual life cycle of a plant or animal that can be defined by a start and end point'. Phenophase intensity, then, quantifies the degree to which a given phenophase is being expressed (continuous phenomenon, although recorded in bins). The phenophase categories employed by NEON are derived from the National Phenology Network (NPN) www.npn.org; although not all reproductive phenophases are monitored by NEON technicians due to labor constraints.

The majority of the phenology observations take place on tagged plants within a single square loop transect, situated within the tower airshed where possible (**Figure 1**).

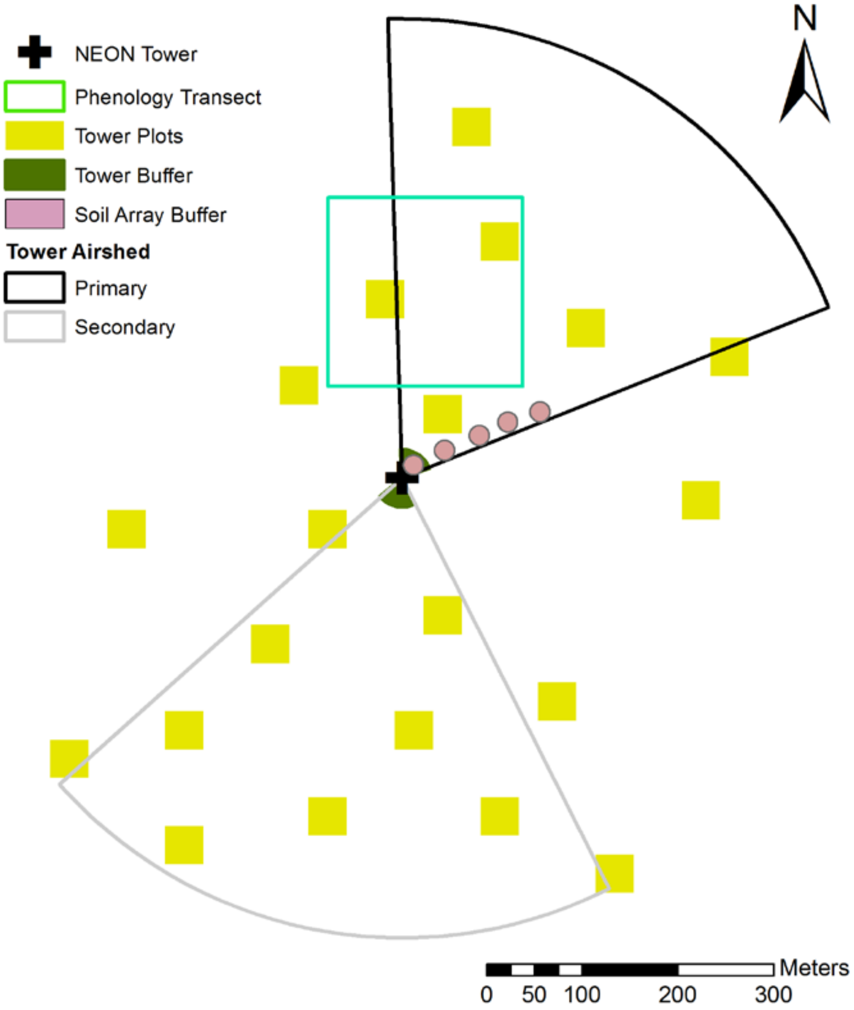

Figure 1. Spatial layout of phenology transect, overlaid with the tower airshed

Sampling occurs at multiple individuals located within 10m of the transect loop. Geolocations of 9 reference points along the transect are measured with high-resolution GPS units in the field (**Figure 2**). Geolocations of individually monitored plants are derived via measurements of relative position as described in this ATBD.

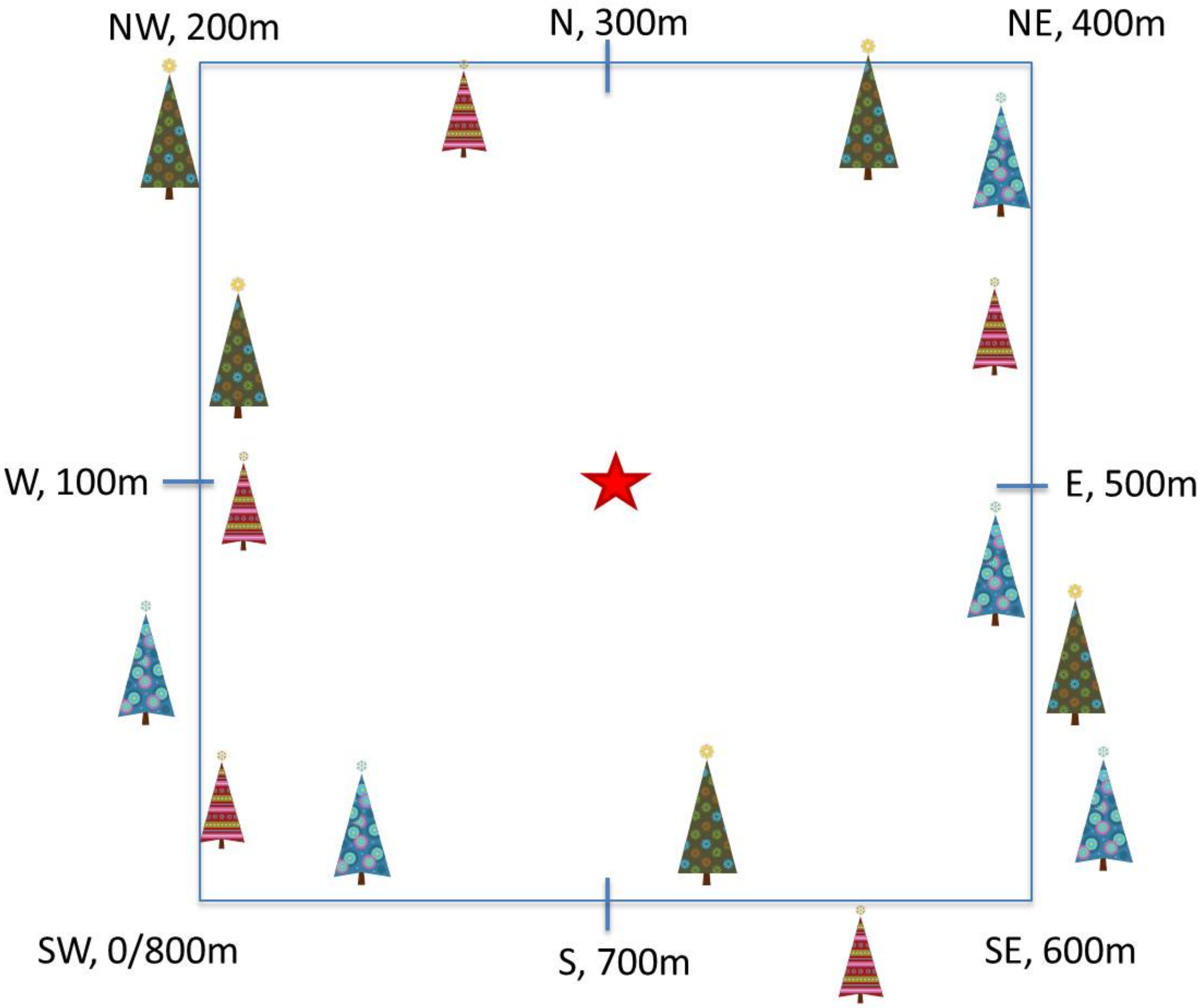

Figure 2. Zoomed in view of phenology transect layout, with transect meter measurements. Point-level geolocations are recorded at 8 reference points along the perimeter, plot-level geolocation at the plot centroid (red star).

3.1 Variables Reported

This ATBD describes the steps needed to generate the L1 data products: plant phenology observations (NEON.DOM.SIT.DP1.10055). Subproducts for this data product are listed below (**Error! Reference source not found.**). Detailed lists of the associated subproducts and metadata products are provided separately, along with example data in publication-ready spreadsheets in the NEON Data Publication Workbook for Plant Phenology Observations (AD[12]). 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), and with the Bird Monitoring Data Exchange standards (http://www.avianknowledge.net; accessed 16 February 2014), and the NPN, where applicable. Geospatial data shall conform to the standards set forth in the NEON Coordinate Systems Specification (AD[05]).

Table 1. List of subproducts produced in this ATBD for the data product, plant phenology observations (NEON.DOM.SIT.DP1.10055). The list is not exhaustive and a variety of supporting data will also be made available.

Number	Field Name	Description
	scientificName	Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
	canopyPosition	Vertical status of an individual relative to neighbors
	plantStatus	Physical status of individual: live, dead, lost
	stemDiameter	Cross-sectional stem diameter
	measurementHeight	Height at which stemDiameter is measured
	maxCanopyDiameter	Maximum canopy diameter of the individual or patch
	ninetyCanopyDiameter	Canopy diameter perpendicular to maxCanopyDiameter
	percentCover	Ocular estimate of cover of the index (e.g., species) as a percent
	height	Highest point of an individual or average height of a patch
	adultLeafLength	The length of an average adult leaf
	diseaseStatus	Disease status
	diseaseType	Specific disease present, if known
	duplicateIndividualYearQF	Quality flag for duplicate records - Per Individual Per Year
	missingRecordsIndividPerYearQF	Quality flag for missing records - Per Individual Per Year
	phenophaseStatus	Status of the phenophase: yes, no, uncertain or missed
	phenophaseIntensity	Intensity value range for the corresponding phenophase

	duplicateIndividualDayStatusQF	Quality flag for duplicate records - Individual Day Status Intensity

3.2 Temporal Resolution and Extent

The finest temporal resolution that plant phenology data will be tracked is at the level of a day. The intersampling period is variable, ranging from 3 days to several months, and is based on phenological activity and growing season duration at the site.

3.3 Spatial Resolution and Extent

The finest spatial resolution at which phenology data will be tracked is per individual (~90 individuals per primary phenology plotID). A single primary phenology plot exists at each site, which is placed in order to be representative of a primary flux footprint. Three sites, a core and two relocatables, are together representative of a domain. Every individual plant monitored will be tied to a specific location where it was collected. Overall, this results in a spatial hierarchy of:

spatial hierarchy = individualID (ID plant) (decimalLatitude, decimalLongitude) → *plotID* (ID of phenology plot within site) → *siteID* (ID of NEON site) → *domainID* (ID of a NEON domain)

3.4 Associated Data Streams

In sites where there is spatial overlap between tower plots for vegetation structure and phenology, the above data streams may be directly linked to measurements on the same individuals in the following data products: *Tree vegetation structure* (NEON.DOM.SIT.DP1.10098) and *Sapling and shrub vegetation structure* (NEON.DOM.SIT.DP1.10071), based on the shared individualID field that uniquely identifies tagged plants. In addition, some plant species are being monitored for foliar chemical properties, and the above data streams will thus be directly linked to archived, air-dried leaf tissue samples. Samples are again linked through individualID, and are available via NEON's collections; chemistry measurements for these species can be found in *Plant foliar chemical properties sampling* (NEON.DOM.SIT.DP1.10051). Finally, plants used during phenophase sampling may also be used to assess alternate methods of assaying the seasonal trajectory of canopy development in the environs of NEON towers. Plants arrayed along the primary phenology transect (identified by subTypeSpecification='primary') are selected to be in an area representative of the flux footprint; where this transect does not include plants within the phenocam view, additional plants within the phenocam view (plotTypeSpecification='phenocam') can be used to link to NEON's phenocam-derived Phenology images (NEON.DOM.SIT.DP1.00009) data products

3.5 Product Instances

Plants will be monitored at all core and relocatable sites. Some species (e.g., clonal species) are monitored by patch, rather than by individual. Algorithmic processing for species monitored by patch rather than by individual is identical, since in the case of species monitored by patch the data are recorded for the patch in aggregate, not for individuals within patches. For the purposes of this document, an ‘individual’ plant may refer to an individual or a patch. The number of individual records will therefore be ~90 individuals (or patches) observed for phenophase status and intensity on a regular rotation during growing season. Annual measurements of size and disease are taken once per year, basic information pertaining to taxonomic identification and spatial location is taken once per plant, although may be updated where the original taxonomic determination is uncertain.

4 SCIENTIFIC CONTEXT

4.1 Theory of Measurement/Observation

Phenology is defined as the seasonal timing of life cycle events. The IPCC (Solomon 2007) notes that it is one of the simplest processes for tracking changes in species' ecology in response to climate change. According to a recent synthesis, the onset of spring phenological events is advancing at a mean rate of 2.3 days per decade worldwide, likely due recent climate warming (Parmesan and Yohe 2003). Plants flower on average ~5 days earlier per °C increase in spring temperature (Wolkovich, Cook et al. 2012), so substantial changes in spring phenology are expected over the life of the observatory.

In addition to being a variable that is sensitive to climate change, phenology is also a potentially important driver of ecological responses to climate change, ranging from the demographic trajectories of individual taxa to changes in biogeographical distribution and ecosystem processes. For example, species whose phenologies track climate variability tend to have improved growth, productivity, or reproductive success in contrast to those that do not (Cleland, Allen et al. 2012) On the other hand, phenologic advancement in response to a warm spring temperatures followed by a late frost can have catastrophic effects on fruit and seed production and canopy development (Inouye 2008, Hufkens, Friedl et al. 2012). In either case, a population's phenological sensitivity may be an early indicator of its demographic trajectory. These sensitivities may constrain broad-scale distribution patterns; phenology appears to be a key biological constraint in process-based species distributions models (Chuine 2010).

4.2 Theory of Algorithm

This document describes the algorithms for assessing the integrity of the L0 data stream generated by the field sampling of plant phenology. The approaches described herein are simple yet necessary components of quality control and quality assurance, including: defining the range of possible values for each variable, specifying the lookup table(s) that contains the accepted values for particular variables, including linking variables, tracking the individual-level data for consistency and accuracy through time; and rectifying and/or flagging conflicting data where possible.

For the phenology entry form, technicians are able to view previously entered data for a given individual (phe_perIndividual form) or individual*day of year combination (phe_statusIntensity, phe_perYear forms). This quality control feature will help accidental duplicate entry of data. It also allows technicians to update (submit a 2nd set of observations) for select fields. The algorithm below rectifies these potential duplicate entries, according to the logic of data entry. For most fields, we assume that that the latest information (by date of observation and/or transactionDate – the date at which the data was entered) is correct.

Geographic coordinates are calculated based on relative positions as measured in the field. Relative positions of plants are recorded as the distance along the transect (traveling in a clockwise direction, beginning in the Southwest corner), as well as the offset direction (left or right, facing clockwise) and distance. Technicians identify relative positions of plants along the primary phenology transect by 3 pieces of information, transectMeter (the linear distance traveled along the transect, in a clockwise direction, starting from the southwest corner), and the ninetyDegreeDistance (offset of the plant from the transect, measured perpendicular to the transect to the left or right (directionFromTransect), as the technician travels clockwise). Primary phenology transects are always square, with the sides aligned in the 4 cardinal directions. The transect is 800 meters long, with reference points marked with high-resolution GPS ever 100m (at the SW, NW, NE, and SE corners, as well as the midpoints of the S,E,N,and W sides) (see Figure 2). Using this information, the technician's real position in geographic space can be calculated as the weighted mean of the 2 reference points bounding the technician's position along the transect (weighed by how close the reported location is to each reference point), plus an adjustment N,E,S, or W from the transect, based on the technicians reported ninetyDegreeDistance and directionFromTransect.

4.2.1 Summary of Algorithm for All data

1. Remove records for species of conservation concern

4.2.2 Summary of Algorithm for the statusIntensity data

1. Resolve duplicate records by selecting last entry in each 24 hour period
2. Update **taxonID**, and **scientificName** based on the latest information collected in the perIndividual data.
3. Generate **dayOfYear** from ISO dates

4.2.3 Summary of Algorithm for the perIndividual data

1. Resolve duplicate entries for the same individual by updating to retain only the last entry of taxonomic information and geolocation as necessary
2. Assign **L1:taxonID**, **scientificName** values using the reported **L0:acceptedTaxonID** and the taxonomic name list for plants (AD[13]).
3. Convert relative positions to geographic coordinates
4. Calculate **dropDate** for plants that are no longer monitored

4.2.4 Summary of Algorithm for perIndividualperYear data

1. Resolve duplicate records by selecting last non-NULL values of measurements in each calendar year:
2. Check for missing records

5 ALGORITHM IMPLEMENTATION

**Throughout the algorithm implementation section of this ATBD, 'nodata', 'NULL', and/or 'NA' indicates a blank cell; unless otherwise specified in the algorithm below, all variables that appear in tables phe_perindividual_pub, phe_perindividualperyear_pub, phe_statusintensity_pub, and phe_definitions_pub can be passed directly from the L0 or CI datastore to the L1 variable with the same name.*

Processing frequency:

*Automated processing should occur for all subproducts contained in phe_perindividual and phe_statusintensity on the 1st of each month, to include all dates up to t-1 month.

*Automated processing should occur for all subproducts contained in phe_perindividualperyear on the 1st of February each year, to include all dates Jan1- Dec 31 of the previous year.

Automated Processing Steps for Validation of Field Collected Data

5.1 Run the following processing steps for NEON Raw Data Ingest Workbook for TOS Plant Phenology observations (*AD[11]:phe_perindividual*)

1. Resolve duplicate entries for the same individual by updating the existing L1 data (adding a new readout) in order to retaining only the last entries of sets of location and taxonomic information
    - i. Create quality flag field **duplicateIndividualIDQF** and populate with zeroes
    - ii. Check for duplicate records over all previous entries (in date range 2012:present), based on exact matches in **individualID**
    - iii. If only one record exists
        - A. pass (or retain, if already transitioned) that record to L1 data and set **duplicateIndividualIDQF** =0.
        - B. Continue to step 2.
    - iv. If >1 records exist:
        - A. Create a single L1 *phe_perindividual_pub* record retaining:
            - AA. The earliest non-NULL value in *phe_perindividual_in:***addDate**
            - BB. The earliest (by addDate) non-NULL value in *phe_perindividual_in:***samplingProtocol**
            - CC. The latest entry (based date; in case of ties use transaction date) of *phe_perindividual_in:***transectMeter, directionFromTransect, ninetyDegreeDistance, growthForm**, and the *phe_perinidividual_in:***recordedBy, measuredBy** values associated with that date

            i. If updates to transectMeter, directionFromTransect, ninetyDegreeDistance have occurred, update **decimalLatitude** ,**decimalLongitude**, **elevation, elevationUncertainty and coordinateUncertainty and geodeticDatum** using the instructions in step (3) below, using the updated inputs from step (CC)

        DD. The latest non-NULL entry for **identifiedDate (**in case of ties use latest transactionDate) of phe_perindividual_in, and the corresponding values of **acceptedTaxonID, identificationQualifier, identifiedBy,** associated with that date.

            i. If updates to **acceptedTaxonID** have occurred, also update phe_perindividual_pub:**scientificName** , **taxonRank** and **taxonID** values based on matching the phe_perindividual_in:**acceptedTaxonID** from step DD to **scientificName** and **taxonRank** values in the taxonomic name list for plants (AD[13]). Assign phe_perindividual_pub:**taxonID** based on phe_perindividual_in:**acceptedTaxonID**.

        EE. Retain last entry (by transactionDate) **remarks**

        FF. Set **duplicateIndividualIDQF** =1

For non-duplicate (first entry of each individualID) records:

2. Assign *phe_perindividual_pub*: **scientificName** and **taxonRank** values based on matching the *phe_perindividual_in*:**acceptedTaxonID** to **scientificName** and **taxonRank** values in the taxonomic name list for plants (AD[13]). Assign *phe_perindividual_pub*: **taxonID** based on *phe_perindividual_in:***acceptedTaxonID.**

3. For each record (i) in *phe_perindividual_in* , generate **decimalLatitude** ,**decimalLongitude**, **elevation, elevationUncertainty and coordinateUncertainty and geodeticDatum** in *phe_perindividual_pub*, according to the following rules (see figure 2 for diagram of reference points and offset measurements).
    a. Determine subTypeSpecification as AD[14]: subTypeSpecification where AD[14]:plotID = plotID(i)
        i. If subTypeSpecification ≠ 'primary' or if subTypeSpecification is NULL
            A. *phe_perindividual_pub:* **decimalLatitude** ,**decimalLongitude**, **elevation, elevationUncertainty and coordinateUncertainty** = NULL
            B. Skip to step 4 to calculate dropDate.
        ii. Else:
            A. If any values in { *phe_perindividual_in:***transectMeter, directionFromTransect,** or **ninetyDegreeDistance**} are NULL:
                AA. **decimalLatitude** ,**decimalLongitude**, **and coordinateUncertainty** = NULL

BB. skip to step HH: generate **elevation** and **elevationUncertainty**

B. Else: calculate the geocoordinates of each plant in UTMs, based on offsets (in meters) from the 8 georeferenced points along the transect (Fig 2):

AA. Assign referencePoint_A, referencePoint_B, and **offsetDirection** for each record as in the following table:

*phe_perindividual_in:***transectMeter**	referencePoint_A	referencePoint_B
0 < **transectMeter** <= 100	SW	W
100 < **transectMeter** <= 200	W	NW
200 < **transectMeter** <= 300	NW	N
300 < **transectMeter** <= 400	N	NE
400 < **transectMeter** <= 500	NE	E
500 < **transectMeter** <= 600	E	SE
600 < **transectMeter** <= 700	SE	S
700 < **transectMeter** <= 800	S	SW

BB. Assign offsetDirection for each record as in the following table:

*phe_perindividual_in:***transectMeter**	*phe_perindividual_in:* **directionFromTransect**	offsetDirection
0 < **transectMeter** <= 200	Left	West
	Right	East
200 < **transectMeter** <= 400	Left	North
	Right	South
400 < **transectMeter** <= 600	Left	East
	Right	West
600 < **transectMeter** <= 800	Left	South
	Right	North

CC. Determine the geolocation of the **transectMeter** for each individual in UTMs as the weighted mean of the northing (in UTMs) of reference_pointA and reference_pointB, and the weighted mean of the easting reference_pointA and reference_pointB, where the weighted mean is defined as:

$$\bar{x} = \sum_{i=1}^{n} w_i x_i$$

where x_i is the northing or easting of the reference points, and

$$w_b = \frac{transectMeter - transectMeter(referencePointA)}{100}$$

aand

$$w_a = 1 - w_b$$

DD. Calculate the geolocation of the actual plant by adding or subtracting the **ninetyDegreeDistance** from the geolocation of the **transectMeter**, according to the **offsetDirection** specified in the table above.

EE. Example: **transectMeter**=85, **ninetyDegreeDistance**=4, **directionFromTransect**=right. Technician is assumed to be 85% of the way between the Sw and W reference points of the plot. If the SW reference point is at (0,0) (easting and northing, in UTMs), and the W reference point is at (-5, 103), for example, the true position of the **transectMeter** is at easting = (0.15*0+0.85*-5)=-4.25 and northing (0.15*0 + 0.85*103)=87.55 Since the technician is between 0-200m on the transect, s/he is traveling north, so a **ninetyDegreeDistance** of 4 to the right is equivalent to an offset of 4m E. The position of the plant in UTMs is (87.55, -0.25).

FF. Convert geolocation to decimal degrees and insert into **decimalLatitude** and **decimalLongitude,** and **geodeticDatum** fields of *phe_perindividual_pub*

GG. Generate **coordinateUncertainty** as the maximum of the coordinate uncertainties of the reference points + 2m, to account for additional error introduced by stretching tables:

  aa. *phe_perindividual_pub:* **coordinateUncertainty** = (max(coordinateUncertainty(referencePointA, coordinateUncertainty(referencePointB))+2

HH. Generate **elevation** and **elevationUncertainty:**

  aa. **elevation** is similarly calculated as the weighted mean of the elevations of the two nearest reference points, with weights proportional to reported distance from the points.

  bb. **elevationUncertainty** is unknowable, in the absence of information on microtopography of the site, so is reported as NULL

iii. Exception handling:

  A. If any of the required 7 inputs for geolocation are NULL (northing and eastings of the 2 reference points, **transectMeter**, **directionFromTransect**, **ninetyDegreeDistance**) **decimalLatitude**, **decimalLongitude**, and **coordinateUncertainty** are not calculated

        B. If any of the required 3 inputs for **elevation** are NULL (elevations of the 2 reference points, **transectMeter**) the **elevation** is not calculated

4. Calculate **dropDate** for plants that are no longer monitored
    a. For each unique **individualID** (i)in *phe_perindividual_in*:
        i. If **dropPlant**='drop', **dropDate**==L0:addDate where **dropDate**==’drop’
        ii. Else: **dateDrop**(i) = NULL
    b. Update L1 **dropDate** into **dropDate** field of *phe_perindividual_pub* in each record where i is found

5.2 Run the following processing steps for NEON Raw Data Ingest Workbook for TOS Plant Phenology observations (AD[11]:*phe_statusintensity_in*)

1. Check for duplicate records
    a. Create quality flag **duplicateIndividualDayStatusQF** and populate with zeros
    b. Determine **dayOfYear** (in Mountain Standard Time UTC-07) of each record dates as the integer number of days since Dec 31 of the previous calendar year.
    c. Check for duplicate records, based on exact matches in (**year** component of date), **dayOfYear**, **individualID**, and **phenophaseName** (example: 2014, dayOfYear 132, individualID = NEON.D10.CPER.123345, openFlowers)
    d. If duplicate records are identified:
        i. If values in all remaining fields (e.g., **phenophaseStatus, phenophaseIntensity**) *except uid*, **enteredBy**, are also the same:
            A. Pass only one of the records into phe_statusintensity_pub
        ii. Else: If there are conflicting values in any of the other fields of these records,
            A. Pass only one of the records into phe_statusintensity_pub
            B. For each measurement in { **measuredBy, recordedBy, phenophaseName, phenophaseStatus, phenophaseIntensityDefinition, phenophaseIntensity, remarks}**
                a. Assign phe_statusintensity_pub {**measuredBy, recordedBy, phenophaseName, phenophaseStatus, phenophaseIntensityDefinition, phenophaseIntensity, remarks**) as the last (by **transactionDate**) entry for each year, dayOfYear, **individualID** and **phenophaseName** combination.

2. Assign **taxonID**, **scientificName**, **growthForm**, and **identificationQualifier** based on the information in the *phe_perindividual_pub*
    a. For each record of *phe_statusintensity_in*:
        i. SELECT **L1:taxonID, scientificName, growthForm** and **identificationQualifier** from *phe_perindividual_pub* WHERE *phe_perIndividual_pub*:**individualID**=*phe_statusintensity_pub*:**individualID**
        ii. Update existing L1s as necessary, in the case of updates to taxonID

5.3 Run the following processing steps for NEON Raw Data Ingest Workbook for TOS Plant Phenology observations (AD[11]:phe_perindividualperyear_in)

1. Check for duplicate records
    a. Create quality flag **duplicateIndividualYearQF** and populate with zeros
    b. Check for duplicate records, based on exact matches in (**year** and **individualID**)
    c. If duplicate records are identified:
        i. If values in all remaining fields (e.g., **height, diseaseStatus**) *except uid*, **enteredBy**, or **remarks** are also the same:
            A. Pass only one of the records into phe_perindividualperyear_pub
            B. Concatenate **remarks** from all entries in *phe_perinidividual_in:***remarks,** using '|' to delimit successive entries
            C. Enter a '1' into the corresponding **duplicateIndividualYearQF**
        ii. Else: If there are conflicting values in any of the other fields of these records,
            A. Pass only one of the records into phe_**perindividualperyear_pub**
            B. For each measurement {patchOrIndividual, canopyPositition, plantStatus, maxCanopyDiameter, ninetyCanopyDiameter, percentCover, height, adultLeafLength}
                a. Assign phe_**perindividualperyear_pub** {**patchOrIndividual**, **canopyPositition**, **plantStatus**, **maxCanopyDiameter**, **ninetyCanopyDiameter**, **percentCover**, **height**, **adultLeafLength**}: as the last (by date) non-NULL entry in each **calendarYear** for that **individualID**.
                b. In case of ties (two entries at same date, use last **transactionDate** within the same date)
            C. Assign phe_perindividualperyear_pub: **stemDiameter** as the last (by date) non-NULL entry in each **calendarYear** for that **individualID**. In case of ties (two entries at same date, use last **transactionDate** within the same date)
            D. Assign phe_perindividualperyear_pub: **measurementHeight** as the measurement height associated with the recordID in step C (above)
            E. Assign phe_perindividualperyear_pub: **diseaseStatus** as the last (by date) non-NULL entry in each calendar Year for that individualID. In case of ties (two entries at same date, use last transactionDate within the same date)
            F. Assign phe_perindividualperyear_pub: **diseaseType** as the diseaseType associated with the by **recordID** in step E (above)
            G. Concatenate **remarks** from all entries in *phe_perinidividual_in:***remarks** for that individualID in that **calendarYear,** using '|' to delimit successive entries

            H. Assign **date** as the date of the record ID with the last non-NULL value in each calendarYear for maxCanopyDiameter.
                a. If no non-NULL values for maxCanopyDiameter for that individualID*calendarYear combination exist, use the date of the recordID last non-NULL value of percentCover.
                b. If no non-NULL values for percentCover for that individualID*calendarYear exist, use the date of the recordID of the last entry for any measurement stream in patchOrIndividual.

            I. Enter a '1' into the corresponding **duplicateIndividualYearQF** to indicated duplicate entries of the same individual per year, rectified algorithmically.
    d. For non-duplicates: Enter a '0' into the corresponding **duplicateIndividualYearQF**

2. Check for missing records. Each individual plant with data in *phe_statusintensity_in* should have at least one record in *phe_perindividualperyear_in*:
    a. Generate quality flag **missingRecordsIndividPerYearQF** and populate with zeros
    b. For each year (y):
        i. For each siteID (s)
            A. Determine the set of unique individualIDs {i} sampled in *phe_statusintensity_in* in siteID(s) between January 1 and December 31.
            B. Determine the set of unique individualIDs {j} sampled in *phe_perindividualperyear_in* in siteID(s) between January 1 and December 31.
            C. If length (i)>length(j)
            AA. Create(i-j) new records in *phe_perindividualperyear_pub*
            BB. Assign values for **individualID** for records created in step AA. based on elements of {i} that are not contained in {j}
            CC. Populate the **domainID**, **siteID**, **plotID** for each record created in step AA according to information for these individuals contained in *phe_statusintensity_in*
            DD. Populate *phe_perindividualperyear_pub:*date for records created in step AA based on the median *phe_perindividualperyear_pub:*date for plants in subset {i}
            EE. If {i} contains no individuals, assign *phe_perindividualperyear_pub:*date as June 30, year =y.
            FF. Generate a unique ID (uid) for each record created in step AA.
            GG. Set **missingRecordsIndivPerYearQF** for all records created in step AA to '1' and add to remarks 'this recorded is generated by the system'.

5.4 Remove records for species of conservation concern from all tables

1. For each record in *phe_perIndividual_pub, phe_statusIntensity_pub, and phe_perIndividualPerYear_pub:*
    a. Use **L1:taxonID** and **domainID** to lookup the dXXNativeStatus for the relevant domain (e.g., d10NativeStatus where domainID=D10), based on AD[13]
    b. If dXXNativeStatus = 'STATE ' or 'FEDERAL':
        i. Remove record from *phe_perIndividual_pub, phe_statusIntensity_pub, and phe_perIndividualPerYear_pub*

6 UNCERTAINTY

6.1 Analysis of Uncertainty

Analysis and reported uncertainty of the phenophase status and intensity classes is uncertain at the time of writing, pending budgeting for resampling in the field. If NEON is unable to fund resampling, repeatability information may be inferred from other groups using the same protocols at a later date. Taxonomic uncertainty for species monitored for plant phenology is expected to be negligible since the tagged individuals will be identified by expert botanists, and individuals of uncertain taxonomy excluded from the selection process.

6.2 Reported Uncertainty

Accuracy and precision of location information is be recorded with the GPS at plot establishment; an additional 2m of potential spatial error is introduced in triangulating to plants off the transect and is included in the calculations above. Taxonomic uncertainty as self-reported by the technicians is noted by **identificationQualifier**.

7 VALIDATION AND VERIFICATION

7.1 Algorithm Validation

Information to be placed her upon guidance from IPT.

7.2 Data Product Validation

Information to be placed her upon guidance from IPT.

7.3 Data Product Verification

Information to be placed her upon guidance from IPT.

8 FUTURE MODIFICATIONS AND PLANS

1. Estimates of elevation based on NEON's AOP data are planned, when NEON's DEM is available.
2. PDA wish list:
   a. Include integrated workflow to phenology reference images.
   b. Keep running list of total number of sampled plants per bout so technicians can check whether they have completed the census.
   c. Add ability to view and/or edit previous census's records in the field
3. Guidelines for uncertainty calculations and reporting for all data products are currently lacking, and need to be addressed, where possible.
4. Processes for reporting fuzzed geolocations for species of conservation concern are under development.
5. Guidelines for releasing revised data products.
6. Future Tests:
   a. Additional modifications to QA/QC tests and uncertainty reporting will be made once (a) plans for assessing of field technician data quality are in place, and (b) a sufficient volume of data from multiple sites has been amassed to test the QA tests proposed here. The aim of such modifications is to implement the suite of tests currently planned by other continental phenology programs (e.g. USA-NPN and PEP-725), including, but not limited to cross-checking the number of entries, the order in which phenophases are observed to occur (aka 'sequence of stages'), the outlier observations pertaining to the dates on which phenophases begin and end and/or intensities, and potentially other quantitative metrics in the annual size measurements (Jurjovic *et al.* 2013).

7. Future derived data products:
   a. Add **taxonRank** to match as implemented assignation of taxonomy for other data products.
   b. Secondary phenology transect within the focal view of the tower phenocams is planned. Specific processing steps for geolocations of individuals within this transect will be added as protocol is developed to identify and locate those individuals.
   c. Additional derived data products will be generated in future releases of this ATBD. One such product is deriving phenophase transition dates for phenophase onset, based on the last observation date on which a phenophase status was 'no' and the first observation date on which a phenophase status was 'yes', as well as the reverse for phenophase end. The planned algorithm is modeled after a prototype dataset provided by the USA National Phenology Network (NPN). Each row of the data product dataset represents a series of consecutive "yes" phenophase status records, beginning with the first day of year on which a "yes" status was recorded for a given phenophase on a given individual (**phenophaseOnsetUpper**), and ending with the last day of year on which the

phenophaseStatus was still 'yes', with 0 records in between of status='no' (**phenophaseEndLower**). If a phenophase status of "no" was recorded prior to the first "yes" or after the last "yes", **dayofYear** values for the last no preceding first yes (**phenophaseOnsetLower**), and the next 'no' after the last 'yes' (**phenophaseEndUpper**), as well as the gap (in days) between the "yes" and "no" records are also given (**phenophaseOnsetSamplingInterval**); **phenophaseEndSamplingInterval**). There may be multiple series (rows) for a given phenophase on a given organism within the same year. Additional summary statistics on (1) whether there were multiple periods of activity throughout the year. From the derived transition dates, additional plausibility quality flags tagging records where the onset and end of phenophases occurred out of the expected order (e.g. example, Leaves appearing earlier than Young leaves) may also be derived.

9 BIBLIOGRAPHY

Chuine, I. 2010. Why does phenology drive species distribution? Philosophical Transactions of the Royal Society B: Biological Sciences 365:3149–3160.

Cleland, E. E., J. M. Allen, T. M. Crimmins, J. A. Dunne, S. Pau, S. E. Travers, E. S. Zavaleta, and E. M. Wolkovich. 2012. Phenological tracking enables positive species responses to climate change. Ecology 93:1765–1771.

Jurkovic, A., T. Hubner, E. Koch, W. Lipa, and H. Scheifinger. 2013. The Pan European Phenology Database PEP725 Data Content and Quality. 9th Data Management Workshop. El Escorial, Spain.

Parmesan, C., and G. Yohe. 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421:37–42.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller. 2007. IPCC Climate Change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Wolkovich, E. M., B. I. Cook, J. M. Allen, T. M. Crimmins, J. L. Betancourt, S. E. Travers, S. Pau, J. Regetz, T. J. Davies, and N. J. B. Kraft. 2012. Warming experiments underpredict plant phenological responses to climate change. Nature 485:494–497.

10 APPENDIX: DATA ENTRY VALIDATION VIA PDA & WEBUI

7.1 All Data: Run the following steps for all fields in the raw data ingest workbook for TOS plant phenology observations (AD[11]):

1. Provide filtered drop-down menus for location fields in the [phe_dataingest_2014] workbook:
2. **domainID** and **siteID**, using the TOS subplot-level spatial data lookup table (AD[14])
3. **plotID**, using the TOS subplot-level spatial data lookup table (AD[14]) and where subType = 'Phenology'
4. Verify that entered values are of the correct **dataType**
5. Verify that entered values are valid, as specified by **entryValidationRules**
6. Generate a unique ID (**uid**) for each record
7. Follow guidelines for fields in which no data have been entered, as specified in **noDataOutcomePDA** and **noDataOutcomeUI**:
    a. If **noDataOutcome** = fail:
        i. do not let user finalize record until a value is provided
        ii. warning message text, 'Please enter a value for [fieldname] to continue', unless an alternative is provided in **warningText**
    b. Else if **noDataOutcome** = warn:
        i. warn user that value(s) are missing prior to finalizing record, and return (once) to screen to correct data) but allow selection of 'OK' to continue without a value
    c. Else if **noDataOutcome** = pass:
        i. allow user to finalize record with no values in this field
8. **dateTime**: Time of observation is not reported in paper datasheets. Default value in webUI to 12:00:00 mountain standard time.
9. Follow guidelines for default values, as specified in **defaultValuePDA** and **defaultValueUI**, and **case**
10. Reviewing data: Require csvs for review to be downloaded prior to submission of phe_statusintensity and phe_perIndividualperYear forms. Data for review should be presented sorted by perIndividual:**transectMeter**.

7.2 phe_perindividual

1. User can select to 'add' (creates new sample), or 'edit' plants (create new post with updated information for previously added plants, default user view contains entry from last post)

Existing plants should be presented sorted by transectMeter.

2. individualID: must be of form NEON.PLA.DXX.SITE.######, with user entering setting only the ###### component of the string
    a. **individualIDs** are uneditable once selected
3. taxonID:
    a. lookup table: AD[13] NEON Taxonomic Name List for Plants
    b. filtering rules: select (unique(**taxonID, scientificName**)) WHERE
        i. **taxonID** is a contained in the list of phenology species (n element of **taxonID** in **species_growthform**
    c. **taxonID** and **acceptedTaxonID** are posted to PDR
    d. User can also elect (with additional click) to select based on full list of domain-specific plants (to accommodate accidental monitoring of the wrong species)
    e. If taxonID is edited, user must re-enter **identifiedDate**
4. Autopopulate **growthForm** based on **taxonID**
    a. lookup table: species-GrowthForm
5. If **taxonID** is manually selected (step 4d, above), user can manually set any **growthForm** based on the full list of **growthForms** in **species-growthForm**

7.3 phe_statusIntensity

1. Form opens upon selecting an **individualID** from the current active list of plants (all plants where **dropPlant** = 'active'
2. Provide identical entry options for date, **measuredBy, recordedBy, samplingProtocol, remarks**
3. growthForm specific prompts appear for phenophase status (i.e. all fields in phe_statusintensity_conditionalvalidationrules.csv where description = ('Status of the identified phenophase').
4. When technician checks 'yes' to a given phenophase, popup box to prompt entry of intensity appears (only if an **intensityQuestion** exists for that **phenophase*growthForm** combination.
5. Radio button to set all phenophase status to 'no' for a given plant is available to speed data entry.

7.4 phe_perindividualperyear

1. Form opens upon selecting an **individualID** from the current active list of plants (all plants where **dropPlant** = 'active'

1. If data for that **individualID**/date is already entered, default user view contains entry from last post in the current calendar year. User can edit and repost information.
2. Autopopulate **patchOrIndividual** if possible based on **growthForm**
3. **growthForm** specific prompts appear for **canopyPosition, stemDiameter, measurementHeight, maxCanopyDiameter, ninetyCanopyDiameter, percentCover**