*Journal of Vegetation Science* **26** (2015) 828–838

# Estimating the missing species bias in plant trait measurements

Brody Sandel, Alvaro G. Gutiérrez, Peter B. Reich, Franziska Schrodt, John Dickie & Jens Kattge

**Keywords**
Bias; Leaf nitrogen; Maximum height; Missing data; Photosynthesis rate; Plant functional trait; Seed mass; Specific leaf area

**Abbreviations**
SLA = Specific leaf area; CWM = community-weighted mean; $N_{mass}$ = leaf nitrogen per unit mass; $A_{area}$ = maximum leaf photosynthesis rate per unit area; NE = naïve estimate; BCE = bias-corrected estimate.

Received 21 October 2014
Accepted 13 March 2015
Co-ordinating Editor: Sebastian Schmidtlein

**Sandel, B.** (Corresponding author, brody.sandel@biology.au.dk)[1],
**Gutiérrez, A.G.** (bosqueciencia@googlemail.com)[2,7],
**Reich, P.B.** (preich@umn.edu)[3,4],
**Schrodt, F.** (fschrodt@bgc-jena.mpg.de)[5],
**Dickie, J.** (j.dickie@kew.org)[6],
**Kattge, J.** (jkattge@bgc-jena.mpg.de)[5]

[1]Department of Bioscience, Aarhus University, Ny Munkegade 114, 8000 Aarhus C, Denmark;
[2]Instituto de Conservación Biodiversidad y Territorio, Facultad de Ciencias Forestales y Recursos Naturales, Universidad Austral de Chile, Valdivia, Chile;
[3]Department of Forest Resources, University of Minnesota, 115 Green Hall, 1530 Cleveland Ave. N., St. Paul, MN 55108, USA;
[4]Hawkesbury Institute for the Environment, University of Western Sydney, Penrith, NSW 2751, Australia;
[5]Max Planck Institute for Biogeochemistry, P.O. Box 10 01 64, 07701 Jena, Germany;
[6]Seed Conservation Department, Royal Botanical Gardens, Kew, Wakehurst Place, Ardingly, West Sussex RH17 6TN, UK;
[7]ETH Zürich, Institut für Terrestrische Ökosysteme, CHN G 76.2, Universitätstrasse 16, 8092 Zürich, Switzerland

## Abstract

**Aim:** Do plant trait databases represent a biased sample of species, and if so, can that bias be corrected? Ecologists are increasingly collecting and analysing data on plant functional traits, and contributing them to large plant trait databases. Many applications of such databases involve merging trait measurements with other data such as species distributions in vegetation plots; a process that invariably produces matrices with incomplete trait and species data. Typically, missing data are simply ignored and it is assumed that the missing species are missing at random.

**Methods:** Here, we argue that this assumption is unlikely to be valid and propose an approach for estimating the strength of the bias regarding which species are represented in trait databases. The method leverages the fact that, within a given database, some species have many measurements of a trait and others have few (high vs low measurement intensity). In the absence of bias, there should be no relationship between measurement intensity and trait values. We demonstrate the method using five traits that are part of the TRY database, a global archive of plant traits. Our method also leads naturally to a correction for this bias, which we validate and apply to two examples.

**Results:** Specific leaf area and seed mass were strongly positively biased (frequently measured species had higher trait values than rarely measured species), leaf nitrogen per unit mass and maximum height were moderately negatively biased, and maximum photosynthetic capacity per unit leaf area was weakly negatively biased. The bias-correction method yielded greatly improved estimates in the validation tests for the two most biased traits. Further, in our two applications, ecological interpretations were shown to be sensitive to uncorrected bias in the data.

**Conclusions:** Species inclusion in trait databases appears to be strongly biased in some cases, and failure to correct this can lead to incorrect conclusions.

## Introduction

Plant functional traits are finding increasing use in a wide range of ecological fields (McGill et al. 2006; Westoby & Wright 2006). They have been used to understand major axes of plant strategies (Reich et al. 1997; Westoby 1998; Wright et al. 2004), to shed light on community assembly mechanisms (Weiher & Keddy 1995; Kraft et al. 2008;


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science

Cornwell & Ackerly 2009), to parameterize dynamic vegetation models (Kattge et al. 2009; Bonan et al. 2012) and to predict community composition or change (Diaz et al. 2001; Sandel & Dangremond 2012; Frenette-Dussault et al. 2013). This has led to increasing efforts to establish plant trait databases, notably TRY – a global plant trait database (Kattge et al. 2011).

Despite being the world's largest database of plant traits, TRY nevertheless lacks measurements on almost all species for almost all traits. The database contains approximately 5.6 million records across about 100 000 plant species and 1100 distinct traits (status as of 4 June 2014). Thus, about a quarter of the world's plant species are represented, and for those, the trait matrix is 1.5% filled. A well-studied trait, specific leaf area (SLA), has 137 689 records for 14 754 species, roughly 3.5% of global plant diversity. Thus, missing data are a ubiquitous feature of trait-based studies, and despite substantial effort applied towards both data collection and sharing, this situation is unlikely to change in the foreseeable future.

The impact of missing data on trait-based analysis is receiving increasing attention (Pakeman & Quested 2007; Pakeman 2014; Taugourdeau et al. 2014). In one common application, the calculation of community-weighted means (CWM), it has been suggested that values will be reliable if the species with known trait values constitute at least 80% of the cover or biomass of the community (Garnier et al. 2004; Pakeman & Quested 2007). Other metrics of interest, such as some functional diversity metrics, are highly sensitive to even small proportions of missing species (Pakeman 2014). Concern about the impact of missing data has led to the development of gap-filling techniques (Taugourdeau et al. 2014), such as simply filling missing species with the mean of non-missing species, or more complex approaches, such as hierarchical probabilistic matrix factorization (Shan et al. 2012) and phylogeny-based methods (Swenson 2014).

These methods, however, still depend on the sample of species with trait measurements being a representative subset of the larger population of species. Unfortunately, this may not be the case – there are a range of factors that could result in a biased selection of species in a trait database. Within particular sites, researchers may have a preference for targeting highly apparent species, much as the Ecological Apparency Hypothesis (EAH) predicts for ethnobotanical use (Albuquerque & Lucena 2005). Thus, we should expect to see an over-representation of large, abundant and long-lived species in trait databases. At a broader scale, researchers are likely to have geographic and habitat biases, such as biases towards grasslands over forests because of their fast dynamics and ease of experimental manipulation. Researchers may avoid or prefer 'strange' habitats, including unusual edaphic conditions, ruderal communities or otherwise extreme conditions (e.g. the 'botanist effect' on species richness; Moerman & Estabrook 2006). Combined with geographic, phylogenetic or among-habitat trait differences, these biases can lead to biased representation of species in a trait database. As a simple example, if forests contain more tall species than grasslands and researchers have a preference for working in grasslands, then data collection will be biased towards short species. Similarly, roadside ruderal plants may have relatively high SLA values (Pierce et al. 2013), and by virtue of being easy to collect, may be disproportionately represented in databases. In contrast, stress-tolerating species, with a tendency for high SLA, may be underrepresented.

Such biases would have consequences for many applications. For example, if missing species differ consistently from measured species, variation in CWMs could reflect variation in trait coverage rather than a true change in functional composition. Despite this, we are aware of no study to date that has attempted to quantify the magnitude of a bias. Here, we use the TRY database to attempt to estimate the direction and strength of the missing species bias for several important plant traits. The bias estimation method leads naturally to a bias correction method that we validate and apply to two example data sets – a global plant distribution data set and a set of vegetation plots.

## Methods

### Data

All plant trait data used here were obtained from the TRY database (details in Kattge et al. 2011). We extracted values for five traits: specific leaf area (SLA), seed mass, leaf nitrogen (N) per unit mass ($N_{mass}$), maximum height (Height) and maximum leaf photosynthesis rate per unit area ($A_{area}$) (data download: 5 Oct 2013). These traits were selected because they are widely used and reasonably well represented in the TRY database. We also used plant growth form information from TRY. In most cases, we used five categories: Tree, Shrub, Graminoid, Herb and Other. For species with more than one potential class (e.g. 'Shrub/Tree' or 'Herb/Shrub'), we used a 'maximum potential' standard, selecting the largest of the possible growth forms (in this case 'Tree' and 'Shrub', respectively). Species with unknown growth forms were classified as 'Other', a category which also included ferns (and their allies) and mosses. For full information on data sources in the TRY data, see Appendix S2.

TRY is a continuously growing database subject to ongoing cleaning. Here, we cleaned the data in two ways. First, we removed duplicate records. Second, we only kept trait records that gave either a measure of an individual or of the central tendency of a set of individuals (e.g.

median, mean). We excluded measures marked as 'maximum', 'minimum' or similar. For SLA, this resulted in 47 050 measurements across 8396 species; for seed mass 61 966 measurements across 26 107 species; for $N_{mass}$ 32 495 measurements across 6741 species; for height 24 339 measurements across 9817 species; and for $A_{area}$ 3374 measurements across 1149 species. All raw trait values were log-transformed. Analyses for each trait were conducted independently (for example, no use was made of trait–trait correlations).

### Estimating trait values for missing species

In outline, our approach proceeds in three basic steps:

1. Convert the raw trait measurement matrix into a matrix containing each species' mean trait value across all of its measurements, and the number of measurements on that species (measurement intensity).
2. Divide species into sets based on their measurement intensities and calculate the mean trait value and median measurement intensity within each set.
3. Fit a curve to the relationship between this median measurement intensity and mean trait value across all sets and project the curve to a measurement intensity of zero to estimate the trait value of unmeasured species.

More formally, say that we have measurements of a trait τ for some set of species. Let $n_i$ denote the number of measurements of τ on species $i$, and $t_i$ denote the mean of those measurements. $S_k$ will indicate the set of species with $k$ measurements of τ. Finally, we define a function $T$ such that $T(S_k)$ is the mean value of $t$ for all species in $S_k$.

Roughly, our approach involves estimating the relationship between $k$ and $T(S_k)$, and projecting that relationship to $k = 0$ to estimate the mean value of τ for species in $S_0$. In practice, there are a few complications to this process. First, in real data sets, the number of species in $S_k$ will become small as $k$ becomes large (there are few species with exactly 47 SLA measurements, for example). Hence, we defined a minimum set size and combined small sets together until they reached this size (Fig. 1). There are various possibilities for defining minimum set size. Here we used the square root of the total number of species with trait measurements as this balances set size with the number of sets. Let $U_i$ indicate the $i$th set of species obtained in this way. Usually, $U_1 = S_1$ and $U_2 = S_2$ (because $S_1$ and $S_2$ are already large enough), while for larger $i$, $U_i$ will contain the union of several distinct $S$ sets (Fig. 1). By analogy to the function $T$ defined above, let $N(U_i)$ indicate the median value of $n$ for species in $U_i$. We can now focus on understanding the relationship between sqrt($N(U_i)$) and $T(U_i)$, towards the goal of predicting the value of $T(S_0)$.

We described this relationship by fitting a spline. The spline was allowed to have 1 *df* for every three $U$ sets,

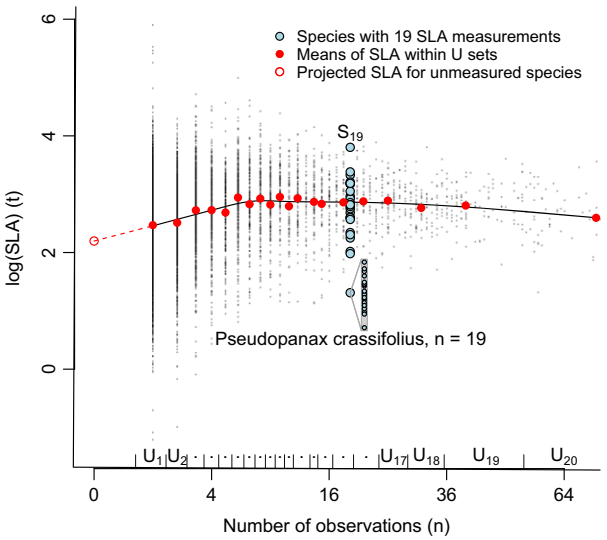


**Fig. 1.** Illustration of the bias-correction method. Each species (small dots) has a certain number of SLA observations ($n_i$) and a trait mean ($t_i$). For example, *Pseudopanax crassifolius* has 19 measurements (individual observations shown as small blue circles at an arbitrary position on the $x$ axis), with a mean value of 1.31. *P. crassifolius* and all other species with exactly 19 measurements (25 species total) constitute the set $S_{19}$. Sets are aggregated together to avoid very small set sizes. In this case, *P. crassifolius* and all other members of $S_{19}$ end up in set $U_{15}$. For each such $U$ set, the mean value of $t_i$ and the median of $n_i$ was then calculated for all species in the set. A spline was fitted to these points (black line) and projected to 0, yielding the estimate of SLA for unmeasured species.

although of course other definitions are possible, yielding more or less flexible fits to the data. Finally, we projected that spline to 0 to estimate the mean value of τ for all species with no observations ($T(S_0)$). We call this value the bias-corrected estimate (BCE) of $T(S_0)$ and compare it to the naïve estimate (NE) obtained by assuming that $T(S_0) = T(S_{>0})$ (i.e. that species with no observations have the same mean value of τ as all species with at least one observation). An illustration of the entire procedure is provided as Fig. 1.

Variations on the BCE and NE can be obtained by generating each one independently for subsets of the species list and then combining them afterwards. For example, plants with different growth forms may have different bias strengths, and also different measurement intensities. Taking these into account could allow better bias correction when the group memberships of missing species are known. Here, we examined growth form specific trait value and bias estimates, using growth form descriptions in the TRY data. We abbreviate these as $NE_{gf}$ and $BCE_{gf}$.

### Researcher degrees of freedom

There are three important decisions in this process that influence the inferred strength of the bias. These are (1)


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science

the flexibility of the spline fit, (2) the target size of species sets and (3) how one transforms the measurement intensity axis. For the case of SLA, we explored the sensitivity to these decisions, asking how much BCEs vary with different possible decisions, and comparing their validation performance using Method 1 (see below).

### Validating trait estimates

We used two methods for validating the bias correction method. In the first, we assessed how well we could predict the mean value of τ for species with one measurement ($T(S_1)$) from all species with more than one measurement (Method 1). In the second, we discarded half of the trait measurements and assessed how well we could predict the value of τ for species that were no longer in the data set (Method 2). We then compared the four resulting estimates (NE, BCE, $NE_{gf}$ and $BCE_{gf}$) to the true mean value. The growth form-adjusted estimates were obtained by estimating trait means for each growth form separately. The overall missing species average was then taken to be the mean of these estimates, weighted by the proportions of each growth form among the unknown species (simulating a common situation in which growth forms are known for all species, but quantitative traits are not).

### Global bias estimate

For each of the five traits (SLA, seed mass, $N_{mass}$, maximum height and $A_{area}$), we obtained an overall bias estimate considering the entire TRY database. Here, growth form correction was not possible, because the growth form composition of the missing species was not known. We defined the strength of the missing species bias (β) as the difference between the NE and the BCE, divided by the SD of the mean ($t$) across all measured species. This gives the strength of the bias relative to the interspecific trait variation. We validated these estimates using both Methods 1 and 2. Growth form correction was possible in the validations, where the growth forms of the 'missing' species are known. As described above, we divided plants into five growth forms: Herbs, Graminoids, Shrubs, Trees and Other, based on growth form descriptions in the TRY data.

### Regional bias estimates

In both this example and the next, we focused only on SLA because SLA was the most biased trait (see Results) and is very widely used. For each plant species with SLA measurements, we downloaded from GBIF (gbif.org, on 24 Feb 2014) a 1° resolution map of its occurrence record density. Of 8396 species with SLA data, 7591 had occurrence records in GBIF (90.4%). These maps were then used to determine the presence or absence of each species in each cell of a 10° global grid. A very coarse grid was used to minimize errors of omission due to the extreme spatial heterogeneity of sampling intensity in GBIF. Nevertheless, many species are poorly represented in GBIF, leading to incorrect representations of their range even on this coarse grid. Further, specimens in GBIF frequently include incorrect coordinates (e.g. flipping the hemisphere of the record). Visual checks of 500 of the range maps indicated suspicious single-cell disjunctions in about 15% of cases. Thus, this example should be taken as illustrative rather than as a definitive map of species distributions with respect to SLA.

We estimated the mean value of log(SLA) for unmeasured species within each grid cell, using only the species in that cell (considering only those grid cells with records for at least 100 species). Note that this inference is possible despite not having a list of the unmeasured species within the cell. For each cell, we calculated the NE and the BCE, and then validated the estimates using Method 1.

### A local example: Yosemite vegetation plots

To complement the coarse global example, we also selected a high-quality local data set of vegetation plots. We downloaded vegetation plots from vegbank (vegbank.org) for the Yosemite area of California ($n$ = 604 plots). Of the 1222 species occurring in these plots, TRY contained SLA values for 237 of them (19.4%). These 237 tended to be the more abundant species, resulting in a higher average trait coverage per plot (39.9%). We used these 237 species to obtain NE, $NE_{gf}$, BCE and $BCE_{gf}$ values for the missing 985 species. In contrast to the first example where BCE values were obtained within each region using only species from that region, we could not do the same here with plots because most plots have too few species. As with the regional example, we validated the estimates using Method 1. Growth forms here provided a coarser grouping than in the global estimation, since we had fewer species and some growth forms had very few representatives. Hence, we divided species into Graminoid, Herb/Other and Shrub/Tree. Here, Other consisted of six fern species and one of unknown form.

This local example differs from the global bias estimate and regional bias estimates described above in that the set of missing species is known. This means that it is possible to use the growth form correction to estimate the missing species bias, and that it is possible to use the missing species bias correction to update CWM estimates because the abundances of all species are known. We selected the least biased estimation method and used it to create improved estimates of SLA CWM for each plot. Finally, we asked if the ecological interpretation of SLA patterns changed


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science 

when bias correction was applied. This could be expected if there is a relationship between environmental variables and trait coverage, for example. We considered two simple environmental variables: mean annual temperature and annual precipitation, extracted for each plot from worldclim (Hijmans et al. 2005) data. Relationships were examined with simple linear models.

## Results

Of the five traits considered, two showed strong biases ($\beta \approx 0.6$, SLA and seed mass; Fig. 2, Appendix S1), two showed moderate biases ($\beta \approx -0.3$, $A_{area}$ and Height; Appendix S1), and one showed very little bias ($\beta = -0.017$, $N_{mass}$; Appendix S1, Table 1). For the most strongly biased trait, SLA, the mean of all species was 2.618 $\log(mm^2 \cdot mg^{-1})$, while the estimated mean of missing species was 2.197, amounting to a bias (β) of approximately 0.63 SD, with respect to the distribution of observed species mean SLA values. SLA and seed mass were positively biased (i.e. we estimated that unmeasured species have lower SLA and seed mass than measured species), while the other three traits were negatively biased (Table 1).

The two validation approaches generally agreed regarding which of the four estimates was most accurate (Table 1). Overall, $BCE_{gf}$ provided the best estimate for SLA, BCE was best for seed mass and $NE_{gf}$ was best for $A_{area}$, Height and $N_{mass}$. Thus, for the two most strongly biased traits, a bias-correction method provided the best estimates, while naïve estimates were preferable when the bias was weak (e.g. $N_{mass}$). However, the NE was never best – it was always preferable to at least take account of the growth form of the missing species when it is available (Table 1).

Decisions made in the bias estimation process have some influence over the BCE, but the general conclusion that the BCE and NE differ substantially was robust (Appendix S1). Across a wide range of minimum group sizes (25 species to 400 species) and spline *df* (3, 6 and 9), the average BCE for log(SLA) was 2.163, with a SD of 0.073, as compared to the NE of 2.618 (Appendix S1). Using a $\log(x + 1)$ transformation of measurement intensities (rather than square root) led to a slight increase in the BCE ($2.243 \pm 0.054$). Across the full range of variations, validation with Method 1 indicated a reduction in error from 0.254 (NE) to $-0.070 \pm 0.058$.

Within $10 \times 10°$ regions across the world, the estimated bias in SLA was overwhelmingly positive (β range −0.28 to 1.43 SD, 91% positive; Fig. 3). The geographic pattern of bias was complex, with hotspots in western Australia, south-central Asia, South Africa, western North America and southwestern South America. Validation with Method 1 indicated that the BCE was nearly always a better estimate than the NE (94.5% of cases; Fig. 4). Further, and as expected, the NEs were systematically biased towards too-high values (in 87.4% of cases, median error = 0.290 SD; Fig. 4). The BCEs removed the bias, with 45.9% of estimated values being too high, and a median error of −0.023 SD.

**Table 1.** Bias-correction estimates and validation for mean values of SLA, seed mass, $N_{mass}$, Height and $A_{area}$. The top section of the table shows the estimated trait mean of all missing species, either by assuming that they are an unbiased subset of the global pool (naïve, NE) or using the bias-correction method (BCE). The bias is given as the difference between the NE and BCE divided by the SD of observed trait values. The bottom sections of the table show results for two validation approaches across four different estimation methods (NE, NE weighted by growth form ($NE_{gf}$), BCE and BCE weighted by growth form ($BCE_{gf}$)). For each test the best estimate (i.e. the value closest to the observed) is shown in bold.

| | SLA | Seed | $A_{area}$ | Height | $N_{mass}$ |
|---|---|---|---|---|---|
| Naïve | 2.618 | 1.140 | 2.276 | 0.173 | 2.910 |
| Bias-corrected | 2.197 | −0.495 | 2.469 | 0.729 | 2.917 |
| Bias (β) | 0.630 | 0.600 | −0.331 | −0.293 | −0.017 |
| Validation Method 1 | | | | | |
| Observed | 2.470 | 0.808 | 2.319 | 0.247 | 2.944 |
| NE | 2.724 | 1.453 | 2.226 | 0.085 | 2.875 |
| $NE_{gf}$ | 2.621 | 1.599 | **2.257** | **0.191** | **2.909** |
| BCE | **2.407** | **1.300** | 2.408 | 0.438 | 2.756 |
| $BCE_{gf}$ | 2.400 | 1.506 | 2.192 | 0.339 | 2.886 |
| Validation Method 2 | | | | | |
| Observed | 2.455 | 0.938 | 2.291 | 0.303 | 2.934 |
| NE | 2.673 | 1.231 | 2.273 | 0.109 | 2.899 |
| $NE_{gf}$ | 2.596 | 1.195 | **2.283** | **0.240** | **2.912** |
| BCE | 2.139 | **1.008** | 2.387 | 0.524 | 2.839 |
| $BCE_{gf}$ | **2.392** | 0.315 | 2.200 | 0.157 | 2.842 |

Species in the Yosemite plots showed a rather moderate bias in SLA measurements, with rarely measured species having lower SLA values (Fig. 5). Validation with Method 1 showed that the BCE provided the closest match to the observed values. When missing species within plots were assigned this BCE estimate, substantial small-scale variation in SLA was lost (Fig. 6). This suggests that some of the small-scale variation in SLA is due not to true differences in environmental conditions or community composition, but to variation in trait coverage among plots. Furthermore, large-scale variation among plots was also due to trait coverage, as coverage was positively related to mean annual temperature and negatively related to precipitation (Appendix S1). As trait coverage was confounded with environmental variation in this case, failing to remove the bias resulting from variation in trait coverage could lead to spurious inference regarding environmental effects. Indeed, using the raw, uncorrected plot CWM values suggested a weak but significant positive relationship between mean annual temperature and log(SLA) ($P = 0.03$), while this relationship disappeared after removing the missing species bias ($P = 0.335$). The multiple $R^2$ from a regression of plot


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science

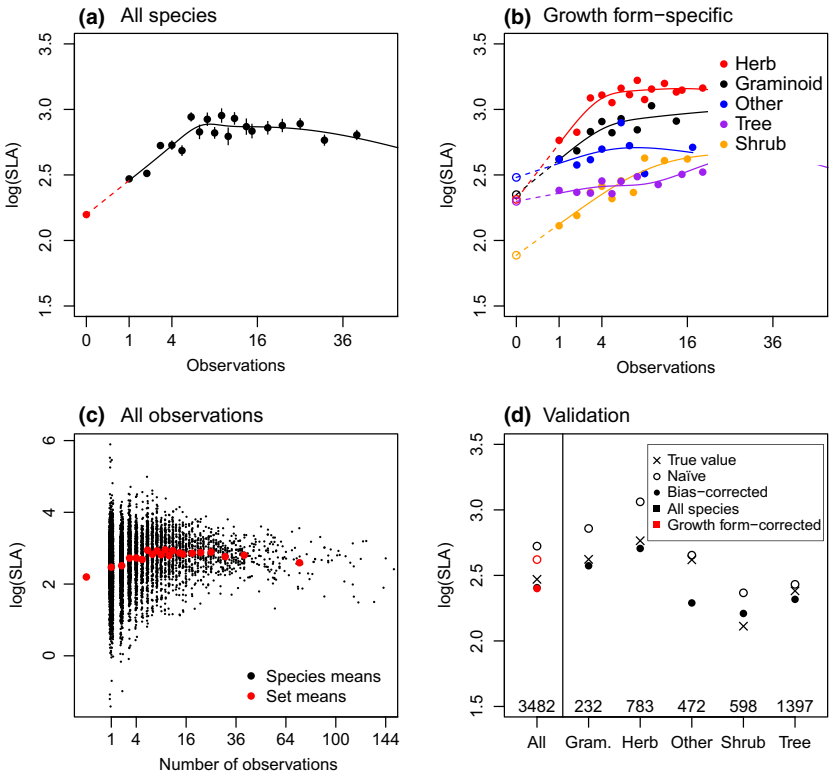


**Fig. 2.** Estimation of log(SLA) for species with no measurements. In (**a**), species are binned (into sets *U*) according to SLA measurement intensity, and a spline was fit between sqrt(*N*(*U*)) and *T*(*U*) (black line). That spline was projected to zero (red) to estimate the mean log(SLA) of all unmeasured plant species. Growth forms differ in their SLA, as well as the strength of the relationship between number of observations and mean log(SLA) (**b**). Herbs and shrubs show particularly strong biases, while trees show very little. All species observations (excluding six species with more than 144 trait measurements) are plotted in (**c**), with the within-group means and projected trait value plotted as red dots. Panel (**d**) shows validation of four missing species estimates using Method 1 (see Methods). Red dots indicate the growth form-corrected versions (the black BCE dot is nearly obscured by the red $BCE_{gf}$ dot). Bias-corrected values provided a closer match to true values for all species, Graminoids, Herbs and Shrubs, but the Naïve estimates were closer for the two groups with weakest bias: Trees and Other. The numbers of species in each group are indicated along the *x* axis.

mean SLA against mean annual temperature and annual precipitation also improved somewhat from 0.058 to 0.096 with bias correction, although most of the variance remained unexplained, due to some combination of variability in plot SLA values and unmeasured environmental variables.

## Discussion

In general, one can distinguish three types of missing data – missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR; Rubin 1976; Nakagawa & Freckleton 2008). MNAR arises when the tendency for 'missingness' of a variable is a function of the variable itself, or of some other unmeasured variable (Nakagawa & Freckleton 2008). In this situation, the data generating mechanism cannot be safely ignored; rather we must explicitly model the process leading to missing data. Here, we propose a 'data augmentation' approach (Nakagawa & Freckleton 2008), in contrast to previously used data imputation approaches.

Our results demonstrate that, at least for some plant traits, the tendency for missingness depends on the trait values themselves. Thus, these data are MNAR, and analyses ignoring the missing data will be biased (Nakagawa & Freckleton 2008). This bias can take various forms, including obscuring geographic trait patterns, creating spurious relationships between environmental variables and CWMs (e.g. Appendix S7), and potentially influencing trait–trait correlation patterns.

To avoid these problems, we must explicitly model the process leading to missing data. Our results indicate that missing species have relatively low SLA, low seed mass, high height, high $N_{mass}$ and high $A_{area}$. In part, this is due to biased sampling of growth forms, combined with variations in trait means between growth forms. For example, herbaceous species are better represented in the plant height data than are trees, while at the same time being


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science 

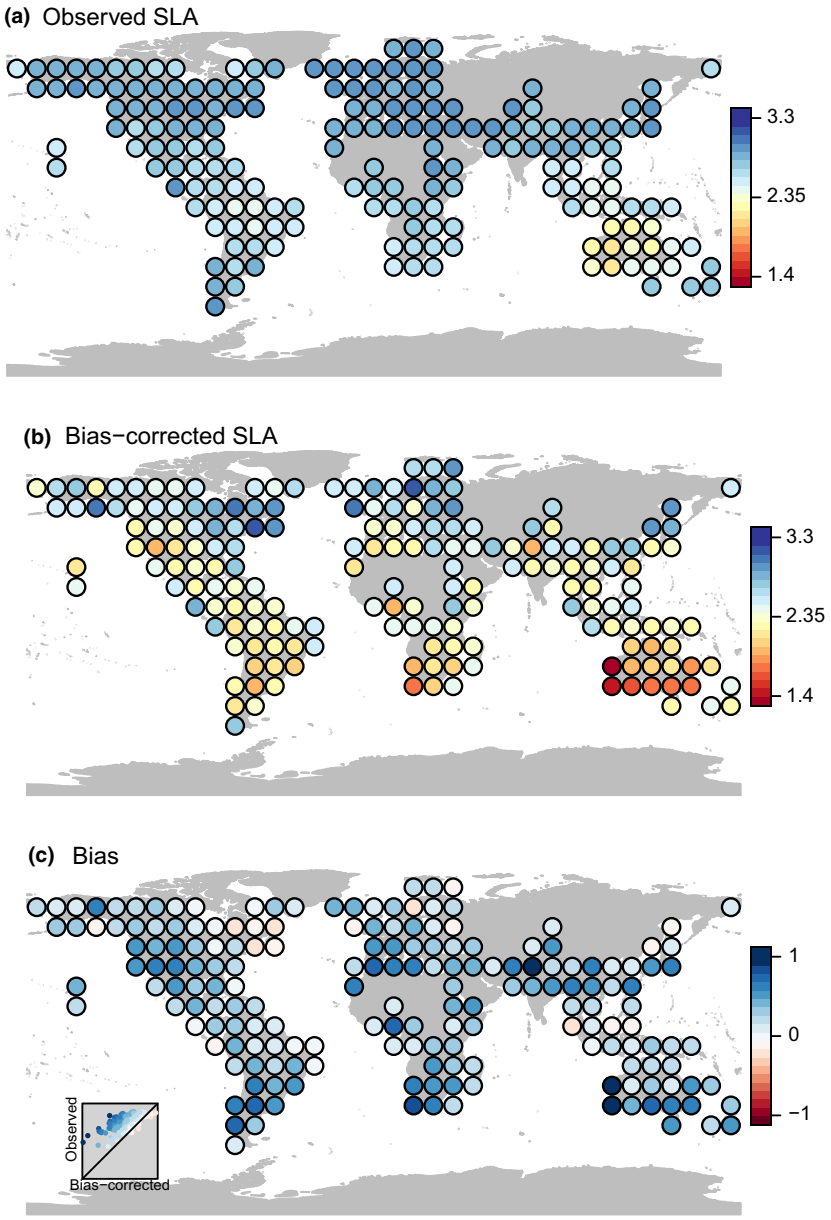


**Fig. 3.** Global patterns in SLA values and bias. Within 10 × 10° windows, we calculated observed mean log(SLA) (**a**), bias-corrected SLA (**b**) and the strength of the bias (**c**). Across nearly all windows, we estimate a positive bias in SLA. There was substantial variation in the strength of the estimated bias. For example, northeastern North America showed very weak biases, while the southwest had relatively large estimated biases.

shorter. Together, this leads to a bias towards more frequent measurement of short species. However, such growth form-dependent patterns were not the entire explanation, as consistent biases were also detected within growth forms. Somewhat surprisingly, all groups except for herbs showed consistent biases towards short species, in contrast with predictions of the EAH. The EAH also predicts that larger, more apparent species should show weaker biases than more cryptic species. Indeed, for SLA, trees showed the weakest bias. On the other hand, for the next most biased trait (seed mass), trees showed relatively strong biases while herbs and shrubs were weakly biased.

The directions of these biases make sense, given some likely biases in data collection. There may be a preference for working on short-lived species, perhaps because they are easier to measure or manipulate, which would lead to better representation of short-statured species. The bias towards measuring high-SLA species may reflect similar biases towards relatively weedy species. It could also result from a collecting bias, where researchers are more likely to


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science

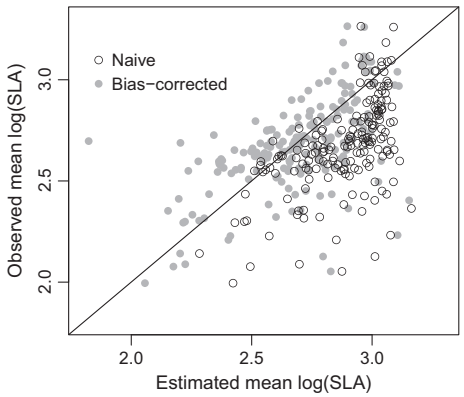


**Fig. 4.** Validation of the global bias-correction estimates, by predicting to species with one trait observation (Method 1, see Methods). Each point represents one 10 × 10° window, with an observed mean SLA value across species with a single SLA measurement. The Naïve estimates for these species means were consistently too high, and the bias-correction method removed this bias.

collect species from somewhat disturbed environments, which in turn have species with relatively high SLA values (Kühner & Kleyer 2008). Finally, it may also reflect a preference for measuring broad-leaved, rather than needle-leaved species. The fact that we detected a bias towards measuring leaves with large SLA, while there was a very weak reverse bias for $N_{mass}$, is somewhat surprising, given that these two traits are typically positively related (Wright et al. 2004). This suggests that a future avenue for research will be a focus on measurement bias in multivariate trait space – testing the idea that certain trait *combinations* are less likely to attract measurement. Such biases, if they exist, would have important consequences for our understanding of trait–trait correlations, with implications for the interpretation of physiological trade-offs and constraints.

While not perfect, our bias correction was successful in greatly reducing bias for the two most-biased traits (seed mass and SLA). Further development of this or similar bias-correction methods seem like a productive avenue for future research. In particular, it will be useful to develop insights into which settings in the bias-correction process lead to the least biased results. Here, we showed that minimum group size and spline flexibility have some influence on the BCE, although not enough to overturn the main result that SLA is biased. More objective approaches to selecting these parameters would be highly desirable.

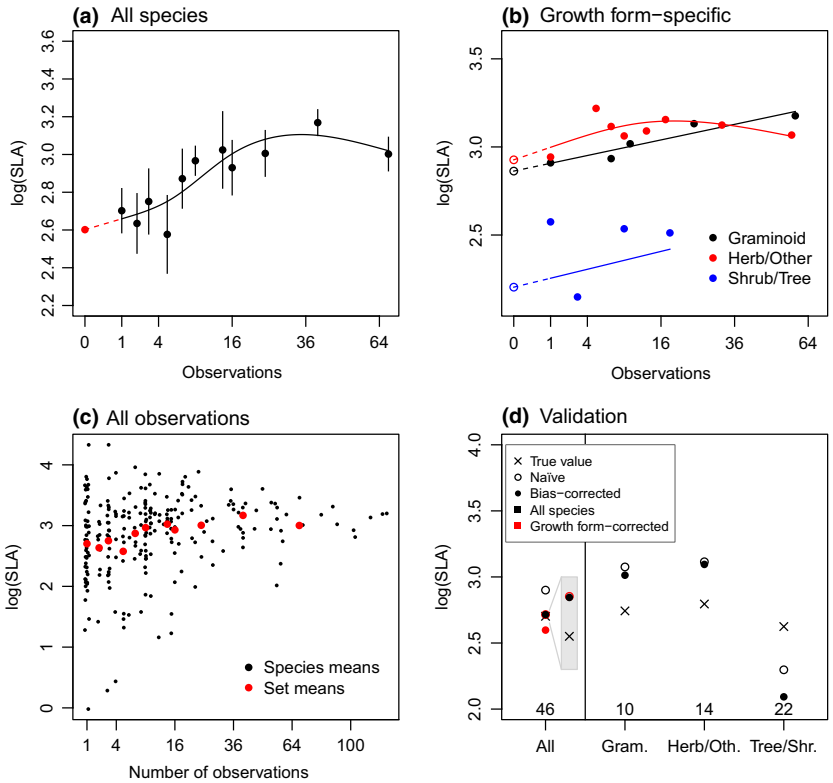


**Fig. 5.** Bias estimation for SLA within Yosemite vegetation plots. Details are as in Fig. 1, except that we used a coarser division of functional types. The overall bias correction (a) and growth-form specific method both indicated moderate positive bias (b) All individual species means are shown in (c) The grey box in (**d**) shows a zoom-in to reveal details among the points. Overall, the missing species bias appears to be fairly moderate in this case, although the bias-correction method still performed best in the validation (**d**).


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science 

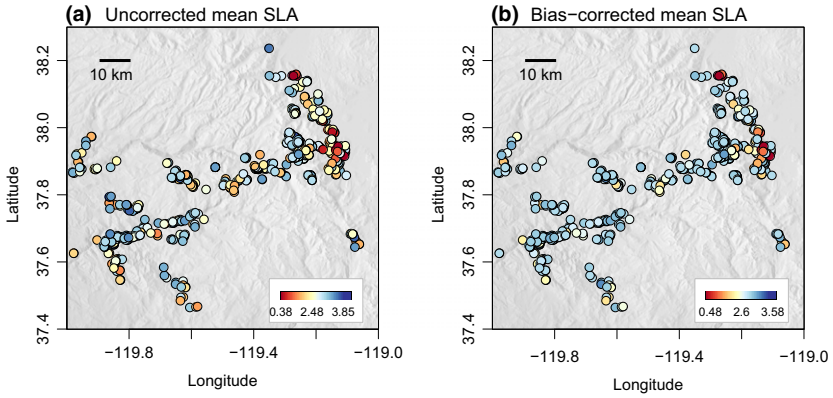


**Fig. 6.** Uncorrected (**a**) and bias-corrected (**b**) plot mean SLA values in and around Yosemite. The bias correction smooths the pattern, reducing fine-scale variation among plots due to variation in trait coverage in these plots.

The two examples demonstrate that the influence of the bias can be large and ecologically relevant. In the regional example, the bias correction revealed a southwest–northeast gradient across North America in SLA that hardly appears in the raw data. In the local Yosemite example, bias correction led to much smoother patterns in SLA, evidently because much of the variation in nearby plots was due to variation in trait coverage. Further, because trait coverage was correlated with environmental variables, removing the bias altered the perceived trait–environment relationships.

Our bias-correction method seeks to solve one rather specific problem. That is, we would like to have an unbiased estimate of the mean trait value of missing species. This problem is of interest in calculating CWMs, for example, where all one needs to know is the mean value of missing species. However, in other applications, the mean alone is insufficient. For example, the calculation of functional trait metrics depends on individual trait values for each species – clearly replacing all missing species with the estimated mean of these species would produce a bias towards low diversity. Our approach here could potentially be extended to address this challenge. We focused on estimating the mean trait value of species within sets defined on measurement intensity. It is equally possible to estimate the variance of trait values of species within each set. One could then assume that missing trait values are log-normally distributed with the bias-corrected mean and variance, and draw missing values from this distribution for each species repeatedly to generate a distribution of functional diversity values. Alternatively, it could be possible to combine gap-filling approaches (e.g. Shan et al. 2012; Swenson 2014; Taugourdeau et al. 2014) with the bias-correction method proposed here, or to use this method to evaluate the degree to which gap-filling methods remove bias (which could occur if they take phylogenetic relationships into account; e.g. Shan et al. 2012; Swenson 2014).

## Recommendations

In general, ecologists should be aware that trait databases likely contain a biased subset of species, and the missing species are likely missing not at random (Rubin 1976). Thus, existing studies with large fractions of missing species should be interpreted with caution. Future trait-based studies should seek to assess the strength of the missing species bias, given the trait data set and species list in question. In particular, we make the following recommendations:

1. When possible, researchers should use the approach demonstrated here to estimate the strength of the missing species bias and validate that estimate. To that end, we are providing R scripts for performing this analysis as an appendix to this paper. Data availability will provide a constraint here – the bias estimation depends on having repeated trait measurements across a large number of species, something that may only be available from large databases.
2. If a bias is detected and the application requires only the mean trait of the missing species, researchers should use a bias-correction method such as that presented here. It is crucial to validate the bias-correction method. Of the five traits considered here, the NE was never least biased, but the $NE_{gf}$ was. Thus, a simple correction based on growth forms (or other groupings such as clades) may be sufficient to remove bias. This will need to be assessed on a case-by-case basis, although the fact that the BCE performed best on the most biased traits may prove to be general.
3. Other approaches for examining the potential importance of a bias should also be pursued. In the Yosemite example, the problem in determining trait–environment relationships occurred because of the combination of a trait bias and variation in trait coverage, which itself was correlated with the environment. Without a relationship between trait coverage and the


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science

environment, the missing species bias would not bias the trait–environment relationship. Thus, testing the relationship between coverage and the environmental variables of interest should become routine.

4. Finally, although biases can be estimated statistically, there is ultimately no better solution than further development of trait databases. Thus, we strongly encourage our colleagues to continue to measure functional traits and to contribute their measurements to shared databases.

## Conclusion

Missing data are and will continue to be a ubiquitous feature of trait-based ecology studies. To date, most efforts were focused on filling gaps in traits, largely ignoring gaps on the species side of the matrix. Gap-filling approaches are one promising way forward, but when data are MNAR, any biases in the representation of species in the original data will likely be propagated to the gap-filled data. We proposed a method for detecting and correcting the missing species bias based on the simple assumption that the relationship between measurement intensity and trait values is informative about the difference between species with no measurements and some measurements. This method and similar approaches should provide a useful tool to make unbiased inferences in the face of MNAR trait and species data.

## Acknowledgements

This study was supported by the TRY initiative on plant traits (http://www.try-db.org). The TRY initiative and database is hosted, developed and maintained by J. Kattge and G. Bönisch (Max Planck Institute for Biogeochemistry, Jena, DE). TRY is currently supported by DIVERSITAS/Future Earth and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. We thank Sebastian Schmidtlein, Anja Linstädter and an anonymous reviewer for their comments and suggestions on the manuscript.

## References

Albuquerque, U.P. & Lucena, R.F.P. 2005. Can apparency affect the use of plants by local people in tropical forests? *Interciencia* 30: 506–510.

Bonan, G.B., Oleson, K.W., Fisher, R.A., Lasslop, G. & Reichstein, M. 2012. Reconciling leaf physiological traits and canopy flux data: use of the TRY and FLUXNET databases in the Community Land Model version 4. *Journal of Geophysical Research* 117: G02026.

Cornwell, W.K. & Ackerly, D.D. 2009. Community assembly and shifts in the distribution of trait values across an environmental gradient in coastal California. *Ecological Monographs* 79: 109–126.

Diaz, S., Noy-Meir, I. & Cabido, M. 2001. Can grazing response of herbaceous plants be predicted from simple vegetative traits? *Journal of Applied Ecology* 38: 497–508.

Frenette-Dussault, C., Shipley, B., Meziane, D. & Hingrat, Y. 2013. Trait-based climate change predictions of plant community structure in arid steppes. *Journal of Ecology* 101: 484–492.

Garnier, E., Cortez, J., Billes, G., Navas, M.L., Roumet, C., Debussche, M., Laurent, G., Blanchard, A., Aubry, D., (...) & Toussaint, J.-P. 2004. Plant functional markers capture ecosystem properties during secondary succession. *Ecology* 85: 2630–2637.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.

Kattge, J., Knorr, W., Raddatz, T. & Wirth, C. 2009. Quantifying photosynthetic capacity and its relationship to leaf nitrogen content for global-scale terrestrial biosphere models. *Global Change Biology* 15: 976–991.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., (...) & Wirth, C. 2011. TRY – a global database of plant traits. *Global Change Biology* 17: 2905–2935.

Kraft, N.J.B., Valencia, R. & Ackerly, D.D. 2008. Functional traits and niche-based tree community assembly in an Amazonian forest. *Science* 322: 580–582.

Kühner, A. & Kleyer, M. 2008. A parsimonious combination of functional traits predicting plant response to disturbance and soil fertility. *Journal of Vegetation Science* 19: 681–692.

McGill, B.J., Enquist, B.E., Weiher, E. & Westoby, M. 2006. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution* 21: 178–185.

Moerman, D.E. & Estabrook, G.F. 2006. The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography* 33: 1969–1974.

Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution* 23: 592–596.

Pakeman, R.J. 2014. Functional trait metrics are sensitive to the completeness of the species' trait data? *Methods in Ecology and Evolution* 5: 9–15.

Pakeman, R.J. & Quested, H.M. 2007. Sampling plant functional traits: what proportion of the species need to be measured? *Applied Vegetation Science* 10: 91–96.

Pierce, S., Brusa, G., Vagge, I. & Cerabolini, B.E.L. 2013. Allocating CSR plant functional types: the use of leaf economics and size traits to classify woody and herbaceous vascular plants. *Functional Ecology* 27: 1002–1010.

Reich, P.B., Walters, M.B. & Ellsworth, D.S. 1997. From tropics to tundra: global convergence in plant functioning. *Proceedings of the National Academy of Sciences of the United States of America* 94: 13730–13734.

Rubin, D.B. 1976. Inference and missing data. *Biometrika* 63: 581–590.

Sandel, B. & Dangremond, E. 2012. Climate change and the invasion of California by grasses. *Global Change Biology* 18: 277–289.

Shan, H., Kattge, J., Reich, P., Banerjee, A., Schrodt, F. & Reichstein, M. 2012. Gap filling in the plant kingdom – trait prediction using hierarchical probabilistic matrix factorization. In: Langford, J. & Pineau, J. (eds.) *Proceedings of the 29th International Conference on Machine Learning*, pp. 1303–1310. Omnipress, Madison, WI, US.

Swenson, N.G. 2014. Phylogenetic imputation of plant functional trait databases. *Ecography* 37: 105–110.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. & Amiaud, B. 2014. Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. *Ecology and Evolution* 4: 944–958.

Weiher, E. & Keddy, P.A. 1995. Assembly rules, null models, and trait dispersion: new questions from old patterns. *Oikos* 74: 159–164.

Westoby, M. 1998. A leaf–height–seed (LHS) plant ecology strategy scheme. *Plant and Soil* 199: 213–227.

Westoby, M. & Wright, I.J. 2006. Land-plant ecology on the basis of functional traits. *Trends in Ecology & Evolution* 21: 261–268.

Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J.H.C., (...) & Villar, R. 2004. The worldwide leaf economics spectrum. *Nature* 428: 821–827.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Supporting information including illustrations of bias-correction estimates for additional traits, an assessment of sensitivity to decisions made in the bias-correction process, and relationships of Yosemite vegetation plot SLA with climate.

**Appendix S2.** Full references for all trait data used.

**Appendix S3.** R functions and example script for implementing bias correction.


Doi: 10.1111/jvs.12292 © 2015 International Association for Vegetation Science