Handling missing values in trait data

Article

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

www.reading.ac.uk/centaur

Global Ecology and Biogeography   A Journal of Macroecology   WILEY

Handling missing values in trait data

Thomas F. Johnson[1] [ID] | Nick J. B. Isaac[2] | Agustin Paviolo[3,4] |
Manuela González-Suárez[1] [ID]

[1]Ecology and Evolutionary Biology, School of Biological Sciences, University of Reading, Reading, UK

[2]Biodiversity Science Area, Centre for Ecology and Hydrology, Wallingford, UK

[3]Instituto de Biología Subtropical, CONICET-Universidad Nacional de Misiones, Misiones, Argentina

[4]Asociación Civil Centro de Investigaciones del Bosque Atlántico, Misiones, Argentina

Correspondence
Thomas Frederick Johnson, Ecology and Evolutionary Biology, School of Biological Sciences, University of Reading, Reading, RG6 6UR, UK.
Email: thomas.frederick.johnson@outlook.com

Funding information
Natural Environment Research Council, Grant/Award Number: J71566E

Editor: Franziska Schrodt

Abstract

Aim: Trait data are widely used in ecological and evolutionary phylogenetic comparative studies, but often values are not available for all species of interest. Traditionally, researchers have excluded species without data from analyses, but estimation of missing values using imputation has been proposed as a better approach. However, imputation methods have largely been designed for randomly missing data, whereas trait data are often not missing at random (e.g., more data for bigger species). Here, we evaluate the performance of approaches for handling missing values when considering biased datasets.

Location: Any.

Time period: Any.

Major taxa studied: Any.

Methods: We simulated continuous traits and separate response variables to test the performance of nine imputation methods and complete-case analysis (excluding missing values from the dataset) under biased missing data scenarios. We characterized performance by estimating the error in imputed trait values (deviation from the true value) and inferred trait–response relationships (deviation from the true relationship between a trait and response).

Results: Generally, *Rphylopars* imputation produced the most accurate estimate of missing values and best preserved the response–trait slope. However, estimates of missing data were still inaccurate, even with only 5% of values missing. Under severe biases, errors were high with every approach. Imputation was not always the best option, with complete-case analysis frequently outperforming *Mice* imputation and, to a lesser degree, *BHPMF* imputation. *Mice*, a popular approach, performed poorly when the response variable was excluded from the imputation model.

Main conclusions: Imputation can handle missing data effectively in some conditions but is not always the best solution. None of the methods we tested could deal effectively with severe biases, which can be common in trait datasets. We recommend rigorous data checking for biases before and after imputation and propose variables that can assist researchers working with incomplete datasets to detect data biases and minimize errors.

1 | INTRODUCTION

Trait data describe the characteristics of individuals of a population or species (Webb et al., 2010). Trait-based analyses have been essential for improving our understanding of ecological and evolutionary processes, for example, identifying negative impacts of climate change on biodiversity (Lancaster et al., 2017; Pacifici et al., 2017), common life-history strategies among invasive species (Allen et al., 2017; González-Suárez et al., 2015), and evolutionary changes in reproductive traits (Baker et al., 2020). Large-scale modelling studies like these are increasing in popularity and often require trait data for numerous species and across taxonomic groups (Ríos-Saldaña et al., 2018). However, trait datasets can contain many missing values, and these values can be missing with a bias (Roth et al., 2018; Sandel et al., 2015). For example, in a widely used mammalian trait dataset (Jones et al., 2009), species with smaller body mass values are more likely to have missing data for other traits, and this bias in missing data can impact inferences in comparative analyses (González-Suárez et al., 2012).

The literature recognizes three broad types of missing data mechanisms: (a) missing completely at random (MCAR), where there is no bias, and records represent a random sample; (b) missing at random (MAR), where missing data can be explained by available variables (for example, we know about the bias and can account for it statistically); and (c) missing not at random (MNAR), where missing data cannot be explained by available information (for example, we either do not know about the bias or we lack associated information that could account for it statistically) (Little & Rubin, 2002).

Currently, there are at least 160 packages for handling missing data available on the R-CRAN repository (Josse et al., 2020). A simple, common approach is "complete-case analysis", that is, to exclude all observations with any missing values. This approach is robust when there is no bias (MCAR missing data); bias in the missing values can lead to erroneous inferences. Imputation, estimating missing values, is an alternative approach to handle missing data that can bypass this disadvantage (Little & Rubin, 2002). Imputation methods range from simple approaches, such as filling missing values with an average, to more complex approaches, such as estimating missing values using statistical models (e.g., regression and random forest). Models can also be made more complex e.g. adding hierarchical information, allowing censored observations and weighting observations. There are also approaches designed specifically for handling values with extreme bias (MNAR), in addition to methods for imputing missing response (sometimes called outcome or dependent variable) values (for a more comprehensive description of methods, see Molenberghs et al., 2015).

Imputation can be applied to any dataset but is particularly useful for trait data because traits are often correlated (e.g., body mass is correlated with body length) and shaped by evolutionary history. Therefore, correlations and phylogenetic information can be used to predict missing trait values more accurately (Penone et al., 2014; Swenson, 2014). Previous studies have suggested that imputation in ecological and evolutionary studies generally outperforms complete-case analysis (Kim et al., 2018; Little & Rubin, 2002; Penone et al., 2014). However, imputation can only be successful if it accounts for the mechanism by which data are missing. If the imputation model cannot account for this mechanism (e.g., under extreme biases like MNAR), it is plausible that imputation might even amplify error in inference.

In this manuscript, we evaluate the performance of different approaches for handling missing trait data, considering the following questions. How effective is imputation at estimating missing values and making inference? Which imputation method is best? Is imputation better than complete-case analysis? How does the amount of missing data and presence of bias affect results? Expanding on previous comparisons of imputation methods, we introduce two new bias types, compare six additional imputation methods, evaluate the implications of including the response variable within the imputation and propose steps for detecting erroneous imputation. Our study is most relevant for phylogenetic comparative studies but still applies to wider missing data scenarios.

2 | MATERIALS AND METHODS

2.1 | Data simulation

We simulated 40 datasets, each with 500 species, using the *simtraits* function (Goolsby et al., 2017). Each dataset included four trait-predictor variables ("traits" hereafter) and one response variable. The 40 datasets represent 10 replicates (seeds 1–10) of four dataset types reflecting the combination of two correlation levels among traits (weak Pearson, $R^2 = .2$, or strong, $R^2 = .6$) with two response–trait slopes (no relationship, *c*. 0, or positive, *c*. .7). Traits were simulated under a Brownian model of evolution, with a Gaussian distribution of values ranging from zero to 10 to mimic the distribution of real trait data on a logarithmic scale (a transformation often used in comparative studies). The impact of phylogenetic signal strength on imputation performance was already tested by Kim et al. (2018) and Molina-Venegas et al. (2018); therefore, we standardized Pagel's λ between the phylogeny and traits at approximately one. The response was simulated as a product of a trait, rather than through the phylogeny, and has a Gaussian distribution ranging from zero to 10. We aimed to represent response variables used in comparative analyses such as extinction risk or population trend, rather than allometric relationships.

From each of the 40 original datasets, we removed trait values to create scenarios reminiscent of real trait datasets. Values were removed from between 5 and 80% of the species (in 5% intervals), across 11 distinct bias types (or missing data mechanisms); see the Supporting Information (Appendix A, Table A1.1). As a control, one mechanism was to remove trait values completely at random, simulating the MCAR category. Two mechanisms stratified deletion with trait values removed evenly over the phylogeny and trait range. The remaining mechanisms explored four bias types likely to occur in trait datasets: (a) Trait (large trait values more likely to be missing); (b) Response (trait values more likely to be missing in species with larger responses); (c) Trait & response (trait values more likely to be missing in species with large trait and large response values); and (d) Phylogeny (trait values more likely to be missing in certain clades) (Figure 1; Supporting Information, Appendix A, Table A1.1).

Within each of these four bias types, we tested two bias severity levels: weak (a conservative lower-end estimate for how much bias exists in trait data) and severe (an upper-end estimate aimed at testing how methods perform under the most extreme biases). Under a weak trait bias, the distribution of trait values becomes marginally skewed and the central point is shifted, but the range of values is largely preserved. Under a severe trait bias, the distribution is truncated and the range reduced with extreme skew and shift in the central point. The weak and severe biases replicate the MAR and MNAR categories, respectively. The Supporting Information (Appendix A1) provides a comprehensive description and justification of the bias severities. In total, across all dataset types, levels of missing data (missingness) and bias types, we generated 7,040 datasets.

2.2 | Data imputation

Testing all available imputation methods was not feasible; instead, we expanded upon previous ecological and evolutionary imputation
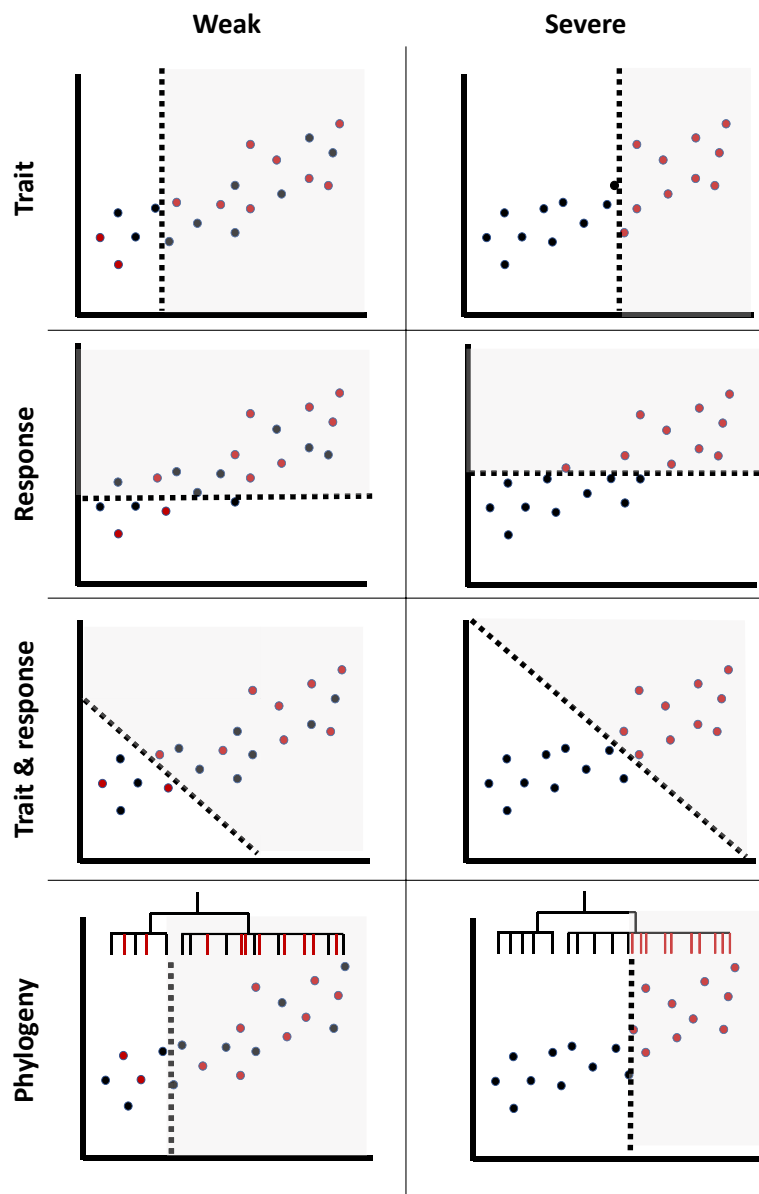

FIGURE 1 Schematic illustration of the effects of different biases. Panels contain an example scatterplot (x axis = trait; y axis = response) depicting a positive trend. In each panel there are 20 points, each representing a species, of which 50% are missing trait values (shown in red). Dotted lines illustrate a removal threshold based on the percentage of missing data (missingness) and bias type, which shows a different mechanism by which data are missing: Trait = large trait values more likely to be missing; Response = trait values more likely to be missing in species with larger responses; Trait & response = trait values more likely to be missing in species with large trait and large response values; Phylogeny = trait values more likely to be missing in certain clades. For each bias type, we illustrate two severities: left panels show weak severity, in which species are split into two groups and species in the shaded area are 1.33 times more likely to be removed; and right panels show severe severity, in which values are removed systematically from large to small (all values removed from the shaded area). For further descriptions of these biases, see the Supporting Information (Appendix A1) [Colour figure can be viewed at wileyonlinelibrary.com]

studies (Penone et al., 2014; Poyatos et al., 2018) to compare methods already identified as effective with new, promising methods. In total, we evaluated the performance of nine imputation methods available from three R packages (R v.3.5.0, R Core Team, 2018): (a) *BHPMF*, Bayesian hierarchal probabilistic matrix factorization (Schrodt et al., 2015); (b) *Rphylopars* (Goolsby et al., 2017); and (c) *Mice*, multiple imputation chained equations (Van Buuren & Groothuis-Oudshoorn, 2011). We summarize these approaches below, providing a more detailed description in the Supporting Information (Appendix A2).

BHPMF is a machine learning technique that takes a sparse trait matrix and uses Bayesian probabilistic matrix factorization to impute and estimate uncertainty in values, but it is not fully Bayesian in that imputation and analysis are not conducted simultaneously (Molenberghs et al., 2015). *BHPMF* provides a multilevel hierarchal framework, which can control for spatial and taxonomic structures (for a comprehensive description of *BHPMF*, see Schrodt et al., 2015). This hierarchical framework, coupled with the overall flexibility of probabilistic matrix factorization (e.g., it can handle nonlinear relationships and binary categories), makes *BHPMF* versatile and potentially robust. However, unlike other approaches it is unable to make estimates for species for which all trait values are missing.

Rphylopars is a maximum likelihood frequentist method that uses a phylogeny and a sparse trait matrix to estimate simultaneously the across-species (phylogenetic) and within-species (phenotypic) trait covariance (similar to a phylogenetic mixed model) to reconstruct the ancestral state and impute missing values (Goolsby et al., 2017). This method is designed explicitly for phylogenetic imputation and requires a phylogenetic tree, which means that the success of *Rphylopars* imputation depends on the phylogenetic signal in a trait; with low signal, the phylogeny may simply add noise. An earlier version of *Rphylopars* was amongst the most accurate methods examined by Penone et al. (2014).

Mice is the most general and flexible of the imputation packages used in this study, offering 24 different methods of imputation, from which we explored three:

1. Predictive mean matching imputes data by matching observed values between traits, then populates missing values in incomplete traits by adopting information from the matched species. This is the default *Mice* approach for continuous data and was considered the best overall method by Poyatos et al. (2018).
2. Bayesian linear regression uses a linear model between traits to estimate missing values. This method is rarely tested and struggles with nonlinear relationships but is appealing to researchers familiar with linear regression.
3. Random forest uses machine learning to produce and aggregate regression trees of the observed data and impute missing values. A similar imputation method, "missForest", was found to be effective by Penone et al. (2014), with results comparable to *Rphylopars* and *Mice* predictive mean matching.

The three imputation approaches we explore fall into two groups: (a) single imputation (*BHPMF* and *Rphylopars*), where each missing value is populated by one estimate (but can have an associated variance); and (b) multiple imputation (*Mice*), where each missing value is assigned multiple estimates from a stochastic draw of the distribution (Little & Rubin, 2002). If the objective of the imputation is to produce estimates of missing values (for example, to fill gaps in a dataset), single imputation is considered most effective, because the stochastic draws in multiple imputation add error (Van Buuren, 2012). However, if the objective is to model imputed values against another variable, the added error in the multiple imputation is advantageous, because when the trait data are modelled the within- and among-dataset errors are pooled, inflating the standard error and reducing the type 1 error rate (Van Buuren, 2012). Although this makes multiple imputation more robust to type 1 errors, it does not necessarily mean that multiple imputation can predict the slope more accurately within a model, because although this slope will have a greater standard error, it might still have the wrong direction.

2.2.1 | Phylogenetic imputation

Imputation has been suggested to improve when phylogenetic information is incorporated (Kim et al., 2018; Penone et al., 2014). To test this, we imputed missing data with *BHPMF* and *Mice*, incorporating and ignoring phylogenetic information (for *Rphylopars*, a phylogeny is required). *BHPMF* is unable to process phylogenies automatically, but its hierarchal nature can support taxonomies. We created a hierarchical node structure reminiscent of a taxonomy by splitting the phylogeny. For *Mice*, we used phylogenetic eigenvectors that described the relationship between the phylogeny and traits (Diniz et al., 2015). Eigenvectors that were effective predictors of a trait were included as predictors within the imputation. We provide a comprehensive description of these approaches in the Supporting Information (Appendix A3).

2.2.2 | Including a response variable in the imputation

The standard practice in comparative studies that use imputation is to impute values using only the traits and, where relevant, the phylogeny. However, the medical statistics literature recommends the inclusion of every variable you plan to analyse, including the response, within the imputation model (Moons et al., 2006; Sterne et al., 2009). Including a response within the imputation of traits, which will then be modelled against the response in later analyses, appears circular and poor practice. However, in the event that the trait has a response bias, including it within the imputation could control for this bias and shift data from the MNAR to the MAR category, where imputation is more robust. We test this by performing each imputation with the response present and absent.

2.3 | Error calculation

2.3.1 | Imputation error: Is there a difference between the true and imputed values?

We compared true and imputed trait values under each of the nine imputation approaches (using the mean value across the repeated imputations for *Mice*), estimating the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{Im} - y_{Tr})^2}$$

where N is the number of imputed values, ranging from 25 (5% of 500) to 400 (80% of 500), y_{Im} is the imputed value for a given observation, and y_{Tr} is its true value. The units for the RMSE are the same as those of the trait (range 0–10). We show alternative error metrics (mean absolute error, median absolute error and R^2 between true and imputed values) in the Supporting Information (Appendices A5 and B4).

Mice guidelines stress that multiple imputation is not effective at predicting missing values and should instead be used for inference after model averaging. However, in the event that inference is prone to error (where the imputed response–trait relationship deviates from the true relationship), it is important to consider how the imputation of missing values influenced this error. Conversely, it is also plausible that a method could produce inaccurate estimates of missing values but still produce valid inference. Thus, assessing error in both the imputations and the inference (see headings 2.3.2 & 2.3.3 below) provides a more holistic view of the imputation approach, which can help to determine the point at which imputation becomes unreliable.

2.3.2 | Slope error: Is there a difference between the true and imputed response–trait slope?

We fitted linear regressions with the imputed datasets to describe the response–trait relationship, recording the slope and associated standard error. We checked assumptions (e.g., normality) in a subset of these models, which were acceptable regardless of the bias or amount of missing data. Given that *Mice* repeats the imputation process numerous times, we fitted multiple regressions using each of the imputed sets and then averaged the slope coefficients. To estimate "slope error", we calculated the absolute difference between the imputed slope (or the complete-case slope) and the true slope. This "slope error" metric illustrates how wrong the imputed slope could be, with the next step showing whether the estimated slope coefficient differed significantly from the true slope. Previous studies have considered how imputation can alter inference, focusing on allometric relationships between traits (Penone et al., 2014) and the impact on functional diversity metrics (Kim et al., 2018). Here, we explored how imputation affects the relationship between traits and a separate response variable.

2.3.3 | Slope significance: Is the difference between the true and imputed slope significant?

We tested whether imputed slopes differed significantly from the true slope and the complete-case slope using the t statistic (Cohen et al., 2003):

$$t = \frac{Slope1 - Slope2}{\sqrt{SE_{Slope1}^2 + SE_{Slope2}^2}}$$

where $Slope1$ is the true slope coefficient and $Slope2$ is the imputed or complete-case slope coefficient; SE_{Slope1} is the standard error of the true slope and SE_{Slope2} is the standard error of the imputed or complete-case slope. We calculated degrees of freedom as the total sample size from the true relationship dataset plus the imputed or complete-case dataset, minus four. We estimated significance at the 95% level. The *Mice* model slopes were averaged across each of the repeats, and the standard errors were pooled by calculating the within- and among-dataset variation, following Little and Rubin (2002). Incorporating the within- and among-dataset variation inflates the standard errors around the slope. This is a key advantage to the *Mice* approaches, because although slope error could be high, the inflated standard errors around the slope reduce the probability of the imputed slope differing significantly from the true slope, and the likelihood of obtaining type 1 errors.

2.4 | Data analysis

To understand the factors influencing the different error estimates, we fitted regression models with various predictors (details below and in Supporting Information, Appendix A6) and dataset seed as a random intercept effect. We used linear mixed models for numerically continuous responses, with a log_{10}-transformation on imputation error and a square-root transformation on slope error, and logistic mixed models for binary responses (e.g., significant or non-significant difference between the imputed relationship and the true relationship). In all cases, we ensured that model assumptions were met. Summary statistics display the mean $\pm$ *SD*.

2.4.1 | Modelling imputation error

We modelled imputation error as a function of six predictors: imputation approach, bias type, missingness (percentage of missing values in a dataset), response in imputation, initial slope direction (positive or none) and between-trait correlation (Supporting Information, Appendix A, Table A6.1). We included interaction terms between imputation approach and bias type and between imputation approach and missingness. We also tested whether including the response in the imputation improved accuracy by testing an interaction between response in imputation, imputation approach and initial slope direction. In some cases, the imputation resulted in implausible values;

we removed any records with an RMSE > 10 to reduce the effect of these outliers.

2.4.2 | Modelling slope error

We modelled slope errors separately for dataset types with initial positive relationships (response–trait slope c. .7) and with no initial relationship (response–trait slope c. 0). We tested as predictors the imputation approach, bias type, missingness and between-trait correlation, in addition to interactions of imputation approach with bias type and missingness. We ran this model first including complete case as a category within the imputation approach factor to identify scenarios where imputation is worse than complete-case analysis. This required the exclusion of response in imputation as a predictor, because this variable was not applicable for complete-case records. Second, we excluded the complete-case records and tested response in imputation as a factor, including an interaction with imputation approach.

2.4.3 | Predicting imputation and slope error

We predicted imputation error using the variables missingness, phylogenetic clustering and change in mean (difference in mean before and after imputation). To predict slope error and significance, we used the variables missingness, phylogenetic clustering, change in mean and change in slope (between imputation and complete case). For all models, we grouped the datasets with positive and no relationship slopes, because in a real scenario a user would not know the true relationship.

3 | RESULTS

Including phylogenetic information generally improved imputation performance in every method (Supporting Information, Appendix B1); thus, we focused on phylogenetic imputation methods, showing results for non-phylogenetic approaches in the Supporting Information (Appendix B3).

3.1 | Which method performs best?

Imputed values were most accurate with *Rphylopars* (Supporting Information, Appendix B, Table B2.1), which had consistently lower imputation errors in every bias type. However, *BHPMF* was the best approach when missing data exceeded 60% with a severe bias on the trait, and *Mice* random forest, *BHPMF* and *Rphylopars* were comparable when missing data exceeded 40% with a severe bias on the phylogeny (Figure 2; Supporting Information, Appendix B, Figure B4.2). Imputation error results were similar regardless of whether the true response–trait slope was positive or had no relationship

(Supporting Information, Appendix B, Figure B4.1), and results were similar across all imputation error metrics (Supporting Information, Appendix B, Figures B4.2–B4.5).

Rphylopars was also generally the best approach for preserving a response–trait relationship, with a significantly lower slope error than all other imputation approaches and complete-case analysis, regardless of whether the true response–trait slope was positive or there was no relationship (Supporting Information, Appendix B, Tables B2.2–B2.3). However, for a severe bias on the trait or phylogeny, the best method was dependent on the true response–trait relationship. With no relationship, the *Mice* approaches performed best (Figure 3), whereas when the true slope was positive, complete case was the best approach. *Rphylopars* was the fastest imputation approach (Supporting Information, Appendix B, Table B8.1).

3.2 | Are imputed values accurate?

Imputation errors increased with the percentage of missing data, missingness (Est = 0.33, SE = 0.003, t = 103, p < .001), and were affected by bias type (Figure 2). Weak and stratified biases were comparable to no bias datasets, but errors were much greater when data were missing with severe bias (Supporting Information, Appendix B, Appendix B4).

Imputed values were as likely to be overestimated as underestimated, except when there was a severe bias on the trait (largest trait values removed), where, as expected, imputed values were primarily underestimated (Supporting Information, Appendix B, Figures B5.1–B5.5). Although *Rphylopars* had the smallest imputation errors, imputed values were still inaccurate. At 5% missing data, the mean difference between imputed and true values for *Rphylopars* was 0.56 $\pm$ 0.15 with no bias, 0.56 $\pm$ 0.15 in the stratified biases, 0.57 $\pm$ 0.16 in the weak biases, and 1.39 $\pm$ 0.57 in the severe biases, all increasing with missingness (Figure 2).

3.3 | Can imputed data preserve response–trait relationships?

Slope errors increased with missingness (true positive relationship: Est = 0.27, SE = 0.008, t = 33, p < .001; true no relationship: Est = 0.23, SE = 0.004, t = 54, p < .001) and were affected by bias type, with large errors detected when data were missing with a severe bias (Figure 3). Imputed slopes were both over- and underestimates of the true slope when there was no true relationship (Supporting Information, Appendix B, Figures B5.6–B5.10). When the true relationship was positive, *Rphylopars* and *BHPMF* again resulted in both over- and underestimated slopes, but *Mice* approaches consistently underestimated the true relationship, with slopes from the imputed datasets tending towards zero (Supporting Information, Appendix B, Figures B5.11–B5.15).

Although imputation errors were often large, imputing missing values did not always introduce errors in the response–trait
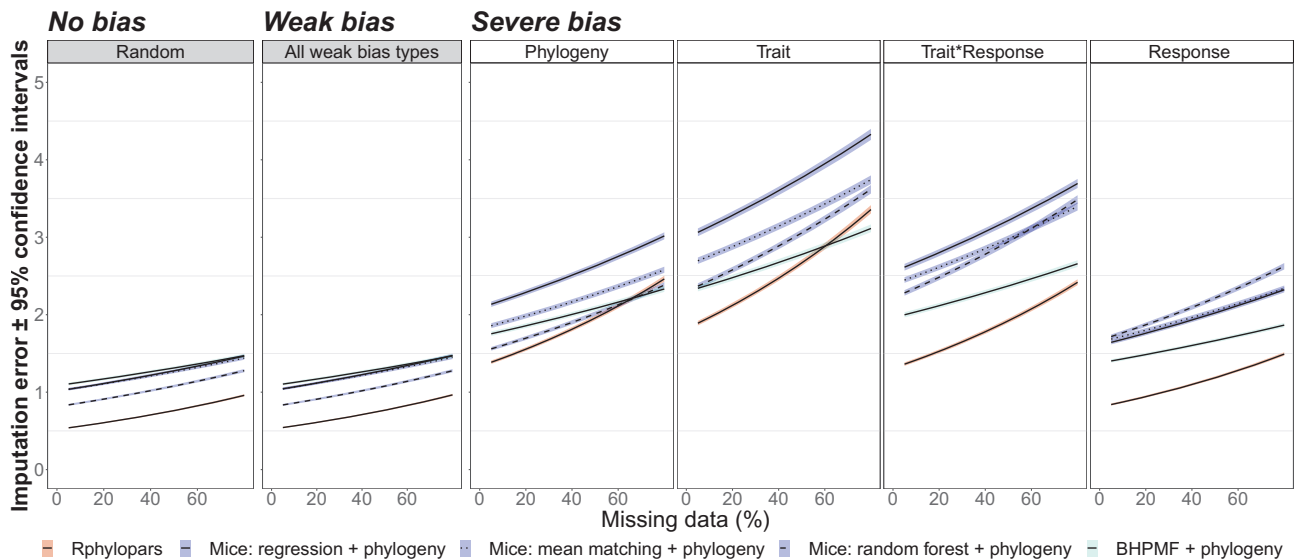
FIGURE 2 Difference between imputed and true trait values (RMSE = root mean square error) for five phylogenetic imputation approaches under varying percentages of missing data (missingness) and bias types. Lines depict the marginal effect of missingness and bias type from a regression model and were averaged across other predictors: seed, response in imputation, between-trait correlation and initial slope direction. For the equivalent plot split based on initial slope direction, see the Supporting Information (Appendix B, Figure B4.1). Confidence intervals were derived from 500 bootstrap simulations and depict the upper and lower bounds (95%) [Colour figure can be viewed at wileyonlinelibrary.com]

relationship. We observed low slope errors in all imputation approaches and all non-severe biases when few data were missing, but as missingness increased the slope error increased exponentially (Figure 3). *Rphylopars* was most robust, with slope errors < 0.05 for all levels of missingness in the no bias, stratified bias and weak bias datasets, regardless of the true response–trait relationship (Supporting Information, Appendix B, Figures B4.6–B4.7). However, *Rphylopars*, alongside all other approaches, had high errors under the severe biases, particularly when the bias acted on both the trait and the response.

Missingness and bias type also influenced whether slopes were significantly different from the true slope in a comparable way to slope error (Supporting Information, Appendix B, Figures B4.8–B4.11).

3.4 | Should the response be included in the imputation?

When the response–trait slope was positive, including the response within the imputation decreased imputation error in all approaches and also decreased slope error substantially in *Mice*, to the extent that it was almost comparable with *Rphylopars* (Supporting Information, Appendix B, Figures B6.1–B6.4). Including the response in the imputation increased slope error in *Rphylopars* and *BHPMF*. When there was no relationship between the trait and response, including the response in the imputation increased imputation and slope errors in every approach, but with a small effect (Supporting Information, Appendix B, Figures B6.2–B6.6).

3.5 | Can we predict when the imputed values and response–trait relationships become inaccurate?

Given that *Rphylopars* was found generally to be the best method, we focused on predicting errors under this method. Missingness, phylogenetic clustering and change in mean were important predictors of slope error, significant differences in slope error and imputation error. Change in slope was also a relevant predictor for slope error and significant differences in slope (Supporting Information, Appendix B, Figure B7.1).

4 | DISCUSSION

Overall, our results show that there is no single best solution to deal with missing data. *Rphylopars* was generally the best approach for predicting missing values and was consistently more accurate than *BHPMF* and *Mice* at maintaining the true response–trait relationship. However, in some scenarios all imputation approaches were outperformed by complete-case analysis, showing that imputation is not always the best option. When using imputation, including phylogenetic information widely reduced errors in our phylogenetically derived trait datasets, but including the response during imputation had mixed effects: increased accuracy for *Mice* approaches, but decreased accuracy for *Rphylopars* and *BHPMF*. Our results suggest that researchers need to assess the available data and consider the need for imputation versus limiting the scope of the study or completing analyses for separate groups. Use of data imputation should be scrutinized, checking for changes in the data before and after imputation (which might indicate biases and assist in detection
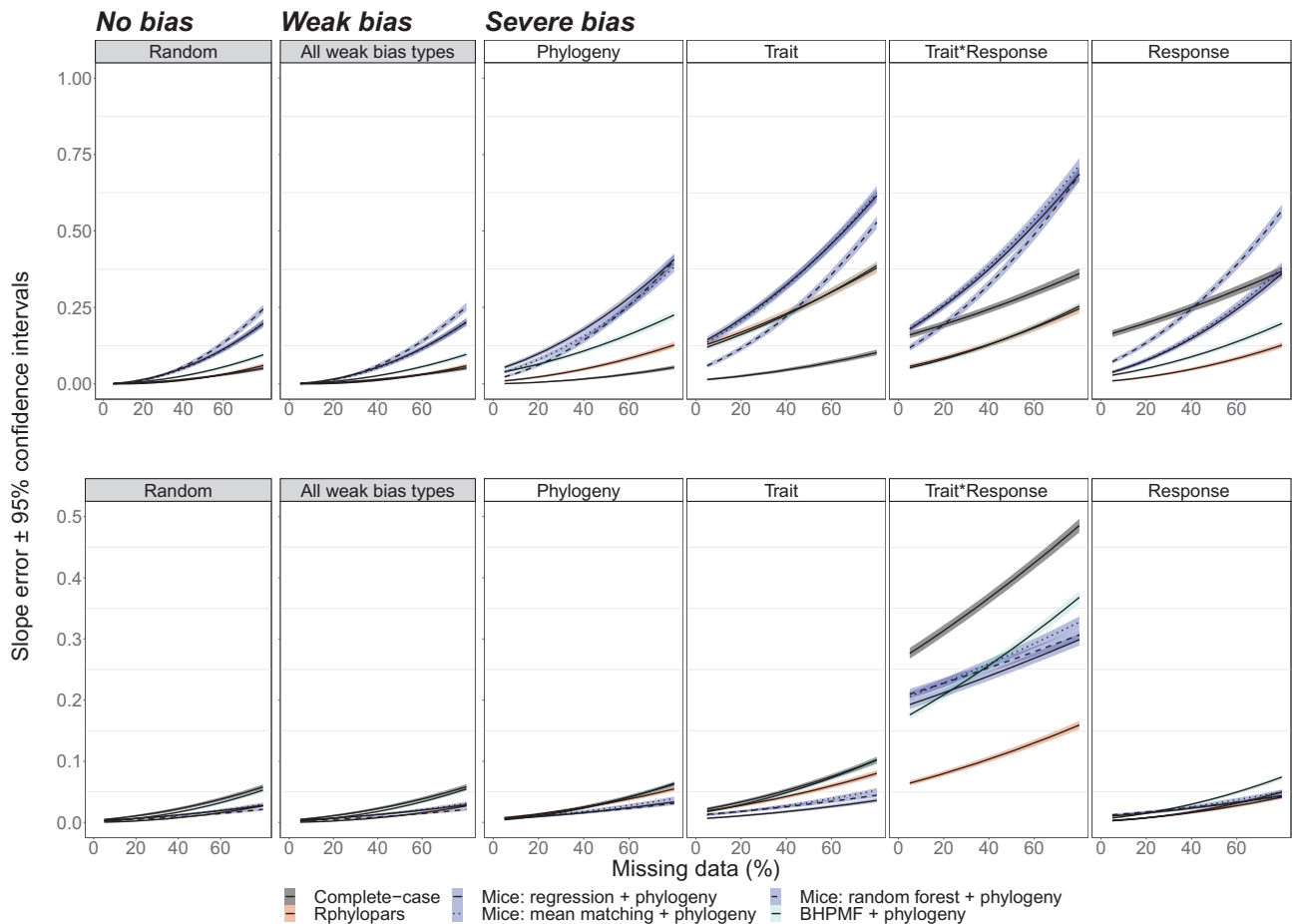
FIGURE 3 Absolute difference between the true response–trait slope coefficient and the slope coefficients in datasets with varying percentages of missing trait values (missingness), removed under a series of bias types. Missing values were imputed under five phylogenetic approaches or treated as complete-case analysis. The top row of panels shows datasets in which the true slope was positive (*r* c. .7), and the bottom row shows datasets with no relationship (*r* c. 0). Lines depict the marginal effect of missingness and bias type from regression models and were averaged across other predictors: seed, response in imputation and between-trait correlation. For plots split based on response in imputation, see the Supporting Information (Appendix B6). Confidence intervals were derived from 500 bootstrap simulations and depict the upper and lower bounds (95%). Note that the range in the *y* axis differs between top and bottom panels [Colour figure can be viewed at wileyonlinelibrary.com]

of imputation and slope errors). Table 1 summarizes our findings as warnings and recommendations.

4.1 | Which method performs best?

Rphylopars was the best overall imputation method in our study. However, we found scenarios where complete-case analysis maintained the response–trait relationship better, particularly compared with *Mice* and *BHPMF* imputation (but also outperforming *Rphylopars* under some severe biases). Our analyses, and others from the medical literature (Mukaka et al., 2016), show that imputation is not always the best solution to handle missing data. Although imputation methods in ecology are not yet widely used, the use of imputation has been recommended over complete-case analysis in recent publications (Kim et al., 2018; Penone et al., 2014). Here, by expanding on the scenarios explored in previous studies, we show that imputation can lead to errors in some conditions. For example,

when there was no true relationship between the response and trait, *Mice* approaches performed well. However, when there was a positive relationship, *Mice* did poorly even after the substantial improvement resulting from inclusion of the response in the imputation (Supporting Information, Appendix B, Figure B6.3), with increases in missing data gradually shifting the positive response–trait relationship towards zero. Further investigation of *Mice* is required, because in this scenario we might expect inflated noise around the slope in *Mice* to cause more type 2 errors (reporting no relationship when one is present), but we would not expect *Mice* systematically to shift the slope itself.

This poor performance of *Mice* is particularly surprising, because we made a concerted effort to optimize the performance of *Mice* (see Supporting Information, Appendix A4). However, the issues we have identified might be relevant only to our scenarios (imputing missing traits for phylogenetic comparative studies) and might not reflect on *Mice* or multiple imputation as a whole, which are considered throughout the literature as the "gold-standard" imputation

Warnings ⚠ and recommendations ✓	
⚠	Carefully select the taxonomic scope of the study, ensuring that species are distributed across the phylogeny and trait space. If any clades or areas of the trait space are nearly or entirely absent, do not draw inferences about them and exclude them from the study to prevent severe biases.
✓	Report which species/clades are included in the study and which species/clades have been removed to limit bias. Provide descriptive statistics or distribution plots for analysed trait values.
⚠	Every imputation approach produced inaccurate values, even with as little as 5% missing data. Slope errors consistently exceeded 0.1 when > 40% of the values were missing or when a severe bias was present.
⚠	Imputation is not always the best approach. Complete-case analysis performs better than the tested imputation methods in some cases.
✓	If using imputation, *Rphylopars* is the best approach for handling missing continuous data, resulting in smaller overall imputation and slope errors.
✓	If using *Rphylopars* or *BHPMF*, do not include the response in the imputation. If using *Mice*, including the response is beneficial.
✓	Include phylogenetic information when using imputation if possible. If a phylogeny is unavailable but a taxonomy is available, use *BHPMF*. If there is no phylogeny or taxonomy information, use *Mice* random forest or the observation-only *BHPMF*.
✓	To assist in detecting biases and the subsequently high imputation and slope errors, assess phylogenetic clustering, in addition to the change in the mean and change in the slope before and after imputation.
✓	Report the amount of missing information that was imputed and where this information falls on the phylogeny, trait and response (if applicable).

TABLE 1 Warnings and recommendations for handling missing trait values

approach (Van Buuren, 2012). Furthermore, despite making an effort to optimize the performance of *Mice*, there are a variety of *Mice* extensions and other multiple imputation approaches that might have fared better and could be tested in future comparisons, such as Multilevel Joint Modelling (Quartagno et al., 2019) or *Mice*: Random indicator method for non-ignorable data (Van Buuren & Groothuis-Oudshoorn, 2011).

One particular issue with *Mice* was the way in which biases interacted with the phylogeny during phylogenetic eigenvector selection. As a control, we estimated the number of eigenvectors when no values were missing. In this scenario, most datasets had between six and 16 eigenvectors, but under a severe trait bias the number of eigenvectors frequently surpassed 20, and under a response bias it rarely reached six. This discrepancy in the number of eigenvectors is likely to explain why incorporating phylogenetic information in *Mice* sometimes resulted in greater imputation and slope errors under a severe bias (Supporting Information, Appendix B, Figure B1.1). Given these findings, we revise the advice of Penone et al. (2014) and Kim et al. (2018), and suggest that phylogenetic information should only be included in *Mice* imputation when data are missing with no bias or a weak bias. Or, other *Mice* options that allow a phylogeny to be incorporated as a hierarchical structure, similar to that used by *BHPMF*, should be used. Further work is needed to establish how different biases

alter phylogenetic eigenvector selection and the downstream imputation and slope errors.

Unlike *Mice*, we suspect that the performance of *BHPMF* could be enhanced further (see Supporting Information, Appendix A4). Most notably, given that *BHPMF* does not allow imputation for species with no trait observations, we forced *BHPMF* to impute values by adding a dummy fully populated variable. This allowed us to compare the performance of *BHPMF* across all biases and levels of missing data and did not clearly affect the performance of *BHPMF* (Supporting Information, Appendix A, Figure A2.3). This feature of *BHPMF* could hinder the generality and taxonomic scale of studies but might also be beneficial if it deters researchers from imputing values in cases with lots of missing data (where imputation errors are greatest). However, removal of species with no trait values represents a form of complete-case analysis that could lead to biases and erroneous inferences.

Categorical traits are a common data type in ecological and evolutionary research and cannot be imputed using *Rphylopars* but can be handled by *BHPMF* and *Mice*. There has been limited assessment of the performance of categorical imputation, and available evaluations have delivered varied results (Akande et al., 2017; Kim et al., 2018; Stekhoven & Buhlmann, 2012). Future work exploring imputation errors and biases with categorical data would be valuable to guide researchers confronting missing values. Future

work could also determine whether machine learning approaches, such as *Mice* random forest and *BHPMF*, would perform better with larger trait datasets (e.g., > 500 simulated species used in the present study).

4.2 | Are imputed values and relationships accurate?

The threshold for deciding whether imputation is accurate depends on the research question. For example, in *Rphylopars*, with 5% of data missing under no bias (best possible scenario), the mean imputation error was 0.56. If we assume the trait data have been ln-transformed (base *e*), such error would mean that the mass of an African Elephant weighing 6,000 kg ($e^{8.7}$) would be imputed with values as low as 3,430 kg ($e^{8.7-0.56}$) or as high as 10,500 kg ($e^{8.7+0.56}$). This error is worrying, especially considering that *Rphylopars* is the most accurate imputation approach and that we used the most favourable missing data scenario in this example. This finding suggests that imputation is not accurate enough to estimate trait values for individual species or records. As such, any imputed values should be interpreted with great caution. Fortunately, trait values are more commonly imputed to establish relationships, in which case our results are less concerning. In linear regressions between a response and imputed traits, the difference between the *Rphylopars* slope and true slope was < 0.05 at every level of missing data (except for severe biases). In many cases, this would be deemed an acceptable amount of error and the same qualitative message, with a trend in the same direction (positive or negative) and not differing significantly from the true slope in most cases (Supporting Information, Appendix B, Figures B4.8–B4.11). However, this error would be large and could lead to qualitatively different messages in the context of debates about the true value of allometric exponents (Isaac & Carbone, 2010). Thus, unless the dataset is complete we recommend that results should be interpreted cautiously, regardless of whether imputation or complete-case analysis is used for the analyses.

Although different errors might be acceptable for different questions, our results show that analysing datasets where values are missing with a severe bias (MNAR) can lead to very wrong conclusions, especially when the bias acted on both the trait and response. This bias type was not tested by Penone et al. (2014), but it is likely to be common in ecology and evolution, because both trait databases (González-Suárez et al., 2012) and response values are biased (Boakes et al., 2010; Troudet et al., 2017). In some cases, a severe trait and response bias shifted a positive response–trait relationship into no relationship, or even a negative relationship (Figure 3). Overall, the methods we tested are unsuitable when a severe bias is present. However, there are imputation options, beyond the scope of the present study, designed specifically for severely biased MNAR data (Molenberghs et al., 2015). These MNAR options add a term to the imputation model to account for the bias. In common methods, such as selection, pattern-mixture and shared parameter models, this term describes a distribution aimed at explaining the mechanism by which data are missing. The parameters in these distributions

(sometimes informed by expert opinion) can have a substantial impact on results, meaning that sensitivity analysis becomes increasingly important. If a severe bias is suspected and the missing data mechanism cannot be accounted for by incorporation of additional data (e.g., other traits, phylogeny, or other spatial or temporal information), these MNAR methods should be explored. However, the main challenge will be in the detection of the severe bias in the first place. Familiarity with the dataset, accompanied by careful checks of the distribution of the data across space, time and the phylogeny, in addition to the trait and response ranges, might help. Furthermore, we recommend accounting for biases in missing datasets before designing research, especially in phylogenetic comparative studies, where severe biases could be reduced simply by trimming the scale of the study and its conclusions to the better represented groups.

4.3 | Should the response be included in the imputation?

Including the response within the imputation substantially decreased imputation and slope errors in *Mice* (Moons et al., 2006; Sterne et al., 2009) and made its performance almost comparable to *BHPMF* and *Rphylopars*. However, for *BHPMF* and *Rphylopars*, including the response had little effect or a negative effect. We are unsure why including the response might negatively affect the performance of *BHPMF*, but for *Rphylopars* we hypothesize that it is attributable to the way in which the phylogeny is incorporated. If the response is not associated to the phylogeny, including the response might skew the phylogenetic–trait covariance matrix, affecting the performance of *Rphylopars*. In contrast, the phylogenetic eigenvectors that are appended to the *Mice* imputation act more like weakly correlated traits; therefore, the benefit of adding a highly correlated response variable is clear. From this, it seems broadly advisable to include the response within *Mice* imputation and to exclude it from *Rphylopars* and *BHPMF*. However, caution is needed, because we suspect that these conclusions might contain caveats that warrant further research. For example, under the severe trait and response bias when there was no true relationship between the response and trait, imputation resulted in a significant negative slope, particularly when the response was used in the imputation (Supporting Information, Appendix B, Figures B6.5–B6.6). This is evidence that including the response in the imputation of trait values, which will then be modelled back against the response, can cause a circularity problem. Nevertheless, when using *Mice*, this detrimental effect was small compared with the overall gains from incorporating the response in the imputation.

4.4 | Can we predict whether imputation is advisable for a given dataset?

Within our work, we identify four ways in which data should be scrutinized before and after imputation to assist with bias

detection, measuring: missingness, phylogenetic clustering, a change in mean and a change in slope. These metrics should not be used as a free pass to claim that the imputation is valid, because no method consistently detected bias (for example, finding no change in slope could occur if both imputation and complete-case analyses are equally wrong). Instead, these metrics should be used alongside careful scrutiny of the data, viewing the imputation process holistically.

Our proposed protocol includes four steps:

1. Explore the data to consider representation of the group of interest (in both trait and response) and assess the potential for severe bias.
2. Compare the distribution of trait data before and after imputation.
3. Use expert opinion and information on closely related species to determine whether imputed values are plausible.
4. Use available tools to assess imputation results. *Rphylopars* and *BHPMF* currently lack imputation exploration functions, but custom checks can be created and adapted from the wide range offered in *Mice* (Van Buuren & Groothuis-Oudshoorn, 2011). *Rphylopars* and *BHPMF* produce uncertainty estimates for each imputed value, which could be scrutinized or, potentially, added to models to inflate noise and make inference more robust in these single imputation approaches. Furthermore, if a phylogenetic imputation approach is used, it is important to consider phylogenetic signal and branch length; otherwise, the phylogeny might add noise (Molina-Venegas et al., 2018).

Notwithstanding these guidelines, gaps remain in the ecological and evolutionary literature on imputation. Three important future steps would be: (a) to explore imputation methods and errors with categorical traits; (b) to validate imputations with non-simulated trait datasets as they become increasingly populated; and (c) to improve guidance on profiling data pre- and post-imputation. Finally, with recent reports of shifts away from fieldwork and into a more quantitative and modelling-based ecology (Ríos-Saldaña et al., 2018), it is important to note that the foundation for any trait-based analysis is the trait values, which can only become available from fieldwork and data compilation. There is still a crucial need to go out into the field and collect data, particularly on poorly studied species, traits and regions.

ORCID
Thomas F. Johnson https://orcid.org/0000-0002-6363-1825
Manuela González-Suárez https://orcid.org/0000-0001-5069-8900

REFERENCES
Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, *71*, 162–170.

Allen, W. L., Street, S. E., & Capellini, I. (2017). Fast life history traits promote invasion success in amphibians and reptiles. *Ecology Letters*, *20*, 222–230.

Baker, J., Humphries, S., Ferguson-Gow, H., Meade, A., & Venditti, C. (2020). Rapid decreases in relative testes mass among monogamous birds but not in other vertebrates. *Ecology Letters*, *23*, 283–292.

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*, e1000385.

Cohen, J., Cohen, P., Stephen, G., & Leona, S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London, UK: Routledge.

Diniz, J. A. F., Villalobos, F., Bini, L. M., & Bini, L. M. (2015). The best of both worlds: Phylogenetic eigenvector regression and mapping. *Genetics and Molecular Biology*, *38*, 396–400.

González-Suárez, M., Bacher, S., & Jeschke, J. M. (2015). Intraspecific trait variation is correlated with establishment success of alien mammals. *The American Naturalist*, *185*, 737–746.

González-Suárez, M., Lucas, P. M., & Revilla, E. (2012). Biases in comparative analyses of extinction risk: Mind the gap. *Journal of Animal Ecology*, *81*, 1211–1222.

Goolsby, E. W., Bruggeman, J., & Ané, C. (2017). Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, *8*, 22–27.

Isaac, N. J. B., & Carbone, C. (2010). Why are metabolic scaling exponents so controversial? Quantifying variance and testing hypotheses. *Ecology Letters*, *13*, 728–735.

Jones K. E., Bielby J., Cardillo M., Fritz S. A., O'Dell J., Orme C. D. L., ... Purvis A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*, 2648–2648. https://doi.org/10.1890/08-1494.1

Josse, J., Tierney, N., & Vialaneix, N. (2020). CRAN task view: Missing data. Retrieved from https://cran.r-project.org/web/views/Missing-Data.html

Kim, S. W., Blomberg, S. P., & Pandolfi, J. M. (2018). Transcending data gaps: A framework to reduce inferential errors in ecological analyses. *Ecology Letters*, *21*, 1200–1210.

Lancaster, L. T., Morrison, G., & Fitt, R. N. (2017). Life history trade-offs, the intensity of competition, and coexistence in novel and evolving communities under climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1712), 20160046.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, B., & Verbeke, G. (2015). *Handbook of missing data methodology*. London, UK: CRC Press.

Molina-Venegas, R., Moreno-Saiz, J. C., Castro Parga, I., Davies, T. J., Peres-Neto, P. R., & Rodríguez, M. (2018). Assessing among-lineage variability in phylogenetic imputation of functional trait datasets. *Ecography*, *41*, 1740–1749.

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, *59*, 1092–1101.

Mukaka, M., White, S. A., Terlouw, D. J., Mwapasa, V., Kalilani-Phiri, L., & Faragher, E. B. (2016). Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials*, *17*, 341.

Pacifici, M., Visconti, P., Butchart, S. H. M., Watson, J. E. M., Cassola, F. M., & Rondinini, C. (2017). Species' traits influenced their response to recent climate change. *Nature Climate Change*, *7*, 205–208.

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., … Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*, 961–970.

Poyatos, R., Sus, O., Badiella, L., Mencuccini, M., & Martínez-Vilalta, J. (2018). Gap-filling a spatially explicit plant trait database: Comparing imputation methods and different levels of environmental information. *Biogeosciences*, *15*, 2601–2617.

Quartagno, M., Grund, S., & Carpenter, J. (2019). jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal*, *11*, 205–228.

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from https://www.r-project.org/

Ríos-Saldaña, C. A., Delibes-Mateos, M., & Ferreira, C. C. (2018). Are fieldwork studies being relegated to second place in conservation science? *Global Ecology and Conservation*, *14*, e00389.

Roth, T., Allan, E., Pearman, P. B., & Amrhein, V. (2018). Functional ecology and imperfect detection of species. *Methods in Ecology and Evolution*, *9*, 917–928.

Sandel, B., Gutiérrez, A. G., Reich, P. B., Schrodt, F., Dickie, J., & Kattge, J. (2015). Estimating the missing species bias in plant trait measurements. *Journal of Vegetation Science*, *26*, 828–838.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., … Reich, P. B. (2015). BHPMF – A hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, *24*, 1510–1521.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ Clinical Research*, *338*, b2393.

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

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*, 9132.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman and Hall/CRC.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *10*, 1–68.

Webb, C. T., Hoeting, J. A., Ames, G. M., Pyne, M. I., & LeRoy Poff, N. (2010). A structured and dynamic framework to advance traits-based theory and prediction in ecology. *Ecology Letters*, *13*, 267–283.

BIOSKETCH

This aim of this project was to clarify and expand on guidance for handling missing values, specifically in reference to trait data. We wanted to examine the performance of imputation in comparative analyses to identify weaknesses with the approach and encourage robust missing data practices in ecology.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.