GLOBAL CHANGE BIOLOGY
BIOENERGY

GCB Bioenergy (2018) **10**, 61–71, doi: 10.1111/gcbb.12420

TECHNICAL ADVANCE

BETYdb: a yield, trait, and ecosystem service database applied to second-generation bioenergy feedstock production

DAVID LEBAUER[1,2], ROB KOOPER[2], PATRICK MULROONEY[1,3], SCOTT ROHDE[1], DAN WANG[1,4], STEPHEN P. LONG[1,5,6,7] and MICHAEL C. DIETZE[1,8]

[1]*Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA* , [2]*NCSA (National Center for Supercomputing Applications), University of Illinois at Urbana-Champaign, Urbana, IL 61801-2311, USA,* [3]*San Diego Supercomputer Center, University of California at San Diego, CA, 92093-0505, USA,* [4]*International Center for Ecology, Meteorology and Environment School of Applied Meteorology, Nanjing University of Information Science and Technology, Jiangsu, 210044, China,* [5]*Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801-4730, USA,* [6]*Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801-3750, USA,* [7]*Lancaster Environment Centre, University of Lancaster, Lancaster, LA1 4YQ, UK,* [8]*Department of Earth and the Environment, Boston University, Boston, MA, 02215, USA*

Abstract

Increasing demand for sustainable energy has led to research and development on the cultivation of diverse plant species for biomass production. To support the research and development required to domesticate and cultivate crops for bioenergy, we developed the Biofuel Ecophysiological Traits and Yields database (BETYdb). BETYdb is a centralized open-access repository that facilitates organization, discovery, and exchange of information about plant traits, crop yields, and ecosystem functions. BETYdb provides user interfaces to simplify storage and discovery as well as programming interfaces that support automated and reproducible scientific workflows. Presently, BETYdb contains over forty thousand observations of plant traits, biomass yields, and ecosystem dynamics collected from the published articles and ongoing field studies. Over half of these records represent fewer than ten genera that have been intensively evaluated for biomass production, while the other half represent over two thousand plant species reflecting research on new crops, unmanaged ecosystems, and land use transitions associated with bioenergy. BETYdb has been accessed over twenty-five thousand times and is used in the fields of bioenergy and ecosystem ecology to quantify yield potential and ecosystem functioning of crops and unmanaged systems under present and future climates. Here, we summarize the database contents and illustrate its applications. We show its utility in a new analysis that confirms that Miscanthus is twice as productive as switchgrass over a much wider range of environmental and management conditions than covered in previous analyses. We compare traits related to carbon uptake and water use of these species with each other and with two coppice shrubs, poplar and willow. These examples, along with a growing body of published research that used BETYdb, illustrate the scope of research supported through this open-access database.

Keywords: bioenergy crops, database, ecosystem services, meta-analysis, open access, plant traits, yields

Received 4 July 2016; accepted 22 September 2016

Introduction

The demand for renewable energy has stimulated interest in the development of plants that can produce biomass sustainably. Because all plants produce carbon compounds that can be used as bioenergy feedstocks, a key challenge is identifying the species and associated agronomic systems most suited as feedstocks in terms of achieving both economic and environmental sustainability (Davis *et al*., 2009; Somerville *et al*., 2010). Of particular interest are species that can be grown on land that is environmentally unsuitable for food production due to factors such as soil, topography, and climate (Long *et al*., 2015). In this direction, assessments have focused on a broad range of species (Karp & Shield, 2008; Somerville *et al*., 2010), including herbaceous perennial grasses (Lewandowski *et al*., 2003; Somerville *et al*., 2010; Arundale *et al*., 2014), coppice trees and shrubs (Wang *et al*., 2013a,b), and arid-adapted succulents (Davis *et al*., 2014; Yang *et al*., 2015).

Correspondence: David LeBauer, tel. +1 217 300 0266,
fax 217 265 6800, e-mail: dlebauer@illinois.edu

© 2017 The Authors Global Change Biology Bioenergy published by John Wiley & Sons Ltd
This is an open access article under the terms of the Creative Commons Attribution License,
which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Unlike key food crops such as maize, soybean, and wheat, experimental assessment of the agronomic and ecosystem service potential of emerging biomass crops is limited to relatively few and often disparate field trials, and there has been no harmonized, open-access database for accessing, sharing, and archiving data in a single format. Indeed, the diversity of potential production systems makes the problems of both capturing and sharing relevant data more challenging than for established food crops. Nevertheless, accessing these data and associated metadata is critical for data synthesis, meta-analysis, and statistical and process-based modeling (Surendran Nair *et al*., 2012; LeBauer *et al*., 2013). By supporting research that builds on existing knowledge, the development of open databases will provide a stronger foundation for analyses in agronomy, ecology, economics, industry, and policy. These analyses include identifying the most viable opportunities for realizing bioenergy production in terms of both plant and land resources. Furthermore, such databases are a key to defining traits for selection of species and genotypes that are most suited for biomass production for given locations and purposes.

Current repositories of biomass crop production and ecosystem service data provide standardized metadata formats that enable users to find datasets of interest and download those that are freely available. Examples of such repositories include the Bioenergy KDF (bioenergykdf.net; all urls accessed June 28, 2016) and DataOne (search.dataone.org), which is a portal to archived ecological datasets. Such repositories make it efficient to publish, share, and find relevant, data but are not designed to harmonize data structures. Because data are deposited in repositories in diverse formats, combining these data for synthetic analysis is very challenging, requiring the user to develop specific code for each analysis or dataset. To allow cross-study analysis, each dataset must be evaluated and translated into a common format. This process of combining heterogeneous data into a single data structure is known as *harmonization*. Harmonization can only be achieved by creating a database where quality assessment and quality control begin at the time of deposit, when the data and metadata are translated into a common format. This is not provided by the current databases for second-generation biomass crops and was therefore a key requirement in developing BETYdb.

Existing plant trait and crop yield databases that are harmonized include the USDA Plants (NRCS 2014), USDA Quick Stats (quickstats.nass.usda.gov), and TRY databases (Kattge *et al*., 2011). These databases, however, do not focus on bioenergy crops. The USDA Plants database contains a taxonomy of plant species found in the United States along with trait and agronomic information aggregated at the species or subspecies level (NRCS 2014). The USDA Quick Stats database provides comprehensive yield, agronomic, and economic data for crops in the United States, but lacks data from experimental trials. The TRY database curates primary ecological trait data, but it does not focus on yield and ecosystem services of biomass crops, and in contrast to USDA Plants, users must apply for access to data in TRY. Because we were unable to identify any single resource to meet the research needs of the bioenergy research community, we developed a new, harmonized, and open-access repository, the Biofuel Ecophysiological Traits and Yields database (BETYdb, betydb.org).

We constructed BETYdb to centralize and standardize the available information on the physiological traits and yields of diverse plant species, with a focus on the identification and analysis of potential bioenergy feedstocks. BETYdb combines information from existing repositories, databases, publications, and archives into a centralized public repository. It provides a consistent, open, distributed, and extensible platform with agronomically meaningful context. In addition, the database provides user-friendly Web, programmatic, and map-based interfaces that allow users to contribute, query, and analyze data without any requirement of programming experience. To date, it contains over forty thousand curated observations from over six hundred sources in a consistent, harmonized format that allows users to access and contribute data.

The specific focus of BETYdb is on yield and on the plant traits affecting yield and ecosystem services, in particular carbon, water, and energy balance. These have already been used to parameterize, calibrate, and validate a range of biomass crop, ecosystem, and life cycle analysis models (Miguez *et al*., 2009; Davidson, 2012; Davis *et al*., 2012; LeBauer *et al*., 2013; Wang *et al*., 2013a, 2015; Dietze *et al*., 2014; Milbrandt *et al*., 2014; Saha & Eckelman, 2015).

Here, we describe the BETYdb database and web application and how it may be accessed and used, with examples of its capability and applications in bioenergy feedstock research.

Implementation

Software

The Biofuel Ecophysiological Traits and Yields database is a relational database, that is, a set of related tables that organize the data into primary data and contextual metadata tables that store information of study location, experimental design, management, and source of data (Fig. 1 and Table 1). The software is implemented in POSTGRESQL (v. 9.3, PostgreSQL Global Development

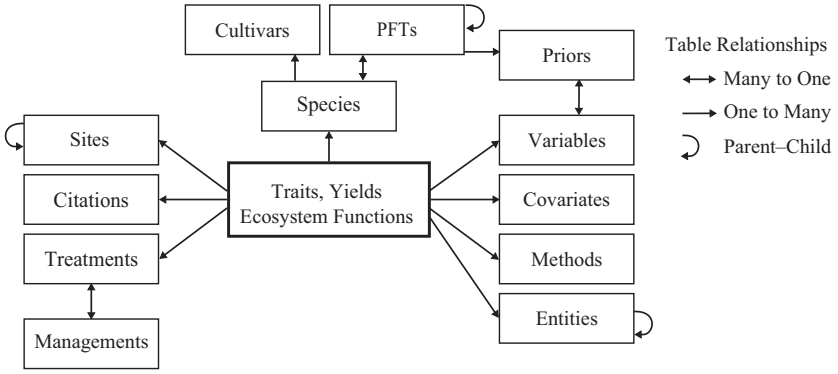

Fig. 1 A simplified entity–relationship diagram for key tables in Biofuel Ecophysiological Traits and Yields database (BETYdb). For clarity, the separate Traits and Yields tables are represented by a single box. The key data contained in the Traits and Yields tables are related to auxiliary tables which contain specific contextual data. Plant taxonomic and functional relatedness are captured in the Cultivar, Species, and Plant Functional Type tables. A full technical schema is available at betydb.org/schemas, and more details on the tables and their relationships are provided in Appendix S3.

Group postgresql.org) with a Web front end in Ruby on Rails (Ruby 2.15 ruby-lang.org; Rails 3.2, rubyonrails.org). The central BETYdb data repository is stored on a RedHat Enterprise Linux (Red Hat Inc, Raleigh, NC, USA) server at the University of Illinois, and the Web interface is available at betydb.org. In addition, BETYdb runs on Apple OSX and Ubuntu Linux operating systems. The code for the Web front end and database schema is version controlled using Git and hosted on GitHub (github.com/pecanproject/bety). Each release of the BETYdb software is assigned a version (Semantic Versioning 2.0.0, semver.org) and archived with a digital object identifier (DOI) at Zenodo (doi: 10.5281/zenodo.51898); the most recent version is BETYdb 4.9.

Database structure

The primary data are stored in two tables, *Traits* and *Yields* (italics indicate database table names). *Traits* and *Yields* store the time of observation as well as the mean, sample size, and uncertainty estimate, if any, and are linked to tables that contain contextual information such as the location of the study, species, genotype, experimental treatments, and agronomic managements (Table 1, Fig. 1, Appendix S3). The *Traits* table contains both measurements at the plant level (traits) and ecosystem level (services), whereas the *Yields* table is dedicated to storing above ground biomass measurements of key candidate bioenergy crops (Fig. 1, Table 2, Appendix S3). *Traits* can also be linked to *Covariates*, *Methods*, and *Entities*. *Covariates* provide information required to interpret a measurement, for example, leaf temperature or depth of a soil core sample. *Methods* describes specific methods, protocols, and instruments with a citation. *Entities* groups related measurements within the same unit of study and can be hierarchically structured, for example, to track a leaf on a plant or a plant in a plot (Fig. 1, Table 1).

Managements quantifies experimental treatments and agronomic inputs. For example, the dates and amounts of fertilizer applied, or the date and density of planting, are contained here (Table 1). Whereas the *Managements* table provides more detail about the timing and quantity of an agronomic or experimental condition, *Treatments* identifies experimental treatments by name (i.e., a single experimental treatment may involve multiple management actions). *Managements* are linked to *Treatments*, and *Treatments* are linked to *Traits* and *Yields*.

Species is derived from the USDA Plants database (NRCS 2014), but includes additional species, varieties, and interspecific hybrids of biomass crops; *Cultivars* stores information about cultivars and defined genotypes (Brickell & ICNCP, 2009). While the taxonomic classification of plants is formally defined (Brickell & ICNCP, 2009), classifications based on the ecological concept of functional relatedness depend on the scope and context of a particular analysis. Plant functional types (PFTs) are groups of plants that share functional characteristics, such as traits or habitats (PFTs, Smith *et al*., 1997). Within BETYdb, *PFTs* apply more generally to any grouping of *Species* or *Cultivars*. Thus, *PFTs* can be used in the traditional sense to represent a broad group of functionally related plants (e.g., perennial grasses), but they can also represent a clade (e.g., *Panicum*), a single species (e.g., *Panicum virgatum*), or a subspecific grouping such as cultivar (e.g., *P. virgatum* cv. Alamo) or ecotype (e.g., 'upland' switchgrass). Users are able to use, copy, and modify existing PFTs or define new PFTs to address a specific scientific question. The use of PFTs supports the hierarchical analysis of traits within and across groups and species, as demonstrated in the trait meta-analysis below.

Table 1 Key tables in BETYdb, including names, key fields, content, and use. See also Fig 1 and Appendix S3

Table	Key Fields	Contents	Use
Traits, Yields	Mean, sample size, variance estimate, date, time, site*, citation*, species*, treatment*	Trait, yield, and ecosystem service data, including values and summary statistics	Stores primary data
Variables	Name, definition, units	Definitions, descriptions, units, and allowable ranges of specific traits and ecosystem services data contained in the database	Defines primary data, covariates, and priors
Covariates	Variable*, level	Context required to interpret a particular data point, for example the time, temperature, or location of a measurement	Provides contextual information such as temperature at time of measurement
Plant functional types	Name, definition, citation*	Functionally related species and cultivars	Associates-related cultivars and species for use in data synthesis including data summaries, crop model parameterization, and QA/QC
Species	Scientific name	USDA Plants database, amended with additional species	Defines species-level taxonomies. Links to traits, yields, PFTs, and cultivar tables
Cultivars	Species*, name, citation	Specific genotype bred for cultivation	Provides taxonomic resolution below the level of species, including cultivars, varieties, and genotypes
Priors	Variable*, citation*, phylogeny, distribution	Probability distributions that quantify knowledge of a variable in the absence of information at the level of functional type, species, or cultivar	Provides estimates of trait values based on quantitative syntheses and expert knowledge, and supports QA/QC and data analysis
Treatments	Name, definition	Qualitative descriptions of treatments described in the primary publication	Categorizes experimental treatments (treatments are quantified by the management table)
Managements	Date, citation*, type, level, units	Quantitative record of management activities performed on all plots or specific experimental interventions	Defines dates of planting, harvest, and other farm management. Provides information about the timing and quantity of experimental treatments
Sites	Name, latitude, longitude	Location and basic climate and soil information. Location is typically stored as a point, but can also be stored as a polygon or other shape	Often a point (latitude and longitude) but is also used with bounding boxes to define field boundaries and plots within a field. Enables geospatial queries and joins with external data as well as hierarchical or nested plots
Citations	Author, year, title, doi	Unique reference for source of information, not necessarily published	Used in many tables to record source of data or information – not necessarily published
Entities	Parent, name, description	Links-related trait records	Identifies measurements made on the same unit of replication (e.g., plant within a plot, or plot within a block)

*Denotes fields in one table that reference another table of the same name.

Sites stores information about the location of a measurement. Typically, a site is defined by a latitude and longitude point, but can also be a polygon or other geospatial object. For example, data can be stored at plot scale and then aggregated to field scale. Sites allow data within BETYdb to be evaluated in the context of climate, soil, remote sensing, and other geospatial databases.

Citations uniquely identifies the source of data, for example, a publication or point of contact; each record in *Traits*, *Yields*, *Sites*, and *Managements* is linked to a record in *Citations*.

Data entry

The protocols for data organization and entry are defined in the data entry workflow (Appendix S1). This

Table 2 The number of trait and yield records for each of ten genera under consideration for bioenergy feedstocks production, followed by the total number of records divided between these genera and others not actively undergoing biomass research and the total number of records

Genus	Traits	Yields	Total
Saccharum	1579	3578	5157
Populus	3226	865	4091
Miscanthus	2666	1021	3687
Panicum	629	2087	2716
Salix	1208	532	1740
Pinus	1466	6	1472
Acer	1142	3	1145
Quercus	882	18	900
Liquidambar	486	10	496
Agave	380	10	390
Biomass crops	13 664	8130	21 794
Other species	18 285	310	18 595
Total	31 949	8440	40 389

workflow uses a sequence of web pages that standardize the entry and description of data from heterogeneous sources (Fig. 2). These pages guide the user through the process of identifying, organizing, and entering metadata into the *Citations*, *Sites*, *Treatments*, and *Managements* tables. Then, trait and yield data can either be entered through additional web forms, through a bulk upload interface, or via the application program interface (API, Appendix S2).

The data entry workflow is designed to facilitate the integration of data from many different sources. Sources of data include technical publications, existing databases, the results of previous syntheses, and primary data provided by individual researchers and consortia. To initiate the database, we located previously published data, journal, and book articles listed in the Web of Knowledge and Google Scholar containing relevant data about Miscanthus, switchgrass, prairie grasses, willow, Agave, and poplar (Miguez *et al.*, 2009; Wang *et al.*, 2010, 2013a, 2015; Davis *et al.*, 2014). Less extensive searches have been conducted for other species (Table 2). Data and metadata were extracted from these figures and tables in these publications, organized into spreadsheets, and entered or uploaded into BETYdb. Where necessary, authors were contacted for clarification and additional data.

Database contents

To date, BETYdb contains over forty thousand trait and yield data records along with experimental metadata. Reflecting the focus of research to date, approximately half of these records come from six key genera being evaluated as lignocellulosic feedstocks (Table 2, Fig. 3a, b) and the other half represent over two thousand distinct species. Over 99% of the yield data and 70% of the trait data for feedstock genera are publicly available. Over 80% of the records are plant-level traits and ecosystem-level processes, although for key bioenergy crops, there are more yield records than there are for any single trait. Plant-level traits include – but are not limited to – measures of photosynthesis, respiration, plant architecture, growth dynamics, partitioning between plant organs, chemical composition, and phenology. Ecosystem-level measures include pools and fluxes of water, carbon, nitrogen, and energy.

Quality assurance and control

Validation and quality control are performed at multiple points in the process of data entry, harmonization, and use. The first validation step is performed by the software itself: The web-interface and database schema only allow valid data types and biologically meaningful values (Appendix S3). The range of allowed values for each trait is wider than observed values and can be updated in the *Variables* table if necessary. For example, the lower bound on biomass is zero, and the upper bound is twice the maximum validated observation of standing biomass. Next, the *Traits* and *Yields* tables contain a 'checked' field that indicates that a record has been independently reviewed after entry. After data pass automated range and type checks, they are entered into the database as 'unchecked'. If a trait or yield data point has been independently compared to the original source, it is either marked as 'passed' or 'failed'. Data that have failed quality assurance and control (QAQC) are not included in data exports. Data review and revision is prioritized by need, and users have the option of downloading 'unchecked' data. Users are encouraged to report errors and suspicious records.

Data access

To meet the needs of its diverse user community, data can be searched directly via the Web interface as well as programmatically using the API or the Postgres SQL server. The global search interface supports exploratory queries and access to the core data and metadata. In addition, each of the database tables can be searched and downloaded independently for more complex queries. The BETYdb web pages also provide a map-based interface for searching crop traits and yields measured at any

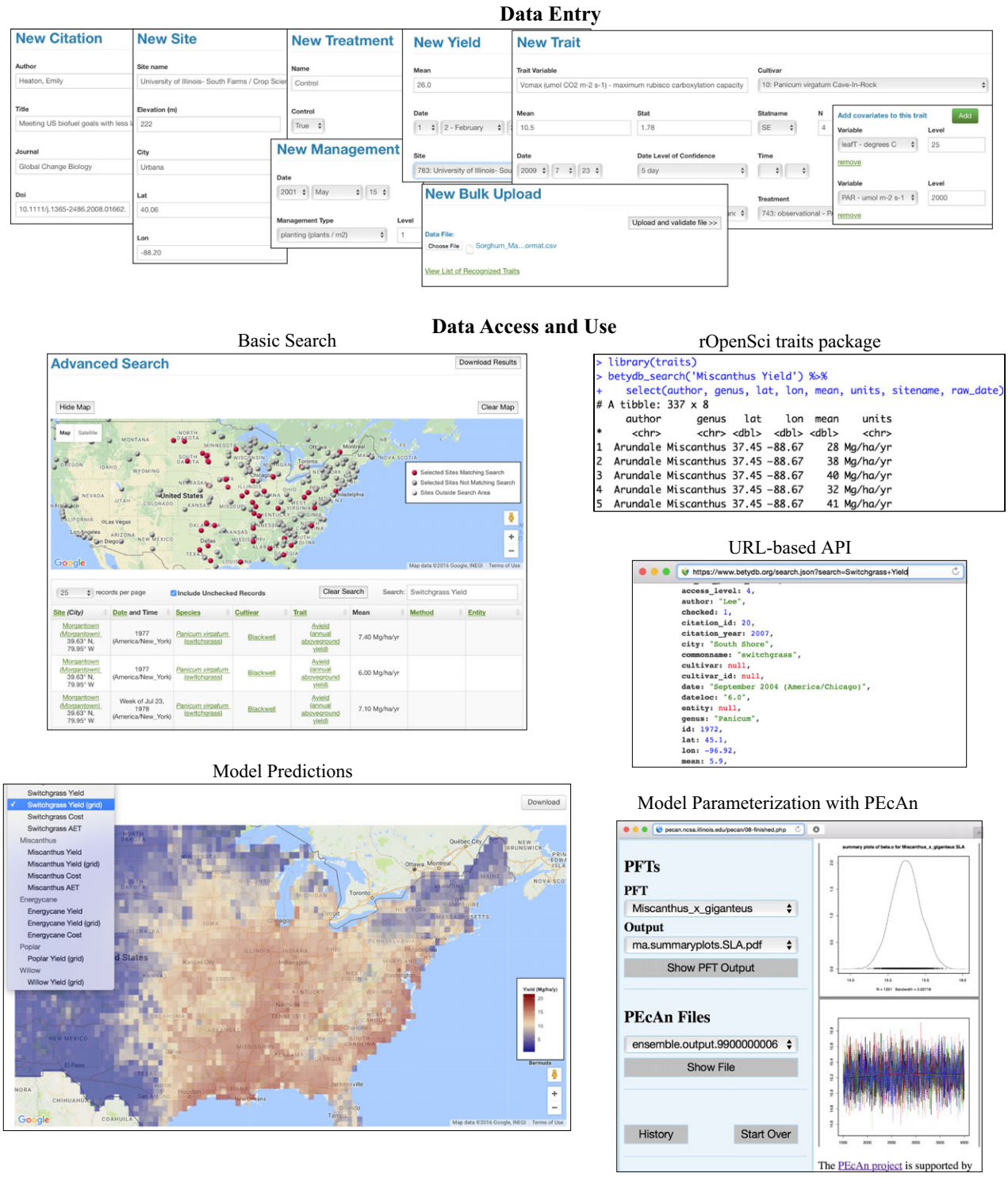

Fig. 2 Screenshots of the user interface showing pages used to enter, search, and download data and model predictions.

geographic location, simply by pointing and clicking. This provides a starting point for comparing modeled and observed yields. The search box returns results that can be exported as comma-separated value (CSV) files. More extensive queries can be submitted directly via the API url in a browser or terminal with results returned in text-based CSV, JSON, or XML formats. The rOpenSci traits package (Chamberlain *et al.*, 2015) uses this API,

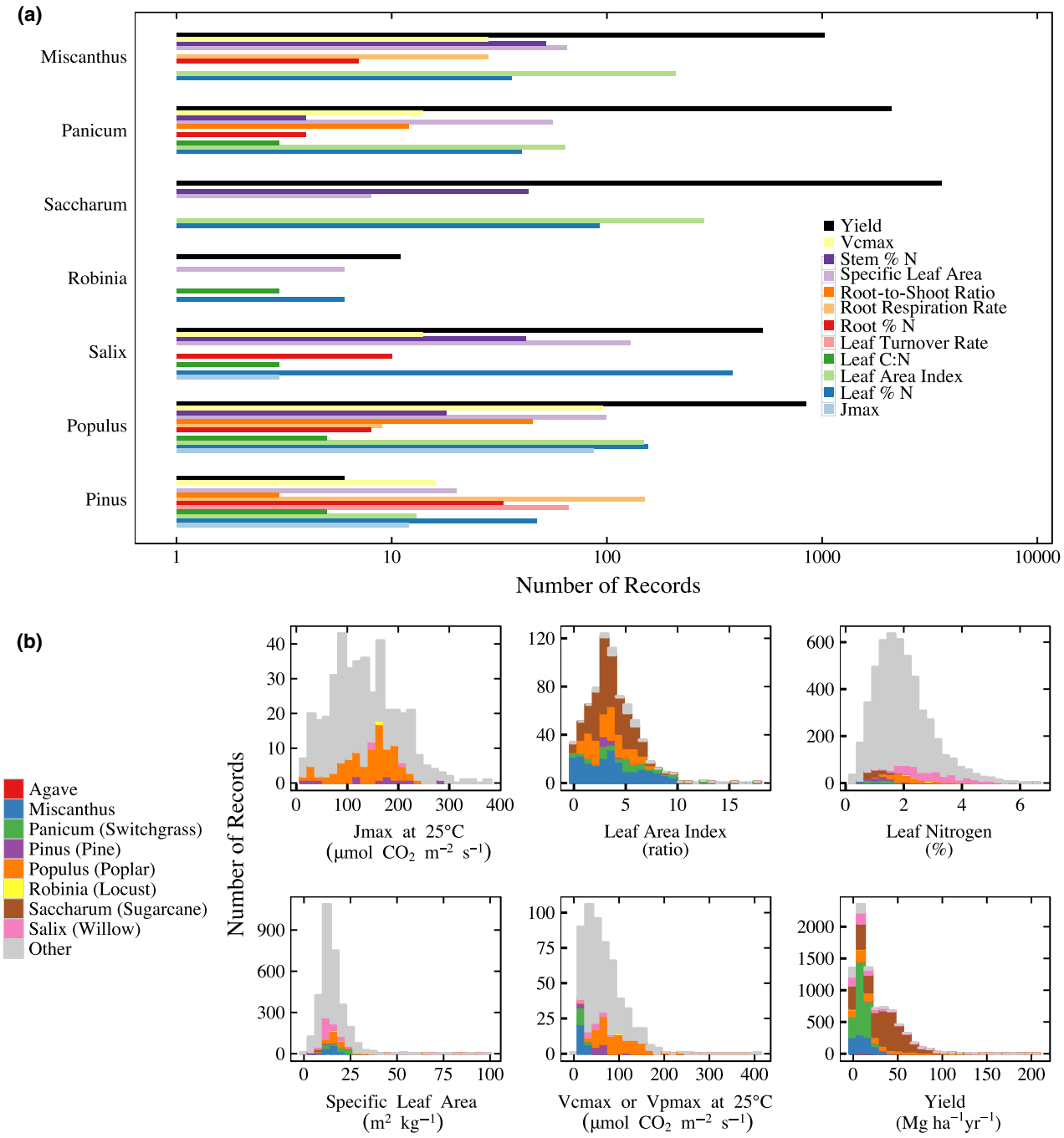

Fig. 3 Summary of available data for a subset of plant genera being evaluated as bioenergy crops and traits related to productivity. (a) Number of records available for yield and eleven physiological traits affecting yield for six genera currently used for biofuel feedstock production. (b) Histograms of yields and key physiological traits related to productivity for all plants (gray) and second-generation lignocellulosic biofuel crops in particular (see legend).

allowing users to query and analyze data from BETYdb entirely within the R programming environment. Furthermore, the entire SQL database is exported daily and made available for download (ebi-forecast.igb.illinois.edu/pecan/dump/betydump.psql.gz). These methods of accessing data are described in Appendix S2.

Example analyses

To illustrate how BETYdb can be used to summarize and evaluate crop physiology and yield potential, we present two new meta-analyses. First, we compare switchgrass and Miscanthus yields and their responses to

Table 3 Meta-analysis results of nitrogen fertilization rate, stand age, summer precipitation, and growing degree days on yield of Miscanthus (top) and switchgrass (bottom)

	Estimate	SE	*T* value	*P*
Miscanthus				
Intercept (Mg ha^{-1} yr^{-1})	9.4	3.5	2.7	0.008
Nitrogen fertilization (Mg ha^{-1} yr^{-1} kg $[N]^{-1}$)	0.028	0.003	8.7	<0.0001
Summer precipitation (Mg ha^{-1} yr^{-1} mm^{-1})	0.014	0.006	2.3	0.02
Growing degree days (Mg ha^{-1} yr^{-1} $°C^{-1}$)	0.0061	0.002	2.8	0.006
Stand age (Mg ha^{-1} yr^{-1} yr $[age]^{-1}$)	−1.4	0.62	−2.3	0.02
switchgrass				
Intercept (Mg ha^{-1} yr^{-1})	4.1	2.6	1.5	0.12
Nitrogen Fertilization (Mg ha^{-1} yr^{-1} kg $[N]^{-1}$)	0.013	0.004	3.0	0.003
Summer precipitation (Mg ha^{-1} yr^{-1} mm^{-1})	0.0081	0.003	2.9	0.004
Growing degree days (Mg ha^{-1} yr^{-1} $°C^{-1}$)	0.0029	0.001	2.7	0.006
Stand age (Mg ha^{-1} yr^{-1} yr $[age]^{-1}$)	−1.2	0.32	−3.8	0.0002

The effect of stand age is evaluated after the third year to allow for establishment.

temperature, precipitation, nitrogen fertilization, and stand age. Second, we summarize and compare four physiological traits related to yield for each of four key bioenergy crops.

Effects of nitrogen, precipitation, temperature, and stand age on yields of switchgrass and Miscanthus

The responses of Miscanthus and switchgrass yield to nitrogen, precipitation, temperature, and stand age across prior studies were first evaluated by Heaton *et al.* (2004). In that study, Miscanthus appeared twice as productive as switchgrass across a wide range of precipitation, temperature, nitrogen fertilization, and stand age over which the species had been studied. At the time, this was surprising as most data were for a single unimproved Miscanthus hybrid, while switchgrass had a longer history of cultivation and of breeding and selection of regionally adapted cultivars. However, at that time there were no side-by-side trials of Miscanthus and switchgrass, and no trials of Miscanthus had been conducted in the United States. In the last decade, many additional trials have been conducted, and now, there are six times as many yield observations for these two species. These new studies include Miscanthus yield trials conducted in the continental United States, many as side-by-side comparisons with switchgrass (Arundale, 2012; Arundale *et al.*, 2014). Here, we use BETYdb and show that this much larger dataset confirms previous findings across a larger, more geographically diverse region.

The statistical analysis follows Heaton *et al.* (2004) to estimate the influence of nitrogen, temperature, summer precipitation, and stand age on yield of Miscanthus and switchgrass. In addition to using the much larger current database, the current analysis differs from the original in the following ways. First, we use temperature and precipitation values for the growing season immediately preceding harvest instead of the long-term annual climatological means used in the original study. These values were extracted from climate reanalyses: MsTMIP (Wei *et al.*, 2013) for data collected through 2010 and DayMet (Thornton *et al.*, 2014) for data collected in 2011. Differences in these two sources were estimated by comparing temperature and precipitation at sites and years for which they were both available, and DayMet values were bias corrected using the slope and intercept of this regression. Second, we included site and year within site as random effects following Wang *et al.* (2010). Finally, we limited the analysis to the range of conditions under which both species were evaluated. The statistical model was fit as a linear mixed-effects model using the R package lme4 (Bates *et al.*, 2015); *P* values were computed from the *t*-statistic.

The positive responses of yield to nitrogen and precipitation for Miscanthus and switchgrass (Fig. 4a, b; Table 3) are consistent with previous meta-analyses (Heaton *et al.*, 2004; Wang *et al.*, 2010) and also with global observations of grassland dynamics (Hooper & Johnson, 1999; LeBauer & Treseder, 2008). The yield of both crops increased with growing degree days, but the response of Miscanthus to temperature was more than double that of switchgrass (Fig. 4c; Table 3).

Yield declined with age following the third year, consistent with Heaton *et al.* (2004). Recent syntheses of Miscanthus yields using nonlinear regression observed that yield declines typically occur around 6 years after planting (Arundale 2012; Lesur *et al.*, 2013). Most of the data from these studies are available in BETYdb, and although the present analysis only tested a linear decline starting in the third year, the nonlinear response of yield to age can be seen in the scatter plot of yield as a function of stand age (Fig. 4d). Incorporating these dynamics into a simulation model would better capture the observed early peaks and the subsequent decline

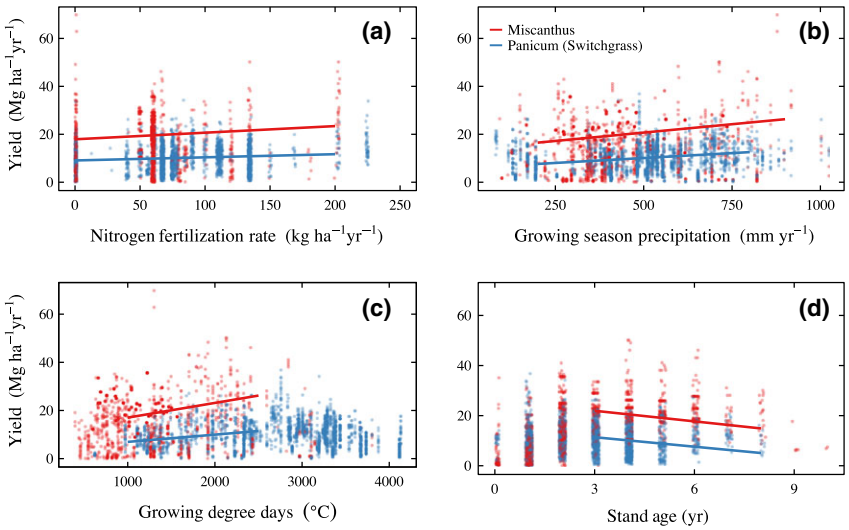

Fig. 4 The response of switchgrass (blue) and Miscanthus (red) yields to (a) nitrogen fertilization, (b) growing season precipitation (mm yr^{-1}), (c) temperature (growing degree days), and (d) stand age. To compare the responses of these crops, the fitted regressions were limited to the range of independent variables where there were data for both species and to records where the covariate of interest was available. Regression statistics are in Table 3.

in yields at some sites. Most importantly, the conclusion of Heaton *et al.* (2004) that Miscanthus out-yields current switchgrass cultivars by about twofold over a wide range of environments, still stands and now has stronger support from over 966 yield estimates across 52 sites (Fig. 4).

Trait meta-analysis: ecophysiological traits of four biofuel crops

We used meta-analysis to summarize available data on three traits across four plant genera currently being evaluated as bioenergy feedstocks. The four genera include two perennial grasses (Miscanthus and switchgrass) and two short rotation coppice shrubs (poplar and willow). The traits evaluated were specific leaf area (SLA), maximum rate of carboxylation (by Rubisco for the two shrubs or PEP carboxylase for the two grasses), and the rate of leaf respiration. SLA is related to the amount of leaf area a plant can produce for a given amount of carbon; the maximum rate of carboxylation sets an upper limit on carbon uptake, and the rate of respiration is the rate of carbon loss measured in the absence of light.

The statistical model included random effects of site, experimental treatment, and a fixed effect to account for growth under controlled conditions using the PEcAn

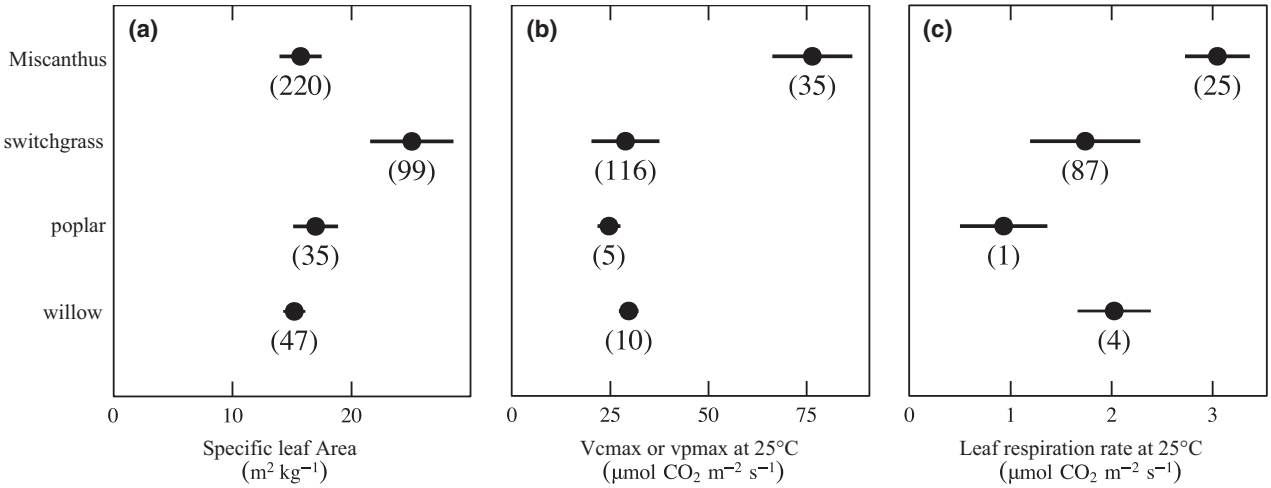

Fig. 5 Estimates of mean values for three physiological traits: (a) specific leaf area, (b) Vpmax or Vcmax, and (c) leaf respiration rate across four genera representing two plant functional types (perennial grass and trees) being evaluated as bioenergy feedstocks. Bars represent ±SD, and sample size is in parentheses.

meta-analysis module (LeBauer *et al.*, 2013). Here, we present the genus level mean trait values (Fig. 5) for each of the three traits.

This analysis shows that even within a functional type, these crops differ in key traits related to productivity. For example, Miscanthus has thicker leaves, higher leaf respiration rate, and higher Vpmax compared to switchgrass. These estimates could be further extended by conducting meta-analysis at the level of species or cultivar, and by comparing estimates of site and treatment effects within and across taxonomic levels.

Discussion

The Biofuel Ecophysiological Traits and Yields database provides a new and comprehensive open-access repository containing data related to the physiology, ecology, and agronomy of terrestrial plant species being used or considered as bioenergy feedstocks. Its harmonized format and Web interface make it simple to enter, view, extract, and analyze data. We have illustrated how BETYdb can be used to identify and evaluate bioenergy feedstocks in terms of potential yield, yield stability, resource use, and ecosystem services. Plant traits underlying yield differences among crop species can be readily extracted and summarized (Fig. 5). Observed yields can be used to predict and compare the potential productivity of promising bioenergy crops under a range of conditions (Fig. 4). Data driven insights into the link between plant traits and crop performance can inform crop improvement, regional distributions, and sustainable management of these crops.

Harmonization of data from disparate sources into a standard format facilitates the comprehensive integration of previous findings into ongoing research (e.g., Figs 4 and 5). The use of meta-analysis to parameterize and constrain crop and ecosystem simulation models is an important advance over the use of fixed parameter values because it provides a statistical framework for synthesis while identifying critical gaps and in turn needs for targeted data collection (LeBauer *et al.*, 2013).

The Biofuel Ecophysiological Traits and Yields database has been used to design field experiments and data collection campaigns, by providing knowledge of levels of variability and key data needs (Davidson, 2012; LeBauer *et al.*, 2013; Wang *et al.*, 2013b; Dietze *et al.*, 2014). Syntheses of data in BETYdb have been used to inform crop selection and farm management decisions by quantifying trade-offs among crops at local and regional scales (Miguez *et al.*, 2012; Wang *et al.*, 2013a,b, 2015; Larsen *et al.*, 2015). Scaling from plant traits to productivity and ecosystem services makes it possible to identify the suite of crop physiological traits associated with suitable biofuel crops and identify traits whose improvement would give the largest benefit in maximizing yield and ecosystem services. Similarly, these data make it possible to identify trait variability that may be exploited in crop improvement, via breeding and engineering, and to identify critical gaps in available data. This can be performed by comparing the geographic and climatological ranges represented by extant field trials with the ranges being considered for production. In addition, the data can be applied more generally in ecosystem science (Davidson, 2012; Dietze *et al.*, 2014), as well as providing parameter estimates and measures of their uncertainty for models of crop productivity (Larsen *et al.*, 2015; Wang *et al.*, 2015). BETYdb also provides access to model results that can be used in downstream analyses of ecosystem services, life cycle analysis, and supply chain optimization (Davis *et al.*, 2009, 2012; Milbrandt *et al.*, 2014; Saha & Eckelman, 2015).

Acknowledgements

This work was supported by the Energy Biosciences Institute and the National Science Foundation under Grants 1062547 and 1457890. The Energy Biosciences Institute was funded by BP America.

References

Arundale R (2012) The higher productivity of the bioenergy feedstock *Miscanthus x giganteus* relative to *Panicum virgatum* is seen both into the long term and beyond Illinois (Doctoral dissertation, University of Illinois at Urbana-Champaign). http://hdl.handle.net/2142/34422.

Arundale RA, Dohleman FG, Heaton EA, Mcgrath JM, Voigt TB, Long SP (2014) Yields of Miscanthus × giganteus and *Panicum virgatum* decline with stand age in the Midwestern USA. *Global Change Biology Bioenergy*, **6**, 1–13.

Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Brickell C, ICNCP (2009) *International Code of Nomenclature for Cultivated Plants: (ICNCP or Cultivated Plant Code) Incorporating the Rules and Recommendations for Naming Plants in Cultivation: Adopted by the International Union of Biological Sciences, International Commission for the Nomenclature of Cultivated Plants*, 8th edn. International Society for Horticultural Science, Leuven.

Chamberlain S, Foster Z, Bartomeus I (2015) traits: Species Trait Data from Around the Web. R package version 0.2.0.

Davidson CD (2012) *The Modeled Effects of Fire on Carbon Balance and Vegetation Abundance in Alaskan Tundra*. University of Illinois, Urbana Champaign.

Davis SC, Anderson-Teixeira KJ, Delucia EH (2009) Life-cycle analysis and the ecology of biofuels. *Trends in Plant Science*, **14**, 140–146.

Davis SC, Dietze M, DeLucia E *et al.* (2012) Harvesting carbon from eastern US forests: opportunities and impacts of an expanding bioenergy industry. *Forests*, **3**, 370–397.

Davis SC, LeBauer DS, Long SP (2014) Light to liquid fuel: theoretical and realized energy conversion efficiency of plants using Crassulacean Acid Metabolism (CAM) in arid conditions. *Journal of Experimental Botany*, **65**, 3471–3478.

Dietze MC, Serbin SP, Davidson C *et al.* (2014) A quantitative assessment of a terrestrial biosphere model's data needs across North American biomes. *Journal of Geophysical Research-Biogeosciences*, **119**, 286–300.

Heaton E, Voigt T, Long SP (2004) A quantitative review comparing the yields of two candidate C 4 perennial biomass crops in relation to nitrogen, temperature and water. *Biomass and Bioenergy*, **27**, 21–30.

Hooper DU, Johnson L (1999) Nitrogen limitation in dryland ecosystems: responses to geographical and temporal variation in precipitation. *Biogeochemistry*, **46**, 247–293.

Karp A, Shield I (2008) Bioenergy from plants and the sustainable yield challenge. *New Phytologist*, **179**, 15–32.

Kattge J, Diaz S, Lavorel S *et al.* (2011) TRY – a global database of plant traits. *Global Change Biology*, **17**, 2905–2935.

Larsen S, Jaiswal D, Bentsen NS, Wang D, Long SP (2015) Comparing predicted yield and yield stability of willow and Miscanthus across Denmark. *Global Change Biology Bioenergy*, **8**, 1061–1070.

LeBauer DS, Treseder KK (2008) Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. *Ecology*, **89**, 371–379.

LeBauer DS, Wang D, Richter KT, Davidson CC, Dietze MC (2013) Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs*, **83**, 133–154.

Lesur C, Jeuffroy MH, Makowski D, *et al.* (2013) Modeling long-term yield trends of *Miscanthus x giganteus* using experimental data from across Europe. *Field Crops Research*, **149**, 252–260.

Lewandowski I, Scurlock JMO, Lindvall E, Christou M (2003) The development and current status of perennial rhizomatous grasses as energy crops in the US and Europe. *Biomass and Bioenergy*, **25**, 335–361.

Long SP, Karp A, Buckeridgec MS, Murphyf DJ, Onwona-Agyemang S, Vonshakh A (2015) Feedstocks for biofuels and bioenergy. In: *SCOPE Bioenergy and Sustainability* (eds Souza GM, Victoria R, Joly C, Verdade L), pp. 302–347. Scientific Committee on Problems of the Environment (SCOPE), Paris, France.

Miguez FE, Zhu X, Humphries S, Bollero GA, Long SP (2009) A semimechanistic model predicting the growth and production of the bioenergy crop Miscanthus × giganteus: description, parameterization and validation. *Global Change Biology Bioenergy*, **1**, 282–296.

Miguez FE, Maughan M, Bollero GA, Long SP (2012) Modeling spatial and dynamic variation in growth, yield, and yield stability of the bioenergy crops Miscanthus × giganteus and *Panicum virgatum* across the conterminous United States. *Global Change Biology Bioenergy*, **4**, 509–520.

Milbrandt AR, Heimiller DM, Perry AD, Field CB (2014) Renewable energy potential on marginal lands in the United States. *Renewable and Sustainable Energy Reviews*, **29**, 473–481.

NRCS, U (2014) The PLANTS Database. National Plant Data Team, Greensboro, NC. Available at: http://plants.usda.gov (accessed 9 February 2014).

Saha M, Eckelman MJ (2015) Geospatial assessment of potential bioenergy crop production on urban marginal land. *Applied Energy*, **159**, 540–547.

Smith TM, Woodward FI, Shugart HH (1997) *Plant Functional Types: Their Relevance to Ecosystem Properties and Global Change*. Cambridge University Press, Cambridge, New York.

Somerville C, Youngs H, Taylor C, Davis SC, Long SP (2010) Feedstocks for lignocellulosic biofuels. *Science*, **329**, 790–792.

Surendran Nair S, Kang S, Zhang X *et al.* (2012) Bioenergy crop models: descriptions, data requirements, and future challenges. *Global Change Biology Bioenergy*, **4**, 620–633.

Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB (2014) Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. Oak Ridge National Laboratory (ORNL).

Wang DAN, LeBauer DS, Dietze MC (2010) A quantitative review comparing the yield of switchgrass in monocultures and mixtures in relation to climate and management factors. *Global Change Biology Bioenergy*, **2**, 16–25.

Wang D, LeBauer D, Dietze M (2013a) Predicting yields of short-rotation hybrid poplar (*Populus* spp.) for the United States through model-data synthesis. *Ecological Applications*, **23**, 944–958.

Wang D, LeBauer D, Kling G, Voigt T, Dietze MC (2013b) Ecophysiological screening of tree species for biomass production: trade-off between production and water use. *Ecosphere*, **4**, 1–22.

Wang D, Jaiswal D, LeBauer DS, Wertin TM, Bollero GA, Leakey AD, Long SP (2015) A physiological and biophysical model of coppice willow (*Salix* spp.) production yields for the contiguous USA in current and future climate scenarios. *Plant, Cell and Environment*, **38**, 1850–1865.

Wei Y, Liu S, Huntzinger D *et al.* (2013) The North American carbon program multi-scale synthesis and terrestrial model intercomparison project–part 2: environmental driver data. *Geoscientific Model Development Discussions*, **6**, 5375–5422.

Yang X, Cushman JC, Borland AM *et al.* (2015) A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist*, **207**, 491–504.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. BETYdb documentation: data entry workflow.

Appendix S2. BETYdb documentation: data access.

Appendix S3. BETYdb documentation: technical guide to structure, development, and deployment of BETYdb.