Paleobiology, 46(4), 2020, pp. 435–444
DOI: 10.1017/pab.2020.30

PALEOBIOLOGY
A PUBLICATION OF THE Paleontological SOCIETY

On The Record

Seven rules for simulations in paleobiology

Joëlle Barido-Sottani, Erin E. Saupe, Tara M. Smiley, Laura C. Soul, April M. Wright, and Rachel C. M. Warnock

Abstract.—Simulations are playing an increasingly important role in paleobiology. When designing a simulation study, many decisions have to be made and common challenges will be encountered along the way. Here, we outline seven rules for executing a good simulation study. We cover topics including the choice of study question, the empirical data used as a basis for the study, statistical and methodological concerns, how to validate the study, and how to ensure it can be reproduced and extended by others. We hope that these rules and the accompanying examples will guide paleobiologists when using simulation tools to address fundamental questions about the evolution of life.

Joëlle Barido-Sottani. Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, 50011, U.S.A.

Erin E. Saupe. Department of Earth Sciences, University of Oxford, Oxford, OX1 3AN, U.K.

Tara M. Smiley, Environmental Resilience Institute, Indiana University, Bloomington, Indiana 47408, U.S.A.

Laura C. Soul. National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, U.S.A.

April M. Wright Department of Biological Sciences, Southeastern Louisiana University, Hammond, Louisiana 70402, U.S.A.

Rachel C. M. Warnock. GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg, Loewenichstraße 28, 91054 Erlangen, Germany. E-mail: rachel.warnock@fau.de

Accepted: 6 July 2020

Introduction

Simulations have a long history in paleobiology (Raup and Gould 1974; Raup 1981, 1982; Holland 1995; Foote 1996). The past decade, however, has seen a proliferation of publications that include them, partly due to the increase in available computational resources, and the parallel development of simulation tools (e.g., FossilSim [Barido-Sottani et al. 2019b], paleotree [Bapst 2012], REvoSim [Garwood et al. 2019]). Simulations can be applied to many different problems within paleontology, including validating method implementation, comparing different protocols, understanding model assumptions and the effects of model violations, generating null models, and exploring the outcomes of different evolutionary and ecological scenarios. Simulations can also be included under the umbrella of resampling methods or Monte Carlo analyses used in paleontology (Kowalewski and Novack-Gottshall 2010).

Designing a simulation study can be daunting, and failure to implement a well-formulated plan can lead to wasted time and computational resources. A poorly designed simulation can produce results that are hard to interpret or even misleading. A well-designed simulation pipeline, however, will help to communicate your process and results clearly. Once an idea has been formulated for a simulation study, it can be tempting to jump straight into implementation. However, designing and testing a cohesive pipeline at the beginning of a study can save time in the long run, help avoid common pitfalls, and allow for better interpretation of simulation findings.

Here, we provide some general guidelines for designing simulations in paleobiology. These guidelines are derived from the authors'

0094-8373/20

personal experiences (and mistakes) in designing and publishing simulations. As such, they reflect the kinds of questions we ask in our own research. Our collective research in this area encompasses the most common applications in paleobiology, including simulations of phylogeny, discrete and continuous trait change, diversification dynamics, fossil preservation and recovery, tectonics, sea level change and geographic distributions, and other processes that operate on macroevolutionary timescales. However, much of the information is generalizable to other areas. A summary of our suggested rules is shown in Figure 1. The citations throughout showcase various aspects of simulation studies done well and provide examples that could be copied and adapted for new questions.

Rule 1. Establish a Clear and Specific Scientific Question

The first step in any simulation study is to establish precisely the scientific question the study will address. This question will inform simulation design and so should be settled early. It is preferable to keep the study question simple and to focus on only a few sources of variation in the simulations. This makes it easier to clearly disentangle the relative contributions of the processes and variables being tested (e.g., compare and contrast the individual *vs.* combined effects of the variables you alter during the simulations).

For example, Soul and Friedman (2017) studied the effect of ancestor–descendant pairs under different preservation rates on measures of the phylogenetic clustering of extinction. Bapst (2014) showed the effect of different paleontological phylogenetic time-scaling approaches on a variety of downstream analyses. Both of these studies used simulations in which one variable was changed while the others were held constant and focused on a few important sources of bias, rather than testing for the effect of every possible bias.

Rule 2. Determine a Clear Analytical Setup, Informed by Empirical Data

Simulations are versatile tools that have been incorporated into paleontological research in a wide variety of ways. Regardless of how you intend to make use of simulations in your study, the analytical conditions should be established early, and most decisions regarding conditions should be informed by empirical data. We note that simulation studies in paleobiology often do not provide explicit empirical justification for all (sometimes any) parameter choices, even if empirical data were ultimately used to inform decisions. For this reason, identifying empirical datasets that are relevant to your research question is an important initial step. Most frequently, empirical data are used within simulation studies to identify suitable values for input parameters, for example, identifying values for rates of character change, sampling, or diversification. Empirical data are often used to ensure that simulations emulate and are generalizable to real-world scenarios or to identify important features of data that should be reproduced by the simulations. You should familiarize yourself with these data early on, so that you can both recognize the critical features that need to be matched by a simulation and assess whether the outcome of your simulation makes empirical sense (i.e., is biologically reasonable).

The next step is to decide what type of simulations can be used to answer your question. In the first instance, this means deciding whether an explicit model will be used to generate the simulated data, or whether the goal of the simulation is to produce something representative of empirical data without attempting to match a specific generating process. For example, to compare different methods for phylogenetic estimation, rather than assume an explicit model of evolution, Puttick et al. (2019) simulated discrete character matrices by generating characters at random until the resulting matrices matched distributions of empirical estimates of homoplasy. See also Fraser (2017), who used a similar approach to simulate latitudinal richness gradients. It is also possible to combine these approaches and use an explicit model to simulate data that resemble empirical data (Novack-Gottshall 2016; Saupe et al. 2019b).

If your simulation is method focused, it is time to decide whether an empirical case study should be included. Including an empirical

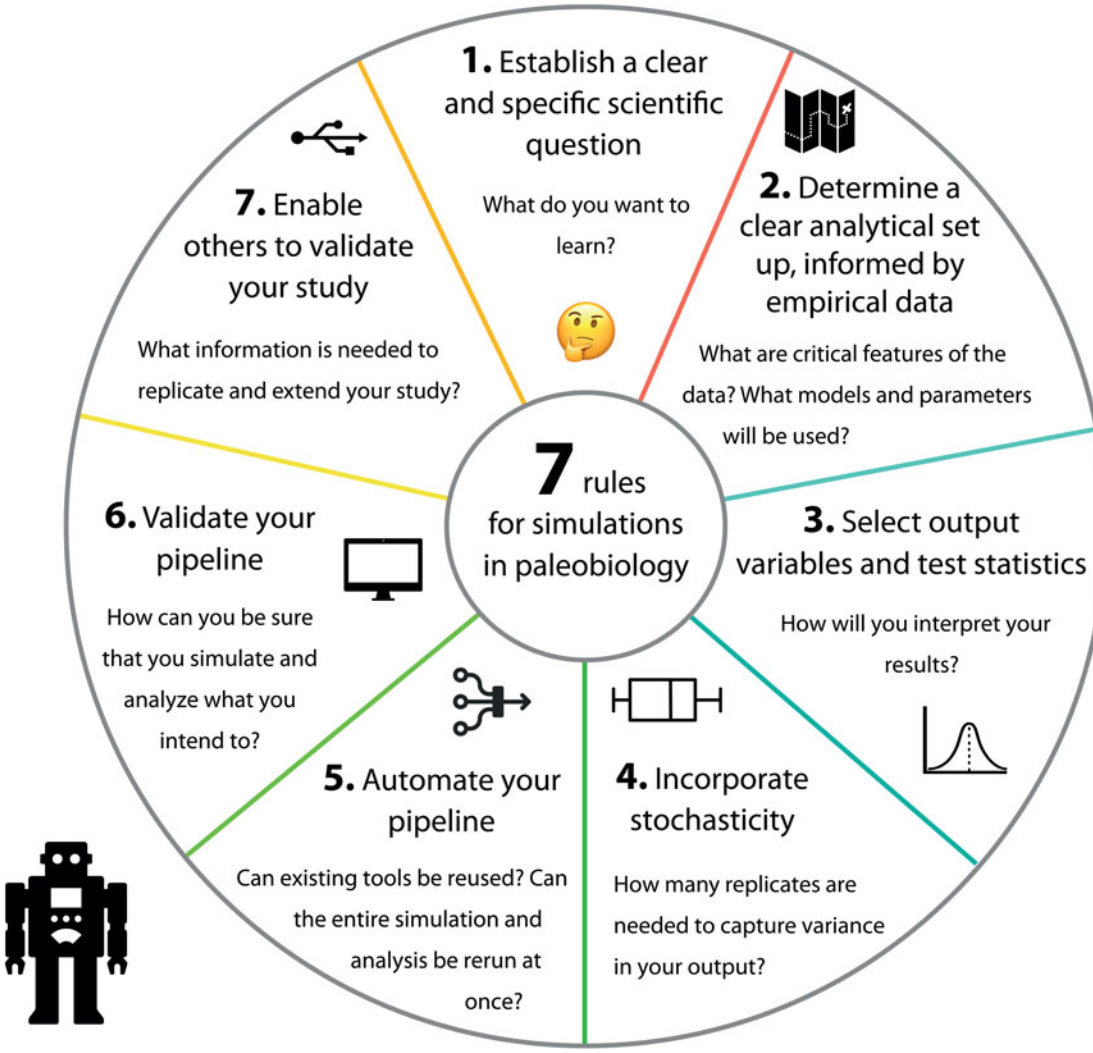

FIGURE 1. Visual summary of our proposed rules for good simulation study design. Each rule is associated with the main question(s) that the rule is designed to address.

case study can illustrate how results apply to real-life scenarios, which can help in communicating the impact or potential application of your results to other researchers. For example, Soul and Friedman (2017) used simulations to identify evolutionary and analytical scenarios that result in biased estimates of phylogenetic clustering of extinction. They then demonstrated the biological implications of their simulation results using an analysis of Mesozoic tetrapods. Similarly, Barido-Sottani et al. (2019a) examined the impact of fossil-age uncertainty in phylogenetic inference, using both simulations and an empirical case study. Whether to include a case study should be considered carefully, however, as it may not bring value to your paper if the case study is particularly time-consuming to set up or outside the range of your expertise or if the general patterns studied in your simulation have already been shown for empirical data (e.g., see Warnock et al. 2017). Empirical examples can be narrow (e.g., single-taxon application for demonstration purposes, as in Wang et al. [2016]) or broad (e.g., clade-level dynamics, as in Silvestro et al. [2018]), depending on the authors' aims and data availability.

Once the appropriate model(s) and data have been chosen, the remainder of the analytical

setup can be finalized. We provide a non-comprehensive list of considerations and examples for doing so in the following sections. Key decisions required include what the input parameter values are; which input parameters will be varied, and how they will vary within or across simulation iterations; how comparisons will be made between simulation inputs, outputs, and empirical data; and which model(s) you will use, if any, and how they will be configured (for instance, priors in a Bayesian analysis). The reasoning or literature to support all of these decisions should be recorded so that others can easily understand where they came from (see rule 7).

Choosing Input Values.—Your selected empirical dataset(s) can be used to establish realistic simulation conditions. Some input parameters will not be known precisely or cannot be extrapolated easily from the data. However, it is usually possible to constrain biological parameters to within an order of magnitude based on available empirical data. For example, the long-term molecular substitution rate, required to simulate sequence alignments, is not known for many groups of species (e.g., most insects), but the best estimate is known to within approximately one or two orders of magnitude (Papadopoulou et al. 2010). As a further example, we may not know the true preservation and recovery rate through time for a particular clade, but we can estimate that it is an order of magnitude higher in some clades (e.g., Cenozoic planktic forams; Aze et al. 2011) than in others (e.g., Paleozoic crinoids; Foote and Raup 1996). Sometimes the empirical data can be incorporated into the simulation as an input (e.g., see Darroch and Saupe [2018], who used real geographic ranges as an input parameter, or Foote [1999], who used real morphological data as a basis for simulating character evolution). In this case, it is important to check that the simulated datasets at the end of the process do in fact match the original data, as complex simulation processes can bias the simulated data away from the original distribution. Features of empirical data that are known to affect the method(s) tested in the study, or the output variables used, should be matched closely. For example, fossil sampling rate and heterogeneity influence analytical output from many methods applied to the fossil record, so simulations should match the sampling rate of the empirical dataset under investigation or explore a range of sampling rates (Holland and Patzkowsky 1999; Soul and Friedman 2017; Warnock et al. 2017; Smiley 2018). Identifying a broad range of empirical datasets from which to draw realistic input parameter values, and matching the simulation inputs with the full distribution of empirical values rather than the mean or another point estimate, can make the simulated output more representative of real datasets and thus generate more broadly applicable results.

Establishing Criteria for Acceptable Correspondence.—It is likely you will want to measure how closely simulation or analytical outputs match with some other value, be that from empirical data or original simulation inputs. Common reasons to measure such correspondence include checking whether simulated data resemble the intended empirical data, identifying which of a set of candidate models can be used to simulate something resembling real data (Raup 1982; Alroy 2001; Green et al. 2011; Smiley 2018), or establishing the accuracy and precision of an analytical method applied to simulated data (Lane et al. 2005; Bapst 2013, 2014; O'Connor and Wills 2016; Soul and Friedman 2017), among others. It is critical to consider the criteria used to evaluate the match between simulated and target values, the statistical measure of offset between them, and how much deviation is acceptable. If the simulation is designed to resemble a specific empirical dataset, you should be aware of features of the data that the simulations have not modeled or recreated and that may affect your ability to establish correspondence between the simulation outputs and the target data. The ways in which simulated data deviate from empirical data should be noted and discussed. If simulations are being used to make predictions, the data used to identify suitable model inputs should be separate from the data used for testing adequacy of predictions, to avoid circularity.

Establishing Baselines.—It is often useful in simulation studies to include best-case and/or worst-case scenarios suitable for your study context, to establish a baseline level for interpreting your results (see rule 3) and ensure

your pipeline is free of errors (see rule 6). In studies designed to test a modeling approach, the generating process should conform to the assumptions of the model used for inference for at least one set of simulation replicates. This way, if intentional model violations are part of your study question, they will be distinguishable from stochastic variation (see rule 4). For example, to assess the performance of a new inference framework for estimating species divergence times, Heath et al. (2014) first simulated data under the model used for inference (i.e., the fossilized birth–death process) before exploring more complex preservation scenarios that violated this model. Similarly, Wright and Hillis (2014) simulated morphological character data under a simple single-parameter model to establish a baseline level of expected error in phylogenetic estimation from discrete data. A subsequent study (Wright et al. 2016) then examined situations in which the model assumptions were violated.

In studies designed to understand empirical patterns or model a target empirical dataset, it is important to simulate data under a null scenario or model that will serve as a baseline for comparison with the other simulated conditions. This null model should fulfill the assumptions of the method or measure tested, and usually represents an ideal or error-free scenario. For instance, when testing the effects of preferential sampling of larger fossil specimens, a model with preferential sampling was compared with a null model of fully randomized sampling (Hawkins et al. 2018). Additionally, the null model can help test whether the simulation framework is functioning properly and whether the tested methods are performing as expected, before examining the influence of parameters on simulation outcomes.

Rule 3. Select Output Variables and Test Statistics

Output variables should be given careful consideration in advance and decided upon before examining the results. Choosing evaluation criteria in advance ensures they are incorporated into the simulation pipeline and helps to avoid a scenario in which you need to rerun simulations to obtain necessary output. It also avoids a scenario in which the measures are decided upon after examination of results, which affects the statistical validity of your study.

Ideally, output variables will be well established in your field, targeted to the study question, and appropriate for use as a measure of quality (i.e., you should know in advance what counts as a "good" or "bad" result for each measure). For example, coverage is a measure of performance widely used in Bayesian statistics. If the data are analyzed under the same model that generated the data, a coverage of 0.95 is expected, meaning that for 95% of datasets the 95% credible interval (the Bayesian confidence interval) contains the true parameter value. Thus, as a rule of thumb, a value of 0.95 is considered "good," whereas values substantially below 0.95 are considered "bad." However, coverage is not suitable for all parameters, especially complex parameters, such as phylogenetic tree topology. Instead, a more direct metric of comparison is often used, such as that between the estimated and true simulated trees. One example is the Robinson-Foulds (RF) distance (Robinson and Foulds 1979, 1981), which quantifies how many bifurcations differ between two trees. This metric is often chosen, because reducing a phylogeny to a summary statistic is challenging, and the RF distance can be expressed as a percentage for easy interpretation.

Simulation exercises can be useful for establishing a baseline or threshold for variables (see rule 2) before evaluating the primary simulation results. For example, Silvestro et al. (2018) used a simulated training dataset to define corrected Akaike information criterion thresholds that would adequately distinguish between competing models of fossil preservation for individual datasets. These threshold values were then used to assess the performance of model fitting in the software PyRate using the primary set of simulations.

Careful consideration must be given when using statistical tests to make quantitative comparisons between data simulated under different models. Frequentist statistical significance is calculated based on number of independent replicates, which in simulation studies can be arbitrarily large and can therefore result in

minor differences being significant. Additionally, null hypotheses in classical statistical tests are often inappropriate for the kinds of scenarios simulation studies address. For instance, significance testing to establish the influence of changes in parameters on simulated output variables is invalid, as the null hypothesis is that the two generating models are the same, which is known to be false from the outset. Thus, it is preferable to measure not whether there is a difference in analytical results, but what the magnitude of the difference is and whether it is biologically important, that is, calculate the effect size. For a review of these issues, see White et al. (2014).

In contrast, using significance tests when comparing simulated and empirical data is often valid, as the generating model for the empirical data is unknown. However, it is likely that the empirical model is not among the tested simulation models, so the best-fitting model should not be interpreted as necessarily matching the true generating process (Brown 2014). In all cases, the underlying assumptions of the test should be stated clearly (Siegfried 2010). Available statistical tests can be tailored to a given question, along with the simulation input, providing researchers with both enormous flexibility and the responsibility to ensure that the chosen test is appropriate (Kowalewski and Novack-Gottshall 2010).

In general, quantitative variables allow for more detailed comparisons and are less subject to interpretation than qualitative output and are thus preferable. However, qualitative output can be useful to demonstrate patterns that are hard to summarize using numerical measures, such as time series showing the clustering of extinction events driven by stratigraphic biases (Holland and Patzkowsky 2015). In this case, the criteria used to interpret these patterns and their similarities or differences should be detailed clearly to make the study reproducible and extendable.

Rule 4. Incorporate Stochasticity

Simulations should be stochastic. That is, the model or parameter values can be fixed, but the simulated datasets should incorporate random variation. Simulations should be stochastic, because there is variation inherent in biological systems and the generating processes used to model them. By failing to consider this variation, we risk obtaining results that are not representative of underlying processes or do not accurately reflect uncertainty in those estimates. This is true irrespective of whether we explicitly model the underlying process or use a target distribution based on empirical data to generate simulated datasets (see rule 2). Stochasticity is a feature of birth–death models (Nee 2006), but should also be built into other model dynamics; for example, ecospace filling (Novack-Gottshall 2016), morphological evolution (Foote 1999; Puttick et al. 2019), dispersal (Silvestro et al. 2016; Saupe et al. 2017, 2019b), and preservation (Holland 1995; Holland and Patzkowsky 1999).

The adequate number of replicates for each simulation condition will depend on the complexity and degree of randomness involved in the simulation framework. Comparing the tested simulation condition with a null model or best-case scenario, as recommended in rule 2, can help to calibrate the number of replicates required by providing a benchmark on the amount of noise generated by the simulation setup. As a general rule of thumb, most simulation studies in the current literature use between 100 and 1000 replicates; however, this choice is subjective.

Plotting variation in simulation outcomes across replicates is critical to distinguish between differences that can be attributed to study conditions versus differences that are simply due to stochasticity of the simulations. A high number of outliers or a very broad range of results can indicate the number of replicates is too low. Similarly, if the results under the same condition are clustered into several groups, it can indicate an underlying feature of the data is affecting the results.

If computational resources are limited, we argue it is better to focus on only a few sources of variation in simulation conditions and to execute them well (i.e., with large enough sample sizes and adequate run times), rather than obtain unreliable results due to low numbers of replicates. For more advice on selecting replication numbers, see Kowalewski and Novack-Gottshall (2010).

Rule 5. Automate Your Pipeline

In addition to careful study design, particular attention should be paid to the computational aspects of the simulation pipeline. In particular, it is important to consider availability of computational resources before designing the simulations, as these resources can limit the number of conditions or replicates that can be executed. When available, the use of high performance computing clusters is strongly recommended for large simulation studies, as they allow many replicate analyses to be run in parallel (Arora 2016).

Another important consideration is whether preexisting simulation tools can be applied to your study. A variety of software is already freely available, such as the R packages ape (Paradis et al. 2004), caper (Orme et al. 2018), FossilSim (Barido-Sottani et al. 2019b), geiger (Pennell et al. 2014), OUwie (Beaulieu and O'Meara 2020), paleotree (Bapst 2012), and phytools (Revell 2012). Reusing code that has already been extensively tested saves time and limits potential mistakes.

Previous code and packages will most likely only perform part of the simulation or analysis and will need to be tied together in a global pipeline. This pipeline should be automated as much as possible, as it most likely will need to be run more than once, and manual steps are time-consuming and error prone. Automation will minimize the amount of work involved in rerunning the simulation, if (when) mistakes are found or if the simulation design is extended.

Some steps in a simulation pipeline may not be easily automated. In these cases, manual steps should be well documented. For example, if a graphical user interface is required, the software version and options should be noted. If subjective or manual assessment of parameters is required, a single individual should be assigned to the task or the criteria for performing the task should be established beforehand. An example of a manual step in simulation studies is assessing the convergence of a Markov chain Monte Carlo (MCMC) inference.

Your simulation pipeline will eventually form the backbone of a methods section. Smart workflow design and implementation can readily facilitate future simulation studies that probe various links in the pipeline, and also provide an opportunity to revisit simulation parameters and assumptions in a flexible and forward-thinking manner.

Rule 6. Validate Your Pipeline

To guarantee the validity of simulation results, the pipeline should be tested thoroughly at each step. Many errors will not produce an obvious notification or software crash but instead generate invalid or biased simulated datasets or calculate wrong output metrics. Detecting these errors requires the user to think critically about the expected results and actively test that the output matches expectations. For instance, plotting summary statistics of the simulated data and comparing them with empirical values is a good way to ensure the simulation works and the overall design makes sense biologically, even if not explicitly using a target dataset (see rule 2). Capturing mistakes early in a study saves time, so we recommend using a modular design in which the pipeline is progressively assembled from single-step functions that each perform only a small part of the process. These small modules will be much easier to test individually. Pursuant to rule 5, the generation of summary statistics or plots can also be a step in this automated pipeline. Internal checks can be retained by either manual commenting out or a toggle option in your code (e.g., if/then statement to run internal checks), allowing for the trade-off between speed and careful checking. Checkpoints in the technical workflow serve to further assess code performance incrementally, identify rate-limiting steps in the pipeline, and ensure optimization. See Gibert and Escarguel (2017) for a well-illustrated pipeline.

Applying good programming practices will help ensure your pipeline is free of errors. Automated code-analysis tools exist for all commonly used programming languages and can flag common mistakes or bad practices that lead to buggy or misleading code. Internal checks that ensure that some conditions are respected during the execution of the pipeline (for instance, that the value of a parameter fits within specified bounds), can alert you early if an error is introduced. Documenting the

code is also advisable, as this will prevent issues such as missing steps of the pipeline or using the wrong input files. Documentation is particularly important when several people are collaborating on the pipeline or when new analyses need to be run again after an extended period of time. Other good practices, such as the use of a version-control system (e.g., Git), will not prevent errors but can mitigate their impact. First, version control can easily show at which point in time a bug or error was introduced, and thus which analyses need to be rerun. Second, using a shared repository helps ensure that everyone executing the pipeline is running the same version of the code, and thus that bug fixes are applied immediately to all new analyses.

For larger code projects, and particularly for source code that is intended to be published and reused as a package or tool, more extensive validation is needed. Automated tests, which can be run all at once without manual intervention from the developer, will alert you if an error appears or the behavior of the code changes. These tests are extremely valuable on large projects, where many different parts are interconnected and a small change can break seemingly unrelated functionality elsewhere in the code. Although they represent a significant investment of time to set up, they will pay off in time saved in terms of debugging and rerunning. Code reviews, ideally performed by someone who has not been involved in writing the code, are effective in identifying potential issues, misleading documentation, or missing features. These can be done informally within a team or as part of a formal review mechanism, such as the rOpenSci project for R packages.

Rule 7. Enable Others to Validate Your Study

Simulation design can result in biases that are not immediately obvious, even to the author. Therefore, methods need to be transparent to enable readers to determine the robustness of a simulation. Transparency requires that other researchers be allowed to examine your code and the data generated during a simulation project. This way, they can understand how a result was achieved, and how they can extend the simulation to address questions in which they are interested. Linking back to the suggestion to reuse code in rule 5: by expanding upon others' code and making scripts available, we will generate a more comparable and coherent body of research as a community.

The core component of transparency is to think about what someone would need if they were interested in publishing a follow-up study. This can involve documentation of design choices, source code, and all scripts used to process results. In general, decisions made during the simulation design should be documented carefully, along with associated reasoning. Providing justification for parameter choices is important for putting simulation results into context, especially for newcomers to the field, who are less likely to be familiar with empirical estimates. We recommend using comments within relevant scripts and/or as part of a comprehensive workflow documentation to accompany simulation code. For example, Liow et al. (2010) present figures representative of key sequential steps, demonstrating how data are transformed through the simulation pipeline (e.g., from tree simulation to stratigraphic ranges). These figures provide both a visual confirmation that the pipeline is executing as expected and a reader-friendly illustration of the steps involved. In general, documentation will help ensure the study design makes sense, avoids idiosyncratic features, and is appropriate to answer the question. It also makes the study easier to replicate or extend. If one intends simulation code to be a tool broadly used by the paleobiology community, online tutorials, R packages, or user-friendly software are key to facilitating implementation by new users.

Simulation code itself can be provided in a number of ways. Many journals allow code to be uploaded along with the article as supplementary material. Code can also be hosted through revision management websites, such as GitHub or GitLab. These websites are free of charge for both the researcher and anyone wishing to use your code. We also recommend retaining an archival version of the code actually used in the study (e.g., divorced from current working versions on GitHub). Starting conditions of simulations and intermediate data output may be important to document and

archive, depending on the nature of the study. Larger files, such as logs from MCMC analyses, can be uploaded as supplementary material upon paper acceptance at most journals.

By following data and code transparency practices and providing clear documentation of study design, researchers gain the additional advantage of making their work approachable and usable by nonexperts in quantitative or computational paleobiology. In this way, a broader audience can better understand simulation-based methodology, including its benefits, limitations, and implications for research on the fossil record.

Conclusions

As large-scale datasets, advanced computational resources, and increasingly sophisticated quantitative approaches become widely available and easy to use, simulations are becoming an essential part of paleobiology. This includes studies that seek to address predominantly empirical, rather than methodological, questions (Brocklehurst et al. 2018; Lewitus et al. 2018; Saupe et al. 2019a). Simulations make use of datasets for which the generating models and parameter values are known, and thus provide unique insights into the performance of commonly applied methods, the impact of specific data features on results, and insights into processes shaping the natural world. However, simulation studies are also complex and require specific expertise. In this review, we provide general guidelines for future simulation studies to help authors exploit simulations to their full potential. Note that these guidelines require a varied range of skills and expertise that will seldom be possessed by a single researcher. We therefore encourage that simulation studies be collaborative efforts, including data specialists familiar with relevant real-world datasets, computational specialists who have the programming expertise needed to design and implement the simulations, and statistical experts who can provide guidance on interpreting results.

Acknowledgments

We thank S. Holland and P. Novack-Gottshall for comments that greatly improved the article. J.B.S. was supported by funds from the National Science Foundation (U.S.A.), grant DBI-1759909. L.C.S. was supported by the Smithsonian National Museum of Natural History Deep Time Initiative. E.E.S. was supported by the Leverhulme Trust (grant no. DGR01020). A.M.W. was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (P20 GM103424-17). T.M.S. was supported by the Environmental Resilience Institute, funded by Indiana University's Prepared for Environmental Change Grand Challenge initiative. R.C.M.W. was supported by the ETH Zürich Postdoctoral Fellowship and Marie Curie Actions for People COFUND program.

Literature Cited

Alroy, J. 2001. A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. Science 292:1893–1896.

Arora, R. 2016. An introduction to big data, high performance computing, high-throughput computing, and hadoop. Pp. 1–12 *in* Conquering big data with high performance computing. Springer International Publishing, Cham, Switzerland.

Aze, T., T. H. G. Ezard, A. Purvis, H. K. Coxall, D. R. M. Stewart, B. S. Wade, and P. N. Pearson. 2011. A phylogeny of cenozoic macroperforate planktonic foraminifera from fossil data. Biological Reviews 86:900–927.

Bapst, D. W. 2012. paleotree: an R package for paleontological and phylogenetic analyses of evolution. Methods in Ecology and Evolution 3:803–807.

Bapst, D. W. 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods in Ecology and Evolution 4:724–733.

Bapst, D. W. 2014. Assessing the effect of time-scaling methods on phylogeny-based analyses in the fossil record. Paleobiology 40:331–351.

Barido-Sottani, J., G. Aguirre-Fernández, M. H. Hopkins, T. Stadler, and R. C. M. Warnock. 2019a. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth-death process. Proceedings of the Royal Society of London B 286:20190685.

Barido-Sottani, J., W. Pett, J. E. O'Reilly, and R. C. M. Warnock. 2019b. FossilSim: an R package for simulating fossil occurrence data under mechanistic models of preservation and recovery. Methods in Ecology and Evolution 10:835–840.

Beaulieu, J. M., and B. C. O'Meara. 2020. OUwie: analysis of evolutionary rates in an OU framework. https://cran.r-project.org/web/packages/OUwie.

Brocklehurst, N., E. M. Dunne, D. D. Cashmore, and J. Fröbisch. 2018. Physical and environmental drivers of Paleozoic tetrapod dispersal across Pangaea. Nature Communications 9:5216.

Brown, J. M. 2014. Predictive approaches to assessing the fit of evolutionary models. Systematic Biology 63:289–292.

Darroch, S. A., and E. E. Saupe. 2018. Reconstructing geographic range-size dynamics from fossil data. Paleobiology 44:25–39.

Foote, M. 1996. Models of morphological diversification. Pp. 62–8 *in* Evolutionary paleobiology. D. Jablonski, D. H. Erwin and J. H. Lipps, eds. University of Chicago Press, Chicago.

Foote, M. 1999. Morphological diversity in the evolutionary radiation of paleozoic and post-paleozoic crinoids. Paleobiology 25:1–115.
Foote, M., and D. M. Raup. 1996. Fossil preservation and the stratigraphic ranges of taxa. Paleobiology 22:121–140.
Fraser, D. 2017. Can latitudinal richness gradients be measured in the terrestrial fossil record? Paleobiology 43:479–494.
Garwood, R. J., A. R. Spencer, and M. D. Sutton. 2019. Revosim: organism-level simulation of macro and microevolution. Palaeontology 62:339–355.
Gibert, C., and G. Escarguel. 2017. Evaluating the accuracy of biodiversity changes through geologic times: from simulation to solution. Paleobiology 43:667–692.
Green, W. A., G. Hunt, S. L. Wing, and W. A. DiMichele. 2011. Does extinction wield an axe or pruning shears? How interactions between phylogeny and ecology affect patterns of extinction. Paleobiology 37:72–91.
Hawkins, A. D., M. Kowalewski, and S. Xiao. 2018. Breaking down the lithification bias: the effect of preferential sampling of larger specimens on the estimate of species richness, evenness, and average specimen size. Paleobiology 44:326–345.
Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences USA 111:E2957–E2966.
Holland, S. M. 1995. The stratigraphic distribution of fossils. Paleobiology 21:92–109.
Holland, S. M., and M. E. Patzkowsky. 1999. Models for simulating the fossil record. Geology 27:491–494.
Holland, S. M., and M. E. Patzkowsky. 2015. The stratigraphy of mass extinction. Palaeontology 58:903–924.
Kowalewski, M., and P. Novack-Gottshall. 2010. Resampling methods in paleontology. Paleontological Society Papers 16:19–54.
Lane, A., C. M. Janis, and J. J. Sepkoski. 2005. Estimating paleodiversities: a test of the taxic and phylogenetic methods. Paleobiology 31:21–34.
Lewitus, E., L. Bittner, S. Malviya, C. Bowler, and H. Morlon. 2018. Clade-specific diversification dynamics of marine diatoms since the Jurassic. Nature Ecology and Evolution 2:1715.
Liow, L. H., T. B. Quental, and C. R. Marshall. 2010. When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? Systematic Biology 59:646–659.
Nee, S. 2006. Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics 37:1–17.
Novack-Gottshall, P. M. 2016. General models of ecological diversification. II. Simulations and empirical applications. Paleobiology 42:209–239.
O'Connor, A., and M. A. Wills. 2016. Measuring stratigraphic congruence across trees, higher taxa, and time. Systematic Biology 65:792–811.
Orme, D., R. P. Freckleton, G. Thomas, T. Petzoldt, and S. A. Fritz. 2018. caper: comparative analyses of phylogenetics and evolution in R, R package version 1.01. https://cran.r-project.org/web/packages/caper.
Papadopoulou, A., I. Anastasiou, and A. P. Vogler. 2010. Revisiting the insect mitochondrial molecular clock: the mid-Aegean trench calibration. Molecular Biology and Evolution 27:1659–1672.
Paradis, E., J. Claude, and Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216–8.
Puttick, M. N., J. E. O'Reilly, D. Pisani, and P. C. J. Donoghue. 2019. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology 62:1–17.
Raup, D. M. 1981. Extinction: bad genes or bad luck? Acta Geològica Hispànica 16:25–33.
Raup, D. M. 1982. Biogeographic extinction: a feasibility test. Pp. 277–281 *in* L. T. Silver and P. H. Schultz, eds. Geological implications of impacts of large asteroids and comets on the Earth. Geological Society of America, Boulder, Colo.
Raup, D. M., and S. J. Gould. 1974. Stochastic simulation and evolution of morphology—towards a nomothetic paleontology. Systematic Biology 23:305–322.
Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.
Robinson, D., and L. Foulds. 1979. Comparison of weighted labelled trees. Pp. 119–126 *in* A. F. Horadam and W. D. Wallis, eds. Combinatorial mathematics VI. Lecture Notes in Mathematics. Springer, Berlin.
Robinson, D., and L. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147.
Saupe, E. E., N. Barve, H. L. Owens, J. C. Cooper, P. A. Hosner, and A. T. Peterson. 2017. Reconstructing ecological niche evolution when niches are incompletely characterized. Systematic Biology 67:428–438.
Saupe, E. E., A. Farnsworth, D. J. Lunt, N. Sagoo, K. V. Pham, and D. J. Field. 2019a. Climatic shifts drove major contractions in avian latitudinal distributions throughout the Cenozoic. Proceedings of the National Academy of Sciences USA 116:12895–12900.
Saupe, E. E., C. E. Myers, A. T. Peterson, J. Soberón, J. Singarayer, P. Valdes, and H. Qiao. 2019b. Spatio-temporal climate change contributes to latitudinal diversity gradients. Nature Ecology and Evolution 3:1419–1429.
Siegfried, T. 2010. Odds are, it's wrong: science fails to face the shortcomings of statistics. Science News 177:26–29.
Silvestro, D., A. Zizka, C. D. Bacon, B. Cascales-Minana, N. Salamin, and A. Antonelli. 2016. Fossil biogeography: a new model to infer dispersal, extinction and sampling from palaeontological data. Philosophical Transactions of the Royal Society of London B 371:20150225.
Silvestro, D., N. Salamin, A. Antonelli, and X. Meyer. 2018. Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. Paleobiology 45:546–570.
Smiley, T. M. 2018. Detecting diversification rates in relation to preservation and tectonic history from simulated fossil records. Paleobiology 44:1–24.
Soul, L. C., and M. Friedman. 2017. Bias in phylogenetic measurements of extinction and a case study of end-Permian tetrapods. Palaeontology 60:169–185.
Wang, S. C., P. J. Everson, H. J. Zhou, D. Park, and D. J. Chudzicki. 2016. Adaptive credible intervals on stratigraphic ranges when recovery potential is unknown. Paleobiology 42:240–256.
Warnock, R. C. M., Z. Yang, and P. C. J. Donoghue. 2017. Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution. Proceedings of the Royal Society of London B 284:20170227.
White, J. W., A. Rassweiler, J. F. Samhouri, A. C. Stier, and C. White. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. Oikos 123:385–388.
Wright, A. M., and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9: e109210.
Wright, A. M., G. T. Lloyd, and D. M. Hillis. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Systematic Biology 65:602–611.