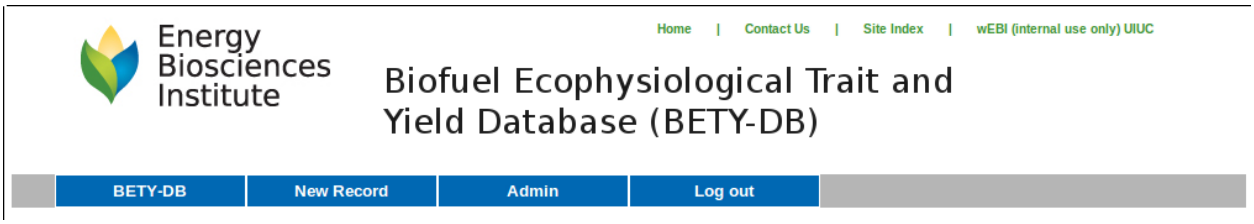
Data Entry Workflow for the Biofuel Ecophysiological Traits and Yields Database (BETYdb)

Andy Tu, David LeBauer

August 25, 2011

List of Figures

List of Tables



1 Overview

This is the userguide for entering data into the BETYdb database. The goal of this guide is to provide a consistent method of data entry that is transparent, reproducible, and well documented. The steps here generally accomplish one of two goals. The first goal is to provide data in a consistent framework that is associated with the experimental methods, species, site, and other factors associated with the original study. The second goal is to provide a record of all of the transformations, assumptions, and data extraction steps used to migrate data from the primary literature to the standardized framework of the database. This second goal not only supports the scientific value of the data itself, it also simplifies the Quality Assurance process.

2 Mendeley: Managing citations

Mendeley provides a central location for the collection, annotation, and tracking of the journal articles that we use. Features of Mendeley that are useful to us include:

- Collaborative annotation & notes sharing, see section 2.2
  - text highlighter
  - sticky notes for comments in the text
  - notes field for text notes in the reference documentation
- Read/unread & favorites:
  Papers can be marked as as **read** or **unread**, and may be **stared**.
- Groups: see section 2.1
- Tagging

2.1 Creating a new group in Mendeley (Project Managers)

Each project has two groups, "projectname" and "projectname_out" for the papers with data to be entered and the papers with data that has been entered. Papers in the _out group may contain data for future entry, for example, traits that are not listed in Table 6.

Each project manager may have one or more projects, each project should have one group. Group names should refer to plant species, plant functional types, or other another project specific name. A list of current groups can be found in Table 1. Please make sure that, at a minimum, Mike Dietze and David LeBauer are invited to join each project folder.

1. open Mendeley desktop
2. `Edit` → `Create Group` or `Ctrl+Shift+M`
3. create group name following instructions above
4. enter group name
5. set `Privacy Settings` → `Private`
6. click `Create Group`
7. click `Edit` → `Settings`
8. check File `Synchronization` → `Download attached files to group`

2.2 Adding and annotating papers (Project Managers)

The 'tag' field associated with each paper can be used to further separate papers, for example by species, or the type of data ('trait', 'yield', 'photosynthesis') that they contain. When naming a group, folders so that instructions for a technician would include the folder and the tag to look for, e.g. "please enter data from projectx" or "please enter data from papers tagged y from project x".

To access the full text and pdf of papers from off campus, use the UIUC VPN service.

If you are managing a Mendeley folder that undergraduates are actively entering data from, please plan to spend between 15 min and 1 hour per week maintaining it - enough to keep up with the work that the undergraduates are doing.

2.2.1 Adding a reference to Mendeley

- if the doi number is available (most articles since 2000)
  1. select project folder
  2. add entry manually
  3. paste DOI number in *DOI* field
  4. select the search spyglass icon
  5. drag and drop pdf onto the record.
- If doi not available:
  1. download the paper and save as `citation_key.pdf`
  2. add using the *Files* field
  3. the citation key should be in `authorYYYYabc` where `YYYY` is the four digit year and `abc` is the acronym for the first three words excluding articles (the, a, an), prepositions (on, in, from, for, to, etc...), and the conjunctions (for, and, nor, but, or, yet, so) with less than three letters.

2.2.2 Annotating a Reference in Mendeley

Each week, please identify and prepare papers that you would like to be entered next by completing the following steps:

1. Use the star label to identify the papers that you want the student to focus on next.

- Start by keeping a minimum of 2 and a maximum of 5 highlighted at once so that students can focus on the ones that you want. Students have been entering 1-3 papers per week, once we get closer to 3-5, the min/max should change.
- Choose papers the papers that are the most data rich.

2. For each paper, use comment bubbles, notes field, and highlighter to indicate:

    - name(s) of traits to be collected
    - methods:
        - site name
        - location
        - number of replicates
        - statistics to collect
        - identify treatment(s) and control
        - indicate if study was conducted in greenhouse, pot, or growth chamber
    - data to collect
        - identify figures number and the symbols to extract data from.
        - table number and columns with data to collect
    - covariates
    - management data (for yields)
    - units in 'to' and 'from' fields of Table 9 used to convert data
    - esoteric information that other scientists or technicians might not catch and that are not otherwise recorded in the database
    - any data that may be useful at a later date but that can be skipped for now.

Comment or Highlight

- sample size
- covariates (see Table 7)
- treatments
- managements
- other information entered into the database, e.g. experimental details

2.3 Finding a citation in Mendeley

To find a citation in Mendeley, go to the project folder. Group folders and personel are listed in Table 1. By default, data entry technicians should enter data from papers which have been indicated by a yellow star and in the order that they were added to the list. Information and data to be collected from paper can be found under the 'notes' tab and in highlighted sections of the paper.

3 Google Spreadsheets: Recording data transformations

Google Spreadsheets are used to keep a record of any data that is not entered directly from the original publication.

- any raw data that is not directly entered into the database but that is used to derive data or stats using equations in Table 9 and Table 5
- any data extracted from figures, along with the figure number
- any calculations that were made. These calculations should be included in the cells.

Each project has a google document spreadsheet with the title "project_data". In this spreadsheet, each reference should have a separate worksheet labeled with the citation key (`authorYYYabc` format). Do not enter data into excel first, this is prone to errors and information such as equations may be lost when uploading or copy-pasting.

4 Redmine: Reporting errors, suggesting features

4.1 Reporting errors in Redmine

4.2 Suggesting features in Redmine

5 BETYdb: Entering new data through the web interface

Before entering data, it is first necessary to (add and) select the citation that is the source of the data. It is also necessary for each data point to be associated with a Site, Treatment, and Species. Cultivar information is also required when available, but is only relevant for domesticated species. Fields with an asterix (*) are required.

Table 1: Current Projects

List of current projects, PI's, Managers, and Technicians.

Folders	Project	PI	Manager	Technicians	Status
arctic	Arctic	M. Dietze	C. Davidson	M. Azimi	active
prairie	Prairie	M. Dietze	X. Feng	*	active
Poplar, Willow, Woody	Hardwood	M. Dietze	D. Wang	*N. Brady	active
sugarcane	Sugarcane	F. Miguez	D. Jaiswal	F. Hussain	active
syntheses	synthesis papers	M. Dietze	D. LeBauer	*D. Bettinardi	complete
face	FACE/NCEAS	M. Dietze	D. LeBauer	*Andy Tu	complete
switchgrass	Switchgrass	M. Dietze	D. LeBauer		inactive

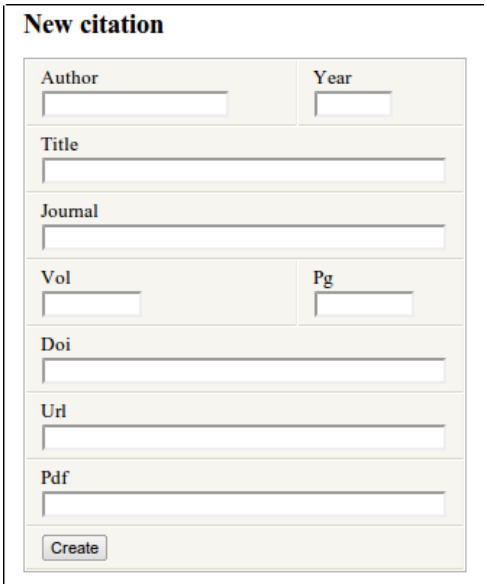


Figure 1: Form for entering a new citation.

5.1 Adding a Citation

Citation provides information regarding the source of the data. This section should allow us to locate and access the paper of interest.

A pdf copy of each paper should be available through Mendeley.

1. Select one of the starred papers from your projects Mendeley folder.
2. The data to be entered should be specified in the notes associated with the paper in Mendeley
3. Identify (highlight or underline) the data (means and statistics) that you will enter
4. Enter citation information (Figure 1)
   (a) Data entery form for a new site: `BETYdb` → `Citations` →`new`
   (b) Author Input the first author's last name only
   (c) Year Input the year the paper was published, not submitted, reviewed, or anything else
   (d) For unknown information, input NA
   (e) URL web address of the article, preferably from publishers website
   (f) PDF URL of the PDF of the article
   (g) DOI is the 'digital object identifier'. If DOI is available, PDF and URL are optional. This can be located in the article or in the article website. Use Ctrl+F 'DOI' to find it. Some older articles do not have a DOI.

5.2 Adding a Site

Each experiment is conducted at a unique site. In the context of BETY, the term 'site' refers to a specific location and it is common for many sites to be located within the same experimental station. By creating distinct records for multiple sites, it is possible to differentiate among independent studies.

1. Before adding a site, search to make sure that site is not already entered in database.
2. search for the site given latitude and longitude
    - if an institution name or city and state are given, try to locate the site on Google Maps
    - if a site name is given, try to locate the site using a combination of Google and Google Maps
    - if latitude and longitude are given in the paper, search by lat and lon, this will return all sites within ± 1 degree lat and long.
    - if an existing site is plausibly the same site as the one mentioned in the paper, it will be necessary to check other papers linked to the existing site.
        - use the same site if the previous study uses the *exact same location* and experimental setup.
        - create a new site if the study was conducted in a different fields (i.e., not the exact same location).
        - create a new site if one study was conducted in a greenhouse and another in a field.
        - do not use distinct sites for seed source in a common garden experiment (see 'When not to enter a new site' below)
3. to use an existing site, click 'edit' for the site, and then select current citation under 'add citation relatinships'
4. if site does not exist, add a new site.

When not to enter a new site: When plants (or seeds) are collected from multiple locations and then grown in the same location, this is called a 'common garden experiment'. In this case, the location of the study is included as site information. Information about the seed source can be entered as a distinct cultivar.

Site name* Site identifier, sufficient to uniquely identify the site within the paper

City Nearest city

State State, if site is in US

Country*

Longitude*

Latitude* Latitude and Longitude must be in decimal form. To convert minute-second to decimal degrees, see the equation in Table 9.

Greenhouse* set Greenhouse = TRUE if plants were grown in a greenhouse, growth chamber, or pots. If a 'warming chamber' or 'greenhouse' is used as the experimental manipulation, but is not used in the control treatments, Greenhouse = FALSE.

Soil Soil class is entered as a categorical variable that describes the texture. If percent clay, sand, and silt are given, Figure 2 can be used to look up the class.

SOM Soil organic matter (% by weight)

MAT Mean Annual Temperature (^{o}C)

MAP Mean Annual Precipitation (mm)

MASL Elevation (meters above sea level, m)

Notes site details not included above

Soilnotes soil details not included above

Rooting Zone Depth Depth of rooting zones in meters

Depth to Water Table Depth to water table in meters

5.2.1 Site Location

If latitude and longitude coordinates not available, it is often possible to determine the site location based on the site name, city, and other information. One way to do this would be to look up a location name in google maps and then locate it on the embedded map. Google Maps can provide decimal degrees if the LatLng feature is enabled, which can be done here. Google Earth can be particularly useful in locating sites, along with their coordinates and elevation. Alternatively, the site website or address might be found through an internet search (e.g. Google).

Use Table 2 to determine the number of significant digits to indicate the level of precision with which a study location is known.

5.3 Adding a Treatment

Treatments provide a description of a study's treatments. Any specific information such as rate of fertilizer application should be recorded in the managements table (section subsection 5.4). In general, managements are recorded when Yield data is collected, but not when only Trait data are collected.

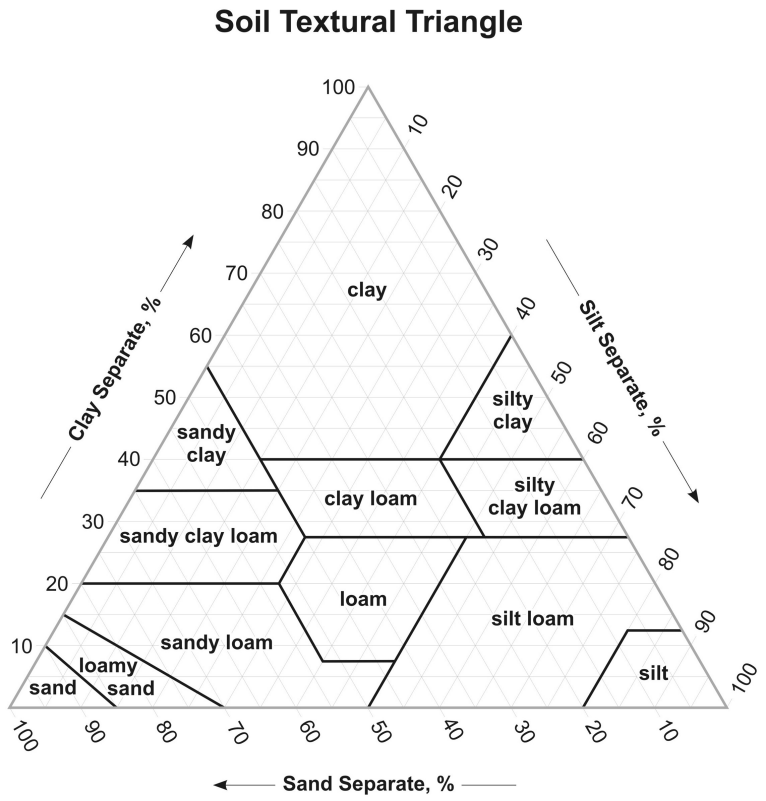


Figure 2: USDA soil classification.

Table 2: Appropriate precision for site latitude and longitude

location detail	decimal degree accuracy
city	0.1
mile	0.01
acre	0.001
10 meters	0.0001

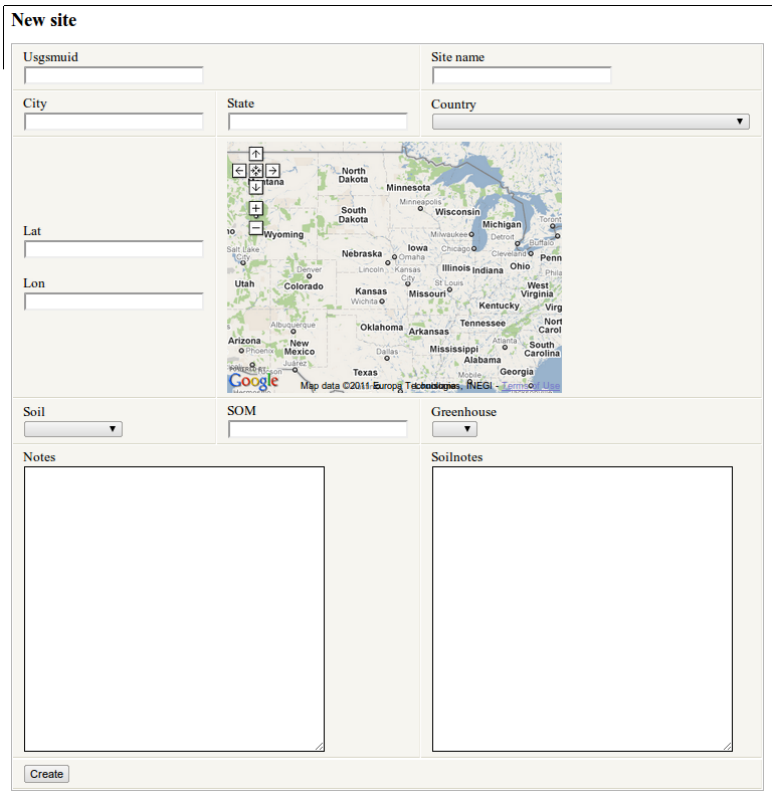


Figure 3: Form for entering a new site.

When not to use treatment: predictor variables that are not based on distinct managements, or that are distinguished by information already contained in the trait table (e.g. site, cultivar, date fields) should not be given distinct treatments. For example, a study that compares two different species, cultivars, or genotypes can be assigned the same control treatment; these categories will be distinguished by the species or cultivar field. Another example is when the observation is made at two sites: the site field will include this information.

A treatment name is used as a categorical (rather than continuous) variable: it should be easy to find the treatment in the paper based on the name in the database. The treatment name does not have to indicate the level of treatment used in a particular treatment - this information will be included in management table.

It is essential that a control group be identified with each study. If there is no experimental manipulation, there is only one treatment. In this case, the treatment should be named 'observational' and listed as control.

To determine the control when it is not explicitly stated, first determine if one of the treatments is most like a background condition or how a system would be in its non-experimental state. In the case of crops, this could be how a farmer would be most likely to treat a crop.

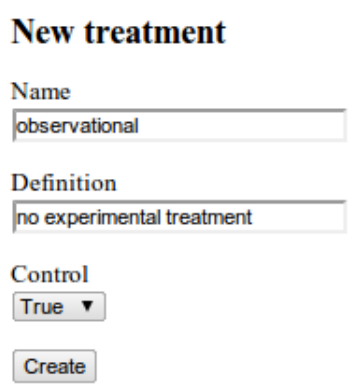


Figure 4: Form for entering a new treatment.

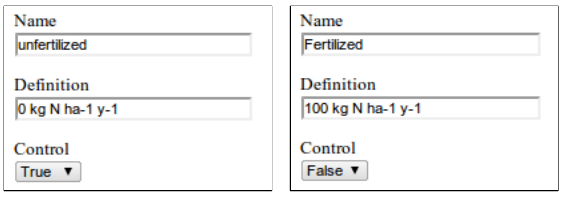


Figure 5: Example treatment form with control and experimental information
Example of data entered into the treatment form for a control (left) and treatment (right)

Name indicates type of treatment; it should be easy for anyone with the original paper to be able to identify the treatment from its name.

Control make sure to indicate if the treatment is the study 'control' by selecting true or false

Definition indicates the specifics of the treatment. It is useful for identification purposes to use a quantified description of the treatment even though this information can only be used for analysis when entered as a management.

5.4 Adding a Management

Managements refers to something that occurs at a specific time and has a quantity; Managements include actions that are done to a plant or ecosystem, for example the planting density or rate of fertilization. Managements are distinct from Treatments in that a Treatment is used to categorically identify an experimental treatment, whereas a management is used to describe what has been done.

Managements are the way a treatment becomes quantified. Each treatment is often associated with multiple managements. The combination of managements associated with a particular treatment will distinguish it from other treatments. The management types that can be entered into BETY are described in Table 3.

Each management may be associated with one or more treatments. For example, in a fertilization experiment, planting, irrigation, and herbicide managements would be applied to all plots but the fertilization will be specific to a treatment. For a multi-year experiment, there may be multiple entries for the same type of management, reflecting, for example, repeated applications of herbicide or fertilizer, for example see Figure 6.

note: At present, managements are recorded for Yields but not for Traits, unless specifically required by the data or project manager.

To associate a management with multiple treatments, first create the management, then edit the management and add treatment relationships.

mgmttype the name of the management being used. A list of standardized management types can be found in Table 3.

Level a quantification of mgmttype.

units refers to the units of the level. Units should be converted to those in Table 3.

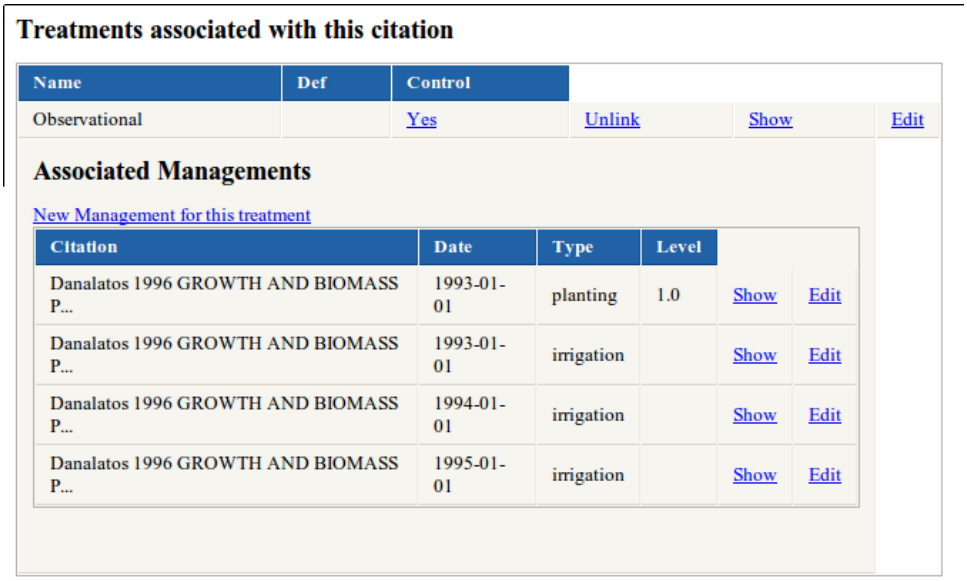
dateloc date level of confidence, explained in subsection 5.5 and defined in Table 6.

5.5 Adding a Trait

In general, a 'trait' is a phenotype; a characteristic that the plant exhibits. The traits that we are primariliy interested in collecting data for are listed in Table 6.

Before adding trait data, it is necessary to have the citation, treatments, and site information already entered. If the correct citation is not identified at the top of the page Figure 8. To add a new Trait, go to the new trait page: `Trait` → `new`.

Figure 6: Form for entering and editing relationships between treatments and managments



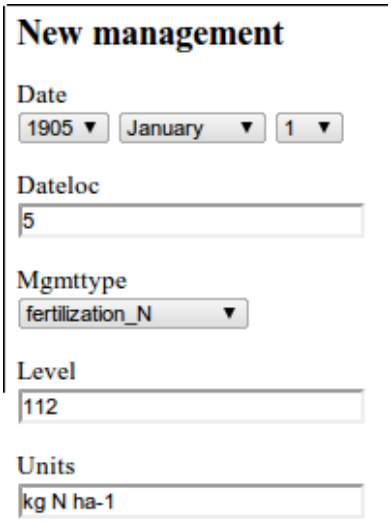


Figure 7: Form for entering new managment
Form for entering management data with example data. This management denotes a nitrogen fertilization rate of 112.0 kg N ha-1.

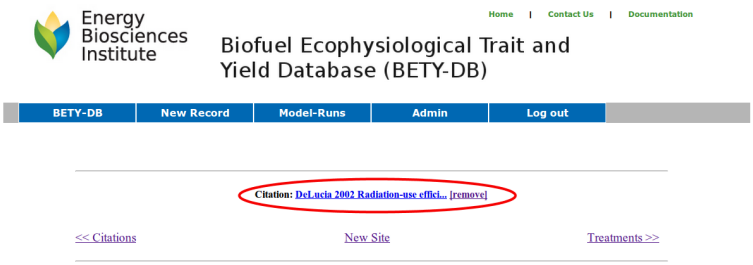


Figure 8: When entering data, the citation is identified at the top of the page.

Presently, we are also using the Trait table record ecosystem level measurements other than Yield. Such ecosystem level measurements can include leaf area index or net primary productivity, but are only collected when required for a particular project.

Figure 9 shows the web form for entering new trait data, and Table 6 provides a list of the traits that we are interested in collecting.

Most of the fields in the Traits table are also used in the Yields table. Here is a list of the fields with a brief description, followed by more thorough explanations:

Species* Search for species in the database using the search box; if species is not found, see Section 5.10 Adding a Species

Cultivar primarily used for crops; If the cultivar being used is not found in drop-down box, see hyperref[sec:addcultivar]Section 5.11: Adding a Cultivar.

DateLOC Date Level of confidence. See Table 4 for values.

Mean* mean is in units of tons per hectare per year (t/ha)

Stat name is name of the statistical method used (usually one of SE, SD, MSE, CI, LSD, HSD, MSD). See section 5.5.2 for more details.

Statistic is the value of the statistic associated with Stat name.

N Always record N if provided. N is the number of experimental replicates, often referred to as the sample size; N represents the number of independent units within each treatment: in a field setting, this is often the number of plots in each treatment, but in a greenhouse, growth chamber, or pot-study this may be the number of chambers, pots, or individual plants. Sometimes this value is not clearly stated.

5.5.1 dateloc

The date level of confidence (DateLOC) provides an indication of how accurately the date associated with the trait or yield observation is known. Table 4 provides the values that should be entered in this field. If the event occurred at a level of precision not defined by an integer in this table, use fractions. For example, we commonly use 5.5 to indicate a one

Figure 9: Form used to enter a new trait.

week level of precision. If the exact year is not known, but the time of year is, use 91 to 97, with the second digit to indicate the information known within the year.

5.5.2 Statistics

Our goal is to record statistics that can be used to estimate standard deviation or standard error. Many different methods can be used to summarize data, and this is reflected in the diversity of statistics that are reported. An overview of these methods is given in Table 5 and a description below.

Where available, direct estimates of variance are preferred, including Standard Error (SE), sample Standard Deviation (SD), or Mean Squared Error (MSE). SE is usually presented in the format of mean(±SE). MSE is usually presented in a table. When extracting SE or SD from a figure, measure from the mean to the upper or lower bound. This is different than confindence intervals and range statistics (described below), for which the entire range is collected.

If MSE, SD, or SE are not provided, it is possible that LSD, MSD, HSD, or CI will be provided. These are range statistics and the most frequently found range statistics include a Confidence Interval (95%CI), Fisher's Least Significant Difference (LSD), Tukey's Honestly Significant Difference (HSD), and Minimum Significant Difference (MSD). Fundamentally, these methods calculate a range that indicates whether two means are diffent or not, and this range uses different approaches to penalize multiple comparisons. The important point is that these are ranges and that we record the entire range.

Another type of statistic is a "test statistic"; most frequently there will be an F-value that can be useful, but this should not be recorded if MSE is available. Only if there is no other information available, record the P-value.

5.6 Adding a Yield

The protocol for entering yield data is identical to entering data for a trait, with a few exceptions:

1. There are no covariates associated with yield data
2. Yield data is always the dry harvestable biomass; if necessary, moisture content can be added as a trait

5.7 Adding a Covariate

Covariates are required for many of the traits. Covariates generally indicate the environmental conditions under which a measurement was made. Without covariate information, the trait data will have limited value.

A complete list of required covariates can be found in Table 7. For all respiration rates and photosynthetic parameters, temperature is recorded as a covariate. Soil moisture, humidity, and other such variables that were measured at the time of the measurement that may be required in order to standardize across studies.

When root data is recorded, the root size class needs to be entered as a covariate. The term 'fine root' often refers to the ¡2mm size class, and in this case, the covariate `root_maximum_diameter` would be set to 2. If the size class is a range, then the `root_minimum_diameter` can also be used.

To add a new covariate, go to the new covariate page: `Covariate` → `new`.

5.7.1 Extracting information from tables and graphs

1. identify the data that is associated with each treatment

   *note:*if the experiment has many factors, the paper may not report the mean and statistics for each treatment. Often, the reported data will be reflect the results of more than one treatment, for example if there was no effect of the treatement on the quantity of interest. In some cases it will be possible to calculate the values for each treatment, e.g. if there are $n-1$ values and n treatments. If this is not the case, the treatment names and definitions should be changed to indicate that data reflect the results of more than one experimental treatments
2. enter the mean value of the trait
3. enter the `statname`, `stat`, and number of replicates, `n` associated with the mean
   - `stat` is the value of the `statname` (i.e. `statname` might be 'standard deviation' (SD) and the `stat` is the numerical value of the statistic)
   - always measure size of error bar from the mean to the end of an error bar. This is the value when presented as $X \pm SE$ or $X(SE)$ and may be found in a table or on a graph.
   - sometimes CI and LSD are presented as the entire range from the lower to the upper end of the confidence interval. In this case, take 1/2 of the interva representing the distance from the mean to the upper or lower bound

5.7.2 Extracting Data From a Figure using R

To extract data from a jpg file in R using the digitize package:

1. save image as a `*.jpg` file
2. open R
3. change the directory that R is using to the one where the image is
4. use R code below to extract data, display it, and save it in a `csv` file steps below
5. upload csv to the project file in google spreadsheet, or open as excel/openoffice and copy/paste to google spreadsheet

```
library('digitize')
calpts <- ReadAndCal('authorYYYabc_fixX.jpg')
   ## click on xaxis min (x1), xaxis max (x2),
   ## yaxis min(y1), yaxis max (y2)
pts <- DigitData(col='red',type='p',n=8)
   ## set n = to the number of points to collect
data <- Calibrate(pts, calpts, x1, x2, y1, y2)
   ## x1, x2, y1, and y2 are the min and max of the x and y axes, respectively.
print(data)
write.csv(data, 'authorYYYYabc_figX.csv')
```

5.7.3 Extracting Data From a Figure using GetData

1. Open pdf in Adobe Reader.
2. Zoom in on the figure
3. Choose `Tools` → `Select and Zoom`
4. Open Paint
5. Paste Picture
6. Save as authorYYYYabc_figX.jpg
7. Open Get Data
8. `File` → `open` open figure
9. select button with two arrows (fourth from left)
10. follow instructions to select x min, x max, y min and y max. If the x-axis has a categorical variable, it does not matter what values you use for x min and x max.
11. make sure to set the correct values for the max and min of each axis, and indicate if the axis is log-scaled
12. select the target button (seven from left)
13. click over center of desired data points and error bars
14. copy data to google spreadsheet (see Section 3
15. calculate SE as the distance between the error bar upper bound and the mean (absolute value of difference between the two points)

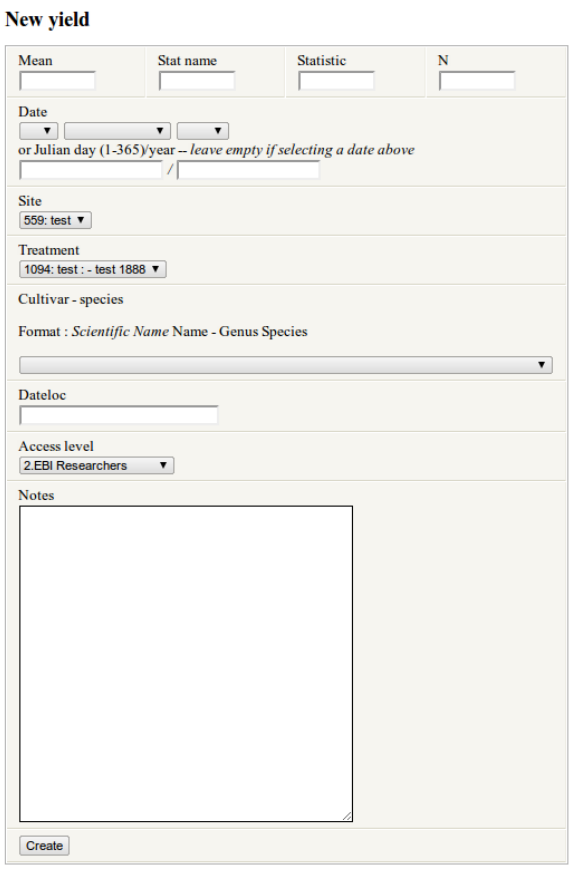


Figure 10: Form used to enter a new Yield.

5.8 Adding a Yield

To add a new Yield, go to the new yield page: `Yield` → `new`. Yield is equivalent to above-ground biomass on a per-area basis, and has units of Mg ha^{-1} y−1

5.9 Adding a PFT

Plant functional types (PFTs) are used to group plants for statistical modeling and analysis. PFTs are associated with both a specific set of priors, and a the subset species for which the traits and yields data will be queried. In many cases, it is appropriate to use default PFTs (e.g. `tempdecid` is temperate deciduous trees)

In other cases, it is necessary to define pfts for a specific project. For example, to query a specific set of priors or subset of species, a new PFT may be defined. For example, Xiaohui Feng defined PFT's for the species found at the EBI Farm parairie. Such project-specific PFTs can be defined as `‘projectname‘.‘pft‘` (i.e. `ebifarm.c4grass` instead of `c4grass`).

5.10 Adding a Species

Species that are found or cultivated in the United States should be in the Plants table. Look it up there first.

5.11 Adding a Cultivar

6 BETYdb: Bulk data upload

Currently the web interface does not support bulk data upload, although this is a planned feature for BETY 2.0.

For bulk data upload, or for a complete view of the tables in BETY, a blank spreadsheet can be found online and can be downloaded in .xls format spreadsheet. Contact David LeBauer or Mike Dietze for more information about using this method of data upload.

7 BETYdb: QA/QC with the web interface

Quality assurance and quality control (QA/QC) is a critical step that is used to ensure the validity of data in the database and of the analyses that use these data. When conducting QA/QC, your data access level needs to be elevated to "manager".

1. open citation in Mendeley
2. locate citation in BETYdb
    - select 'use'
    - select 'show'
    - check that author, year, title, journal, volume, and page information is correct
    - check that links to URL and PDF are correct, using doi if available
    - if any information is incorrect, click 'edit' to correct.
3. check that site(s) at bottom of citation record match site(s) in paper
    - check that latitude and longitude are consistent with manuscript, are in decimals not degrees, and have appropriate level of precision Table 2.
    - click on site name to verify any additional information site information that is present
    - enter any additional site level information that is found
4. select treatments from menubar
    - check that there is a control treatment
    - ensure that treatment name and definition are consistent with information in the manuscript.
    - under "treatments from all citations associated with associated sites", ensure that there is no redundancy (i.e. if another citations uses the same treatments, it should not be listed separately)

- if managements are listed, make sure that managment-treatment associations are correct

5. check managements if there are any listed on the treatments page.
   - if Yield data have been collected, ensure that required managements have been entered
   - if managements have been entered, ensure that they are associated with the correct treatments
6. click Yields or Traits to check data.
   - check that means, sample size, and statistics have been entered correctly
   - if data has been transformed, check that transformation was correct in the associated google spreadsheet (or create a new google spreadsheet following instructions in section 3).
   - for any trait data that requires a covariate Table 7.

8 Acknowledgements

Patrick Mulroony pat@life.illinois.edu implemented the data entry interface. Moein Azimi, David Bettinardi, and Nick Brady, along with other members of the Dietze lab, have contributed to the ongoing development this document and the web interface that it describes.

9 Appendix

9.1 Transformations

9.1.1 Statistics

9.1.2 Variables

9.2 Calculations used in transformations

10 Converting from $\frac{\mu \text{L } O_2}{10\text{mg h}}$ to $\frac{\text{nmol}CO_2}{\text{g s}}$ and $\frac{\mu\text{mol } CO_2}{\text{kg s}}$, including adjustment for temperature

10.1 Objective:

Convert from root respiration data reported in George et al (where O_2 was measured in μL to units of mass.

In the appendix table, George 2003 reports the range of root respiration rates, converted to 15 $^\circ C$ and standard units:

$$[11.26, 22.52]\frac{\text{nmolCO}_2}{\text{g s}}$$

In the original publication Allen (1969), root respiration was measured at 27 °C. The values can be found in table 3 and figure 2. The data include a minimum (Group 2 Brunswick, NJ plants) and a maximum (Group 3 Newbery, South Carolina), which I assume are the ones used by George 2003:

$$[27.2, 56.2]\frac{\mu\text{L O}_2}{\text{10mg h}}$$

10.2 Step 1

Transformed George 2003 measurements back to the measurement temperature using a rearrangement of equation 1 from George, the standardized temperature of 15 °C stated in the Georgeh table legend, and $Q_{10} = 2.075$ from George 2003, and the measurement temperature of 27 °C reported by Allen 1969:

$$R_T = R_{15}[\exp(\ln(Q_{10})(T - 15))/10]$$

$$[11.26, 22.52] * exp(log(2.075) * (27 - 15)/10)$$

Now we have the values that we would have expected to find in the Allen paper, except that the units need to be converted back to the original:

$$[27.03, 54.07]\text{nmolCO}_2\ \text{g}^{-1}\text{s}^{-1}$$

10.3 Step 2 converting the units

10.4 Required constants:

- 1 mol O_2 = 1 mol CO_2 since respiration is $CH_2O + O_2 \rightarrow CO_2 + H_2O$
- density of O_2 at $27°C$: $\frac{7.69\times10^5\ \text{ml O}_2}{\text{g O}_2}$ first assume that Allen converted to sea level pressure (101 kPa), although maybe they were measured at elevation (Allen may have worked at ˜ 900 kPa near Brevard, NC)
- molar mass of O_2: $\frac{32\text{g O}_2}{\text{mol}}$
- treat 10mg, which is in the unit of root mass used by Allen, as a unit of measurement for simplicity

Now convert

$$[27.03, 54.07]\text{nmolCO}_2\ \text{g}^{-1}\text{s}^{-1}$$

to units of $\frac{\mu\text{L O}_2}{\text{10mg root h}}$. The expected result is the original values reported by Allen: $[27.2, 56.2]\frac{\mu\text{L O}_2}{\text{10mg h}}$

$$[27.03, 54.07]\ \frac{\text{nmol } CO_2}{\text{g root s}} \times \frac{1\text{ g}}{100 \times 10\text{mg}} \times \frac{3600\text{ s}}{\text{h}} \times \frac{\text{nmol } O_2}{\text{nmol } CO_2} \frac{3.2 \times 10^{-8}\text{ g } O_2}{\text{nmol } O_2} \times \frac{7.69 \times 10^5\ \mu\text{L } O_2}{\text{g } O_2}$$

The result is:

$$[23.8, 47.8] \frac{\mu\text{L } O_2}{\text{10mg root h}}$$

These are the units reported in the Allen paper, but they appear to be off by the temperature conversion factor, $exp(log(2.075) * (27-15)/10) = 2.4$, e.g. $[11.9, 23.9] \times 2.4 = [28.6, 57.4]$, values which are only 5 and 2 percent larger than the original values of $[27.2, 56.2]$, respectively to be acceptable, but not exact. Since the ratio of observed:expected values are different, it is not likely that Q_{10} or the atmospheric pressure at time of measurement would explain this error.

10.5 convert to units in BETYdb, find k

:

$$\text{k} \times \frac{\mu\text{L } O_2}{\text{10mg root h}} = \frac{\mu\text{mol } CO_2}{\text{kg s}}$$

$$k = \frac{\text{g } O_2}{7.69 \times 10^5\ \mu\text{L } O_2} \times \frac{\mu\text{mol } O_2}{3.2 \times 10^{-5}\text{ g } O_2} \times \frac{10^5\ \times 10\text{mg}}{\text{kg}} \times \frac{\text{h}}{3600\text{ s}} =$$
$$= 1.13$$

11 Calculating MSE given F, df_{group}, and SS

Given:

$$F = MS_g / MS_e \tag{1}$$

Where g indicates the group, or treatment. Rearranging this equation gives:

$$MS_e = MS_g / F$$

Given

$$MS_x = SS_x / df_x$$

Substitute MS_e/df_e for SS_e in the first equation

$$F = \frac{SS_g / df_g}{MS_e}$$

	d.f.	Sum of Squares	*F*-value
Eriophorum vaginatum			
Treatment	2	109.58	0.570
Weeks	10	2151.52	5.095
treatment*weeks	20	1482.43	1.755

Figure 11: Table used to calculate SE from F, from Starr et al. [2008].

Then solve for MS_e

$$MS_e = \frac{SS_g}{df_g \times F} \tag{2}$$

$$df_{\text{total}} = (df_a + 1) \times (df_b + 1)... \times (n) - 1 \tag{3}$$

which depends on the experimental design:

for factors a, b... (usually 1 or 2, sometimes 3) where n is the number of replicates within each treatment combination.

- one-way anova $df_{\text{total}} = an - 1$; where a is the number of treatments
- two-way anova without replication $df_{\text{total}} = (a+1)(b+1)-1$ also known as "randomized complete block design" (RCBD)
- two-way anova with n replicates $df_{\text{total}} = (a + 1)(b + 1)(n) - 1$ aka "RCBD with replication"

11.1 Example

An example application of this is in Starr et al. [2008] table 3 (Figure 11).

The results are from one (two?) factor ANOVA with repeated measures, with treatment and week as the factors and no replication.

We will calculate MSE from the $SS_{\text{treatment}}$ $df_{\text{treatment}}$, and F-value given in the table; these are 109.58, 2, and 0.570, respectively; df_{weeks} is given as 10.

For the 1997 *Eriphorium vaginatum*, the mean A_{max} in table 4 is 13.49.

Calculate MS_e:

$$MS_e = \frac{109.58}{0.57 \times 2} = 96.12$$

12 Bibliography

References

Ch. Korner, G. D. Farquhar, and Z. Roksandic. A global survey of carbon isotope discrimination in plants from high altitude. *Oecologia*, 74(4):623–632, January 1988. ISSN 0029-8549. doi: 10.1007/BF00380063. URL `http://www.springerlink.com/index/10.1007/BF00380063`.

Gregory Starr, SF Oberbauer, and LE Ahlquist. The photosynthetic response of Alaskan tundra plants to increased season length and soil warming. *Arctic, Antarctic, and Alpine Research*, 40(1):181–191, 2008. doi: 10.1657/1523-0430(06-015). URL `http://instaar.metapress.com/index/L5L27V837P506226.pdf`.

Table 3: Managements

This is a list of managements to enter, with the most common management types in **bold**. It is more important to have management records for Yields than for traits. For greenhouse experiments, it is not necessary to include informaton on fertilizaton, lighting, or greenhouse temperature.

Management Type	Units	Definition	notes
burned		aboveground biomass burned	
CO2_fumigation	ppm		
fertilization_*X*	kg X ha^{-1}	fertilization rate, element X	
fungicide	kg ha^{-1}		add type of fungicide to notes
grazed	years	livestock grazing	pre-experiment land use
harvest			no units, just date, equivalent to coppice, aboveground biomass removal
herbicide	kg ha^{-1}		add type of herbicide to notes: glyphosate, atrazine, many others
irrigation	cm		convert volume / area to depth as required
light	W m^{-2}		
O3_fumigation	ppm		
pesticide	kg ha^{-1}		add type of pesticide to notes
planting	plants m^{-2}		convert row spacing to planting density if possible
seeding	kg seeds ha^{-1}		
tillage			no units, maybe depth; *tillage* is equivlent to *cultivate*

Table 4: Date level of confidence (DateLOC) Field

Numbering convention for the DateLOC (Date level of confidence) field, used in managements, traits, and yields table.

Dateloc	Definition
9	no data
8	year
7	season
6	month
5	day
4	time of day i.e. morning, afternoon
3	hour
2	minute
1	second
95	unknown year, known day
96	unknown year, known month
... etc	

Table 5: List of statistical summaries

List of the statistics that can be entered into the statname field of traits and yields tables. Please see David (or Mike) if you have questions about statistics that do not appear in this list. If you have P, or LSD in a study with $n \neq b$ (e.g. not a RCBD, see Table 8), please convert these values prior to entering the data, and add a note that stat was transformed to the table. Note: These are listed in order of preference, e.g., if SD, SE, or MSE are provided then use these values

statname	name	definition	notes
SD	Standard Deviation, s	$\sqrt{\frac{1}{N}\sum(x_i - \bar{x})^2}$	$\bar{x}$ is the mean
SE	Standard Error	$\frac{s}{\sqrt{n}}$	
MSE	Mean Squared Error		like SD, but with multiple treatments. in R: $\frac{mean(aov(y\ x)\$residuals^2}{aov(y\ x)\$df}$
95%CI	95% Confidence Interval	$t_{1-\alpha/2,n} * s$	measure the 95% CI from the mean, this is actually $^1/_2$ of the CI
LSD	Least Significant Difference	$t_{1-\frac{\alpha}{2},n}\sqrt{2\text{MSE}/b}$	b is the number of blocks (Rosenberg 2004)
MSD	Minimum Significant Difference		

Variable	units	median (90%CI) or range	definition
Vcmax	μmol CO_2 m^2 s^{-1}	44(12, 125)	maximum rubisco carboxylation capacity
SLA	m^2 kg^{-1}	15(4, 27)	Specific Leaf Area area of leaf per unit mass of leaf
LMA	kg m^{-2}	0.09(0.03, 0.33)	Leaf Mass Area (LMA = SLM = 1/SLA) mass of leaf per unit area of leaf
leafN	%	2.2(0.8, 17)	leaf percent nitrogen
c2n leaf	leaf C:N ratio	39(21, 79)	use only if leafN not provided
leaf turnover rate	1/year	0.28(0.03, 1.0)	
Jmax	μmol photons m^{-2} s^{-1}	121(30, 262)	maximum rate of electron transport
stomatal slope		9(1, 20)	
GS			stomatal conductance (= gs_{max}
q*		0.2–5	ratio of fine root to leaf biomass
**grasses*:	ratio of root:leaf = below : above ground biomass		
aboveground biomass	g m^{-2} *or* g $plant^{-1}$		
root biomass	g m^{-2} *or* g $plant^{-1}$		
**trees*:	ratio of fine root:leaf biomass		
leaf biomass	g m^{-2} *or* g $plant^{-1}$		
fine root (<2mm) biomass	g m^{-2} *or* g $plant^{-1}$		
root turnover rate	1/year	0.1–10	rate of fine root loss (temperature dependent) $year^{-1}$
leaf width	mm	22(5,102)	
growth respiration factor	%	0–1	Proportion of daily carbon gain lost to growth respiration
R_{dark}		μmol CO_2 m^{-2} s^{-1}	dark respiration
quantum efficiency	%	0–1	efficiency of light conversion to carbon fixation, see Farqhuar model,
dark resp factor	%	0–1	converts Vm to leaf respiration
seedling mortality	%	0–1	proportion of seedlings that die
r fract	%	0–1	fraction of storage to seed reproduction
root respiration rate*	CO_2 kg^{-1} fine roots s^{-1}	1–100	rate of fine root respiration at reference soil temperature
f labile	%	0–1	fraction of litter that is goes

variable	required covariates	optional covariates
Vcmax	irradiance and temperature (leaf or air)	
any leaf measurement		canopy height
root_respiration_rate	temperature (root or soil)	soil moisture
root_respiration_rate	root_diameter_max	root size class (usually $< 2mm$)
any respiration	temperature	
root biomass		min size cutoff, max size cutoff
root, soil	depth (cm)	used for max and min depths of soil, if only one value, assume min depth = 0; negative values indicate above ground
gs (stomatal conductance)	A_{max}	†see notes in caption
stomatal_slope (m)	humidity, temperature	?specific humidity? assume leaf T = air T?

Table 7: Traits with required covariates

A list of traits and the covariates that must be recorded along with the trait value in order to be converted to a constant scale from across studies. *notes:* † stomatal conductance (gs) is only useful when reported in conjunction with other photosynthetic data, such as A_{max}. Specifically, if we have Amax and gs, then estimation of Vcmax only covaries with `dark_respiration_factor` and atmospheric CO2 concentration. We also now have information to help constrain `stomatal_slope`. If we have `Amax` but not `gs`, then our estimate of `Vcmax` will covary with: `dark_respiration_factor`, CO_2, `stomatal_slope`, `cuticular_conductance`, and vapor-pressure deficit (`VPD`) (which is more difficult to estimate than CO_2, but still possible given lat, lon, and date). Most important, there will be a strong covariance between `Vcmax` and `stomatal_slope`.

Table 8: How to convert statistics from P, LSD, or MSD to SE

from	to	conversion	rcode	notes
P	SE	$SE = \frac{X_1 - X_2}{t_{1-P/2,2n-2}\sqrt{2/n}}$	`(x1-x2)/(qt(1-P/2,2*n-2)*sqrt(2/n))`	$\bar{X}_{1,2}$ are two means be compared.
LSD	SE	$SE = \frac{LSD}{t_{1-\alpha/2,n}*\sqrt{2b}}$	`LSD/(qt(1-P/2,n)*sqrt(2*b))`	where b is the number blocks, n is the num of replicates, and n = in a Randomized Comp Block Design
MSD	SE	$SE = \frac{MSD*n}{t_{1-\alpha,2n-2}*\sqrt{2}}$	`msd*n/(qt(1-P/2,2*n-2)*sqrt(2))`	

Table 9: Useful conversions for entering site, management, yield, and trait data

from (X)	to (Y)	conversion	notes
DD°MM'SS	XX.ZZZZ	XX.ZZZZ = XX + MM/60 + SS/60	to convert latitude or lo tude from degrees, minu seconds to decimal degre
lb	kg	$Y = X \times 2.2$	
m^2	ha	$Y = X/10^6$	
g/m^2	kg/ha	$Y = X \times 10$	
US ton/acre	Mg/ha	$Y = X * 2.24$	
m^3/ha	cm	$Y = X/100$	units used for irrigation rainfall
% roots	root:shoot (q)	$Y = \frac{X}{1-X}$	%roots = $\frac{\text{root biomass}}{\text{total biomass}}$
μ mol cm^{-2} s^{-1}	mmol $m^{-2}s^{-1}$	$Y = X/10$	
mol m^{-2} s^{-1}	mmol $m^{-2}s^{-1}$	$Y = X/10^6$	
mol m^{-2} s^{-1}	μmol $cm^{-2}s^{-1}$	$Y = X/10^5$	
mm s^{-2}	mmol m^{-3} s^{-1}	$Y = X/41$	Korner et al. [1988]
mg CO_2 g^{-1} h^{-1}	μmol kg^{-1} s^{-1}	$Y = X \times 6.31$	used root_respiration_rate
μmol	mol	$Y = X \times 10^6$	
julian day (1–365)	date	see NASA Julian Calendar	
spacing (m)	density (plants m^2)	$Y = \frac{1}{\text{row spacing} \times \text{plant spacing}}$	
kg $ha^{-1}y^{-1}$	Mg $ha^{-1}y^{-1}$	$Y = X/1000$	
g $m^{-2}y^{-1}$	Mg $ha^{-1}y^{-1}$	$Y = X/100$	
kg	mg	$Y = X \times 10^6$	
cm^2	m^2	$Y = X \times 10^4$	