

REVIEW

Widespread analytical pitfalls in empirical coexistence studies and a checklist for improving their statistical robustness

J. Christopher D. Terry¹  | David W. Armitage² 

¹Department of Biology, University of Oxford, Oxford, UK

²Integrative Community Ecology Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan

Correspondence

J. Christopher D. Terry
Email: christopher.terry@biology.ox.ac.uk

Funding information

Leverhulme Trust, Grant/Award Number: ECF-2022-666

Handling Editor: Miguel Acevedo

Abstract

1. Modern coexistence theory (MCT) offers a conceptually straightforward approach for connecting empirical observations with an elegant theoretical framework, gaining popularity rapidly over the past decade. However, beneath this surface-level simplicity lie various assumptions and subjective choices made during data analysis. These can lead researchers to draw qualitatively different conclusions from the same set of experiments. As the predictions of MCT studies are often treated as outcomes, and many readers and reviewers may not be familiar with the framework's assumptions, there is a particular risk of 'researcher degrees of freedom' inflating the confidence in results, thereby affecting reproducibility and predictive power.
2. To tackle these concerns, we introduce a checklist consisting of statistical best practices to promote more robust empirical applications of MCT. Our recommendations are organised into four categories: presentation and sharing of raw data, testing model assumptions and fits, managing uncertainty associated with model coefficients and incorporating this uncertainty into coexistence predictions.
3. We surveyed empirical MCT studies published over the past 15 years and discovered a high degree of variation in the level of statistical rigour and adherence to best practices. We present case studies to illustrate the dependence of results on seemingly innocuous choices among competition model structure and error distributions, which in some cases reversed the predicted coexistence outcomes. These results demonstrate how different analytical approaches can profoundly alter the interpretation of experimental results, underscoring the importance of carefully considering and thoroughly justifying each step taken in the analysis pathway.
4. Our checklist serves as a resource for authors and reviewers alike, providing guidance to strengthen the empirical foundation of empirical coexistence analyses. As the field of empirical MCT shifts from a descriptive, trailblazing phase to a stage of consolidation, we emphasise the need for caution when building upon the findings of earlier studies. To ensure that progress made in the field of ecological coexistence is based on robust and reliable evidence, it is crucial to subject our

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

predictions, conclusions and generalisability to a more rigorous assessment than is currently the trend.

KEYWORDS

competition, experiments, model selection, modern coexistence theory, uncertainty propagation

1 | INTRODUCTION

The quest to understand how similar species can avoid competitive exclusion is a central goal of community ecology (Chase & Leibold, 2003; Hutchinson, 1959) and has generated an extensive body of theory describing the formal conditions required for pairwise coexistence (Amarasekare, 2019; Barabás et al., 2018; Chesson, 2000). A key strength of this approach, frequently referred to as modern coexistence theory (MCT), is to unite the niche and neutral theories into a consistent theoretical framework (Adler et al., 2007). It defines coexistence as a balance between niche differences and relative fitness differences with the mechanisms giving rise to these differences named *stabilising* and *equalising* mechanisms respectively. In simple theoretical models, such as the Lotka–Volterra competition model, these quantities can be derived using simple ratios of parameters. The combined theoretical simplicity and conceptual power of this approach to unify our understanding of competition and community assembly (Grainger, Levine, et al., 2019) led to its widespread and accelerating adoption as a framework around which to design experiments investigating the roles of species interactions in community assembly.

The standard approach to empirical MCT comprises a series of steps that are tractable for individual research groups to undertake: using small-scale experiments or natural population dynamics in defined plots to parameterise density-dependent population growth functions. The model's parameters are then used to calculate niche and fitness differences, which predict a binary coexistence outcome based on a simple theoretical criterion. Importantly, experiments can be repeated under different controlled treatments to identify shifts in these key quantities and subsequent coexistence predictions. This general approach has been followed by dozens of studies and is expanding rapidly. Such studies have been conducted across a wide range of taxonomic groups (e.g. Mediterranean annuals, Matías et al., 2018; insects, Terry et al., 2021; yeasts, Grainger, Letten, et al., 2019; and pondweeds, Hess et al., 2022) as well as experimental treatments (e.g. rainfall, Van Dyke et al., 2022; temperature, Armitage & Jones, 2020; consumers, Petry et al., 2018; and shared pollination, Johnson et al., 2022).

At its core, the empirical utility of the MCT framework hinges on its assumed ability to correctly classify, through prediction, the categorical outcome of coexistence or exclusion. Predictions and analysis of coexistence are typically made through satisfaction of the well-known inequality defining the potential for mutual invasibility: $\rho < \kappa_1 / \kappa_2 < \rho^{-1}$ where ρ is niche overlap and κ_1 / κ_2 is the relative

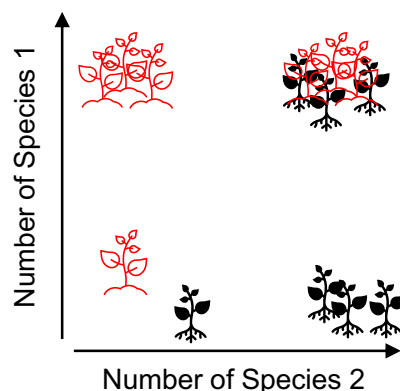
fitness differences between the two competitors. Niche overlap reflects the relative strength of interspecific competition compared to intraspecific competition. Relative fitness differences, on the other hand, reflect the average competitive advantage one species holds over the other. When the mutual invasibility criterion is satisfied, both species are predicted to coexist. These approaches require an assumed functional form of inter- and intraspecific density dependence (such as the Lotka–Volterra or Beverton–Holt competition models) and data from a competitive arena—which can be experimental or observational—between two species (Figure 1). This fitted model is used to estimate each competitor's maximum intrinsic growth rate and competition coefficients. Other taxon- or model-specific values may also be collected, such as seed germination fraction or dormancy rates.

However, empirically characterising competition—even between pairs of species—remains a considerable practical challenge (Hart et al., 2018; Inouye, 2001). Despite frequently following the same fundamental approach (Figure 1), there is a wide divergence in the level of statistical rigour employed across studies (Figure 2), resulting in diminished support for the hypotheses under investigation (Armitage, 2022; Terry, 2023b). In contrast to the apparent simplicity of the results described in the MCT framework, there is no single best data analysis pipeline for empirical applications. Instead, the process of data analysis involves a multitude of decisions that are both data and model dependent, best characterised as a 'garden of forking paths' (Gelman & Loken, 2014). This potential flexibility (termed 'researcher degrees of freedom') encompasses a wide range of choices that must be navigated with care, from selecting among comparable candidate models to determining thresholds for defining 'significant' effects. Particular challenges arise because coexistence inferences are based on quantities which are themselves functions of competition model parameters that are individually hard to empirically estimate and frequently confounded. Furthermore, the core outputs of most studies are essentially unobserved forecasts of coexistence, not observed outcomes, necessitating careful appreciation of the scope of the study and treatment of inherent uncertainty.

Here we present a short checklist of recommended statistical steps (Box 1) and an overview of these recommendations as they relate to making predictions in empirical coexistence studies. While checklists may be perceived as inflexible to authors' preferred experimental and analytical approaches, this is a mischaracterisation of their intended use. A checklist is not a prescription, but rather a prompt for active consideration of one's analytical steps and a

1. Fecundity assays under varying competition and treatments

Fecundity experiment
under different Levels
of inter- and
intraspecific competition



Control

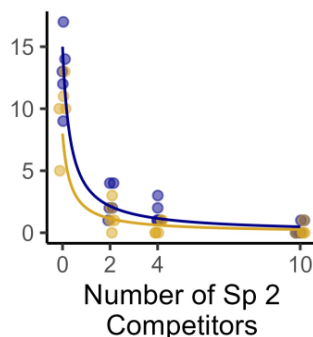
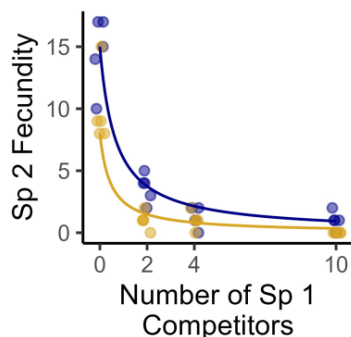
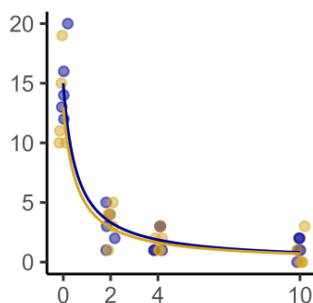
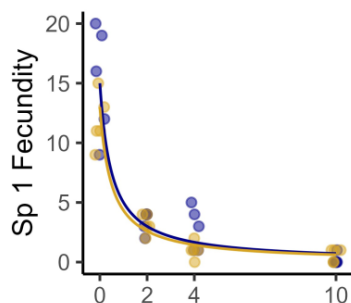


Fertiliser
Treatment

Experimental
Factor

2. Fit population dynamic model to data

Trial ● CTRL ● TREAT



$$\frac{N_{i,t+1}}{N_{i,t}} = \frac{\lambda_i}{1 + \alpha_{ii}N_{i,t} + N_{j,t}\alpha_{ij}}$$

Term	Control	Treatment
λ_1	15	13
λ_2	15	8
α_{11}	2	2
α_{21}	1.5	2.2
α_{12}	1.75	1.75
α_{22}	3	3

3. Determine predicted coexistence determinants

Coexist if: $\rho < \frac{\kappa_i}{\kappa_j} < \frac{1}{\rho}$, where:

$$\text{Relative Fitness Ratio} = \frac{\kappa_i}{\kappa_j} = \frac{\lambda_i - 1}{\lambda_j - 1} \sqrt{\frac{\alpha_{ji}\alpha_{jj}}{\alpha_{ij}\alpha_{ii}}}$$

$$\text{Niche Difference} = 1 - \rho = 1 - \sqrt{\frac{\alpha_{ij}\alpha_{ji}}{\alpha_{ii}\alpha_{jj}}}$$

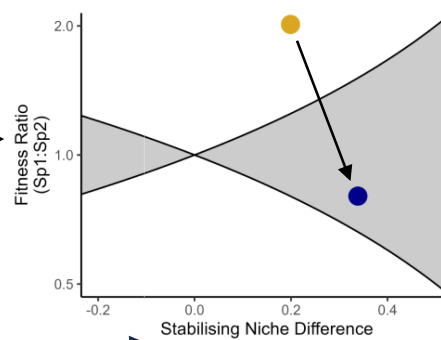


FIGURE 1 Illustration of the three steps common to most empirical modern coexistence theory studies. Here, researchers ask whether a treatment affected the predicted coexistence between a pair of species. In step 1, replicated population growth assays of both species are run under different competitor densities and levels of the treatment. In step 2, a model is fit to these data (here a Beverton–Holt model with intrinsic growth rate λ and competition coefficients α) to estimate the per capita population growth rate per generation. Note that for this idealised example, the data were simulated from this model with Poisson error but the plotted curves are generated using the ‘true’ values of the parameters listed in the table. Lastly, in step 3, predicted coexistence or exclusion is assigned using a theoretically motivated inequality. These identify the conditions for coexistence and label particular quantities as relative fitness differences and stabilising niche differences which can be summarised in the coexistence plane plot. In this case, the treatment is predicted to shift the system from the exclusion of species 2 to coexistence.

reference point for reviewers less familiar with the methods (Parker et al., 2018). While each check is intended to be uncontroversial and consistent with standard protocols for presenting a model's fit to empirical data (Schmolke et al., 2010; Zuur & Ieno, 2016), they are frequently omitted in published papers. We therefore highlight with examples how seemingly innocuous omissions can lead to erroneous or unsupported conclusions specifically within empirical MCT. The experimental work required to parametrise the models is demanding, but this should increase the motivation to make best use of hard-won data. Handling multiple forms of uncertainty simultaneously represents a significant challenge across scientific fields (Milner-Gulland & Shea, 2017; Simmonds et al., 2022), but proven statistical tools exist that can significantly enhance the robustness of results.

All ecological theory requires a set of starting assumptions in order to be operationalised and deployed (Grainger et al., 2021). However, successful deployment of theory in an empirical setting requires these assumptions to be regularly tested. For MCT these assumptions can be split into two categories. First, that competition between the focal species is well described by one's chosen model and the estimated parameters can generate reliable inferences. Second, that the validity of the theoretical framework itself holds in real-world scenarios. These latter assumptions may include the suitability of a pairwise model in multispecies settings (Barabás et al., 2016; Levine et al., 2017), the negligible role of demographic stochasticity in coexistence assessment (Pande et al., 2020; Schreiber et al., 2023) and the lack of interference from confounding environmental factors. Here we focus on this first set of assumptions concerning model suitability as they are directly relevant to ongoing trends in the empirical MCT literature and are often a point of confusion by readers unfamiliar with the MCT approach.

There are a range of alternative approaches to applying MCT in empirical settings (Godwin et al., 2020) and differing interpretations of niche and fitness differences (Song et al., 2019; Spaak et al., 2023). For consistency, our discussion and examples cover the widely used experimental design and interpretations outlined in Figure 1. However, many of the conclusions also apply under alternative definitions of niche and fitness differences (Carroll et al., 2011; Narwani et al., 2013) or approaches that fit models to long-term time-series data (Adler et al., 2006). They are also consequential for other approaches related to MCT such as analytical and simulation-based approaches for estimating fluctuation-dependent coexistence mechanisms (Ellner et al., 2019; Sears & Chesson, 2007), or other approaches to studying coexistence such as structural stability (Saavedra et al., 2017), all of which require the statistical parameterisation of competition models (Box 1).

BOX 1 Summary of key checklist items for robust empirical inferences within the MCT framework

1. Availability and accessibility of underlying data.
 - a. Raw data and computer code are permanently, publicly available in a standard format.
 - b. Raw observations are plotted alongside fitted model predictions.
2. Selecting the core population growth model.
 - a. Basis for selecting the functional form(s) of density-dependent responses is reported.
 - b. Choice of data transformations and error distributions are justified.
 - c. Comparison and selection among alternative candidate model(s) has been carried out in line with study objectives.
3. Addressing uncertainty in model coefficients.
 - a. Uncertainty around parameter estimates is reported.
 - b. Support for the assumption of treatment effects on parameters is presented.
4. Propagation of full uncertainty to final results.
 - a. Parameter uncertainty is propagated into predicted coexistence metrics (niche and fitness differences, invasion growth rates).
 - b. Coexistence predictions are interpreted in light of this uncertainty.

2 | RAW EXPERIMENTAL DATA

2.1 | Check 1a—Availability of raw data accompanied by metadata

The recent shift towards publishing raw observational data underlying experimental results has been a major advance for reproducible science (Culina et al., 2018; Reichman et al., 2011). Publication of raw original data (typically in comma or tab-delimited text format), alongside reasonably annotated code and metadata, has key advantages over just publishing tables of fitted parameter values. Sharing raw data not only enables the complete replication of the original methods but also facilitates the implementation of novel analytical approaches and the testing of

emerging theories (Jenkins et al., 2023). Fortunately, an improvement in raw data availability is a clear trend in our survey of empirical MCT studies (Figure 2), although we encountered recent examples where datasets were described as open, but not linked to from the paper. Relatedly, sharing of computer code, ideally annotated, greatly aids method transparency and reproducibility.

2.2 | Check 1b—Raw data plotted with best fit lines

Despite being a requirement in introductory statistics classes, the graphical presentation of raw data alongside modelled trendlines is surprisingly uncommon (Figure 2). This is unfortunate, as such plots are useful for assessing the accuracy and precision of non-linear population models. The human eye is excellent at pattern detection, and visual assessment aids the more quantitative checks we describe in the following sections. Even where the raw data are archived it should not be expected that the average reader or reviewers generate these plots themselves. While it can be awkward to present raw data from large experiments, in most cases these plots can be moved to [Supporting Information](#) to avoid space restrictions.

3 | MODEL DEPENDENCE OF RESULTS

There are a variety of mathematical functions describing how an individual's reproductive output is affected by various forms of competition. Choosing among many similar, alternative competition models is a longstanding problem in ecology (Ayala et al., 1973; Law & Watkinson, 1987; Martyn et al., 2021), but is a requirement of most approaches for predicting coexistence (Cervantes-Loreto et al., 2023), since all coexistence metrics and their associated uncertainties are contingent on the model itself being a reasonable representation of reality. Each population's monoculture equilibrium and their invasion growth rates when the competitor is at that equilibrium are quantities central to MCT's coexistence metrics. Most often, rather than being directly observed, they are derived from extrapolations of a fitted model. Because of this, relatively small changes in functional form or parameter values can result in qualitatively different predictions about coexistence. It is therefore surprising how few of the surveyed studies explicitly compare or present fit statistics of the models being used to make inferences.

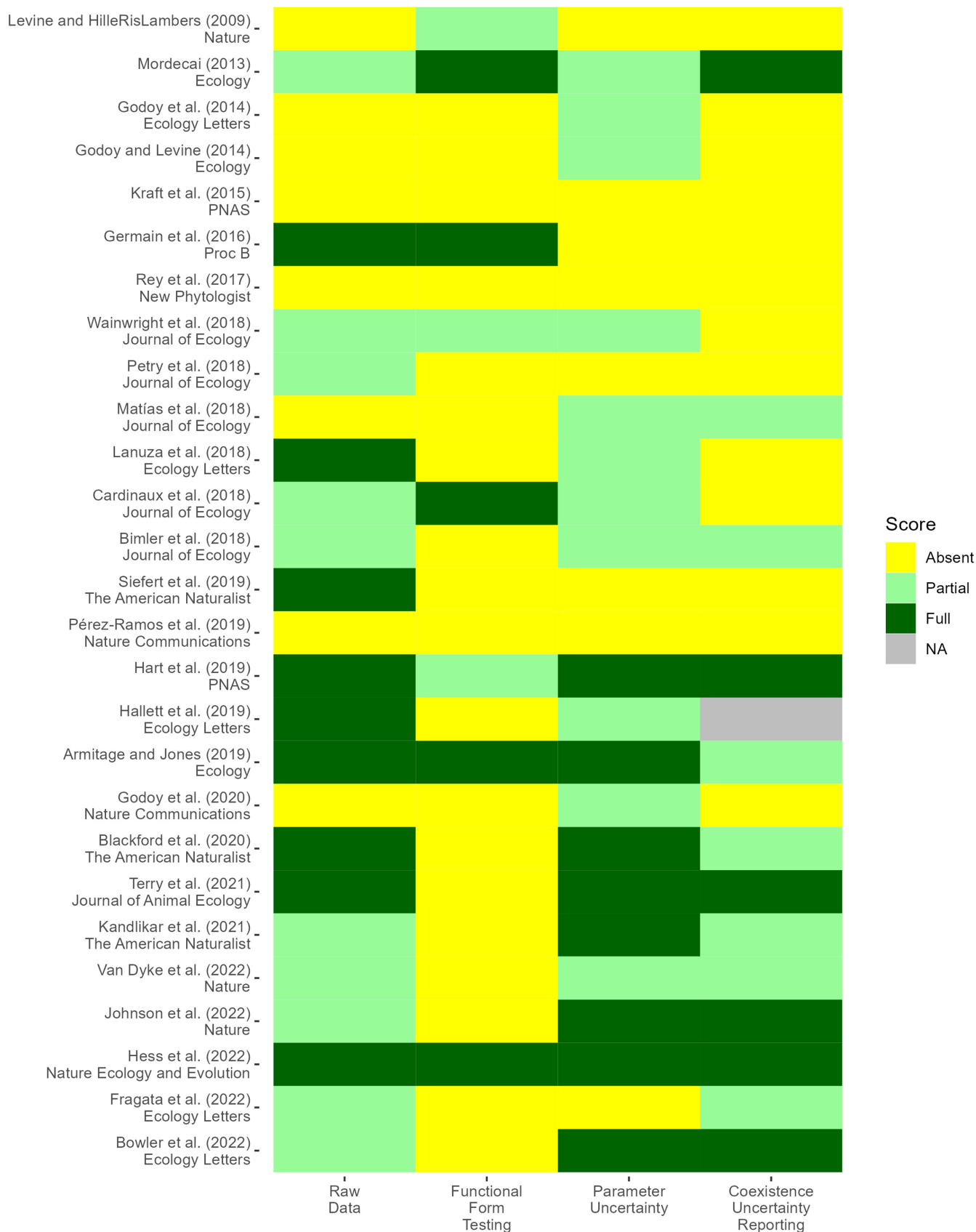
Competition models used in MCT studies tend to be phenomenological (i.e. the myriad mechanisms contributing to pairwise competitive outcomes are collapsed into a single coefficient) and are often chosen based on historical precedent. Because of this, we argue that there is usually no a priori 'best' model for a given system, and even if there is, this choice should still be justified by comparing its fit statistics to plausible alternatives. Model comparison is an active area of statistical research and though it can be challenging to offer universal advice, in general, we suggest the very act of comparing multiple candidate models is a critical step, regardless of the specific metric or approach taken. In this section we will consider how two aspects of the model fit to raw data—the functional form of the response to competition and the error structure can influence results and offer some guidance about the process of model selection in an MCT context.

3.1 | Check 2a—Justification of the competition model's functional form

Across our literature survey, models were frequently described as having been chosen based on simplicity and use in prior studies. In particular, studies of coexistence in annual plant communities nearly unanimously assume the Beverton–Holt model best describes the dynamics of the systems. This simple model assumes a saturating form of density dependence (Figure 1) from which it is convenient to derive niche and fitness differences. Specifically in the context of MCT—which relies on extrapolating fitted models to equilibrium abundances in order to estimate invasion growth rates—reasonable alternative model structures can lead to drastically different inferences of predicted coexistence (Abrams, 2022; Cervantes-Loreto et al., 2023). An example of this using simulated data is shown in Figure 3.

Ideally, competition experiments should be conducted using densities that extend beyond the population carrying capacity (Inouye, 2001). When competition trials are not held at densities at or beyond competitors' carrying capacities, there is an additional imperative to select a competition model that can reasonably be extrapolated to estimate the zero-net-growth conditions at which invasion growth rates are calculated. In such situations, direct model selection may have limited capacity to correctly differentiate models that give very different predicted equilibrium values and it would be necessary to select a model based on precedent and acknowledge the dependence of the results on these choices.

FIGURE 2 Summary of select studies that make use of the classic empirical approach described in Figure 1. A more detailed table, including studies using different approaches such as invasion analysis or fitting to time-series data is included in the accompanying repository. Assessment of whether studies partially or fully meet criteria is necessarily somewhat subjective. The four columns align to the four categories in our checklist: (a) Raw data: Are the raw observed data available and is the model fit plotted with the raw data? (b) Functional form testing: Do the authors test alternative competition models and fully report the model comparison? (c) Parameter uncertainty: Do the authors report the uncertainty in the competition model parameters? Studies reporting just standard errors were graded as 'partial' and full confidence intervals as complete. (d) Coexistence uncertainty reporting: Is the final coexistence prediction reported or tested as a single point estimate, are error bars of some sort reported (partial), or is uncertainty propagated fully to the end result?



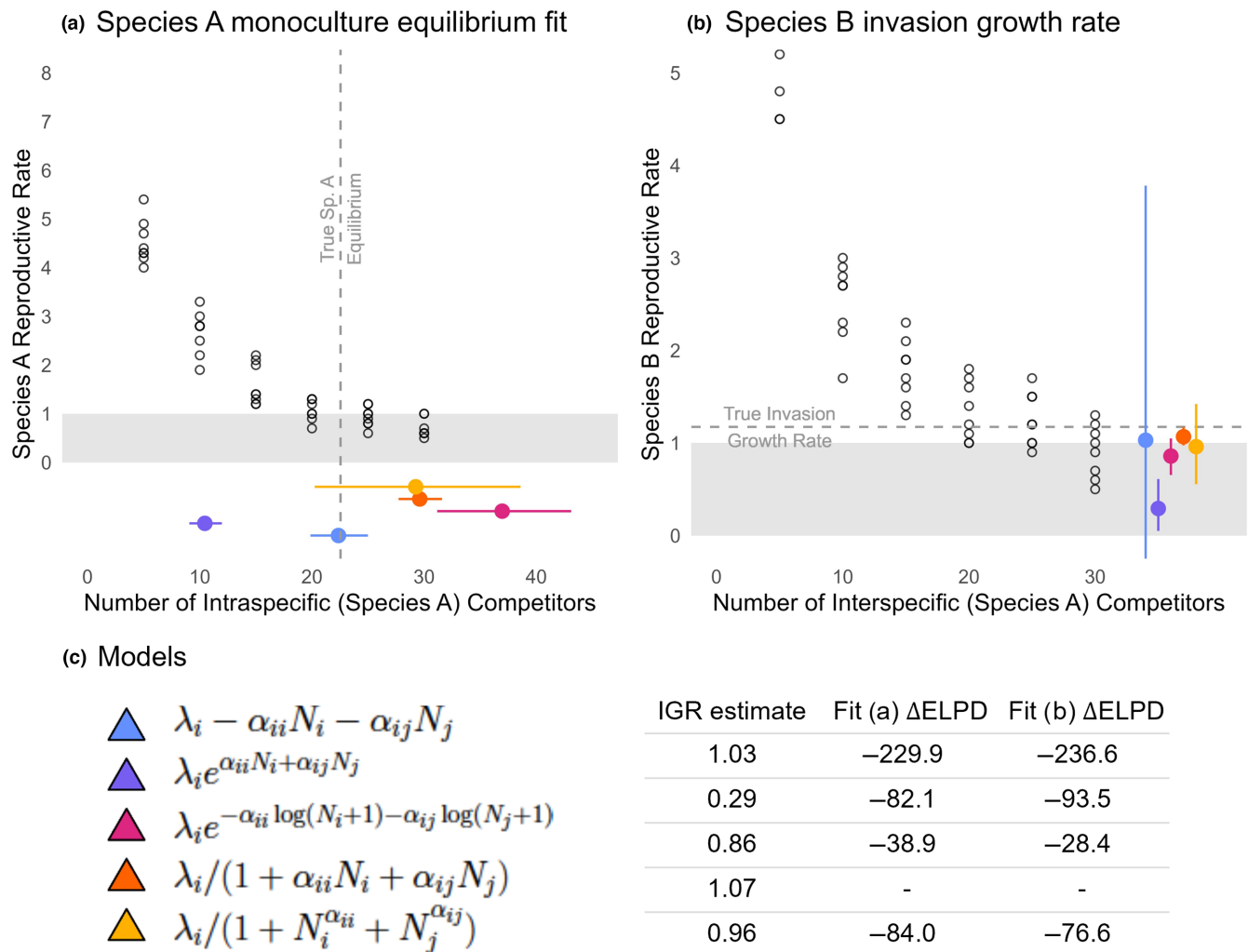


FIGURE 3 Illustration of the influence of model specification on coexistence predictions. Five standard competition models are fit using the *brms* R package to a moderate amount of simulated data. For simplicity, this example investigates only the invasion growth rate (IGR) of species B into a monoculture of species A. Where $IGR < 1$, the species cannot invade, and so is assumed under the framework to be unable to coexist. Here, simulated data are generated using a model where the reduction in seed production is specified with a three-parameter expression $N_{i,t+1} = g \text{Pois} \left[\lambda_i N_{i,t} / (1 + \alpha_{ii} N_i + \alpha_{ij} N_j)^{\theta_i} \right]$, where the germination rate (g) = 0.1, growth rate $\lambda_A = 200$, competition coefficients $\alpha_{AB} = 0.3$, $\alpha_{AA} = 0.4$, shape parameter $\theta_A = 1.3$, $\lambda_B = 300$, $\alpha_{BA} = 0.8$, $\alpha_{BB} = 0.4$, $\theta_B = 1.1$. Poisson noise on the observed seed counts was included. Fitted model trendlines are shown in Figure S2. (a) Estimates of the monoculture equilibrium of species A is dependent on the selected model. (b) Estimates of the fecundity of species B, given species A at a monoculture equilibrium (i.e. its IGR) are also highly model dependent. (c) In this case, the Beverton–Holt model (dark orange) is best supported by ELPD (expected log pointwise predictive density) model comparison for both species and is able to give a qualitatively correct assessment of the IGR. Alternative models, including those that appear to fit well can give qualitatively different predictions.

A key first step in this process is defining the set of models to be compared. Since most competition models are phenomenological, there is essentially no limit to the forms that could be tested, and there are strong arguments that more mechanistic generative models should be developed for specific well-studied groups of organisms (e.g. annual plants, Stouffer, 2022). However, when phenomenological models are all that is available, they tend to fall within the Volterra–Lotka–Gause family of models (e.g. those listed by Law & Watkinson, 1987), which enjoy reasonably straightforward derivations of niche and fitness differences from their shared parameters. We suggest that the diversity of functional forms is more important

than the total number of candidate models. However, identification of alternative candidate models should consider the role of flexibility in the shape of density-dependent responses—see Novak and Stouffer (2021a) for a discussion of these issues in the similar context of fitting consumer functional response models. Further comparison of competition models with null, competition-free or single-coefficient competition models can also provide helpful context on the appropriateness and fit of the corresponding fully parameterised model. When the raw data follow a different shape than that stipulated by classic candidate models, it may be necessary to develop new derivations that better match reality (Godoy & Levine, 2014).

3.2 | Check 2b—Justification of error structure and transformations

Experimental or observational data always contain error components that fitted competition models can only imperfectly capture—the exact same combinations of competitors often result in a wide range of observed fecundities among identical replicates. This error is frequently modelled as observation error, though in many cases, the measurements (e.g. counts of seeds) are likely to be relatively accurate. Instead, inter-replicate variation frequently results from process error in the sense that is caused by real ecological mechanisms such as intraspecific variation, microclimatic differences and stochastic events which differentially affect experimental replicates. Furthermore, it is common for the focal species in competition experiments to have different reproductive strategies and hence different variability between species. Certain plants, for example, produce extremely variable numbers of seeds per individual and within a species the distribution of seed production can also vary from normally distributed to highly skewed or zero-inflated. This variation (and how it is handled) has consequences both for observed population dynamics (Hart et al., 2016; Pascual & Kareiva, 1996) and the fitting and assessment of competition models.

In this context, careful consideration needs to be given to data transformations, error distributions and whether the variance is treated separately for each species. The core problem is that the mean of many transformations of a random variable is not generally equal to the transformation applied to the mean—that is, Jensen's inequality (Ruel & Ayres, 1999). There has been much discussion of this issue in other ecological contexts (Richards, 2008; Warton et al., 2016), yet we found it barely mentioned in our survey of the MCT literature. Because the goal of the model fitting stage is to precisely and accurately estimate parameter values, rather than to simply assess their statistical significance, generic advice and assumptions concerning transformations from a direct null-hypothesis testing context may be less relevant.

Until somewhat recently, least squares and maximum likelihood approaches to parameter estimation were necessitated by computational constraints. In the specific context of nonlinear competition models and noisy measurements, these methods can be unstable and result in convergence errors unless transformations are first applied to the data. In particular, log-transforming both sides of a predictive model and assuming a Gaussian error distribution during fitting has been a common approach in plant competition models (Law & Watkinson, 1987) to normalise the distribution of variances and make nonlinear models easier to fit (Carroll & Ruppert, 1984). This can improve model likelihoods when data are sparse, but introduces biases for MCT beyond the widely appreciated challenges of zero-values and non-compliant forms of heteroscedasticity. The most meaningful 'average' value of seed count from a single generation is an arithmetic (natural scale) mean, not the geometric (log-scale) mean that is estimated from a fit to the logarithm of seed counts. As such, fitting to $\log(\text{seeds})$ will underestimate the

expected mean number of seeds per individual of that species by a factor proportional to $\exp(\sigma^2)$ where σ is the standard deviation on the log-scale.

Particularly where the variance differs between species, these biases in estimates of maximum growth rates (λ) and competition coefficients (α) can have significant consequences for predicted coexistence outcomes. In the simulated example presented in Figure 4 where data are lognormally distributed, the parameter values estimated from the logged data markedly underestimated the population growth rate terms and overestimated the competition terms, particularly of species 1, which had the higher error. This seemingly minor decision to transform the data has the consequence of completely switching the predicted competitive hierarchy from the 'true' outcome. For an example using real data, see Figure S1. In that example, both a Poisson and a Negative Binomial error distribution were able to give good predictions, but the Poisson model was overly confident in its parameter estimates.

The development of efficient Monte Carlo approaches for model fitting such as STAN (Stan Development Team, 2022) and its accessible R frontend *brms* (Bürkner, 2018) mean there is more flexibility in directly modelling error and avoiding the need for transformation. Error models can account for the functional form of uncertainty such as the commonly observed heteroscedasticity in offspring production across gradients of competitor densities. Error models can also be fully pooled (i.e. a single error term fit) or separated for each focal species and treatment level. For counts, a negative binomial or a quasi-Poisson error structure can capture more complex and non-symmetric error patterns, while estimating mean effects on their natural scales.

That said, it is rarely easy to choose an error model a priori solely on theoretical and biological grounds. Visual inspection of model fit, particularly by plotting the distribution of residual errors, can often be helpful to determine candidate models for residual variation. If comparison of different functional forms is undertaken, it is little additional work to include alternative error distributions for each candidate model. Care must be taken to compare the data on the same response scales, since many statistical programmes automatically use non-identity link functions when certain families of error distributions are specified. In these cases, each model's parameters must be back-transformed before coexistence metrics can be calculated.

3.3 | Check 2c—Model comparison and selection

It is well established, although often taken for granted, that the goals and outcomes of model selection depend on the objectives of the study (Aho et al., 2014; Tredennick et al., 2021). This is mostly widely understood in terms of the classic metrics: Akaike's information criterion (AIC) is generally preferred for its capacity to identify a model with the best predictive power, and the Bayesian information criterion (BIC), which penalises free parameters more than AIC, is favoured for accurately identifying the 'true' model. Although the

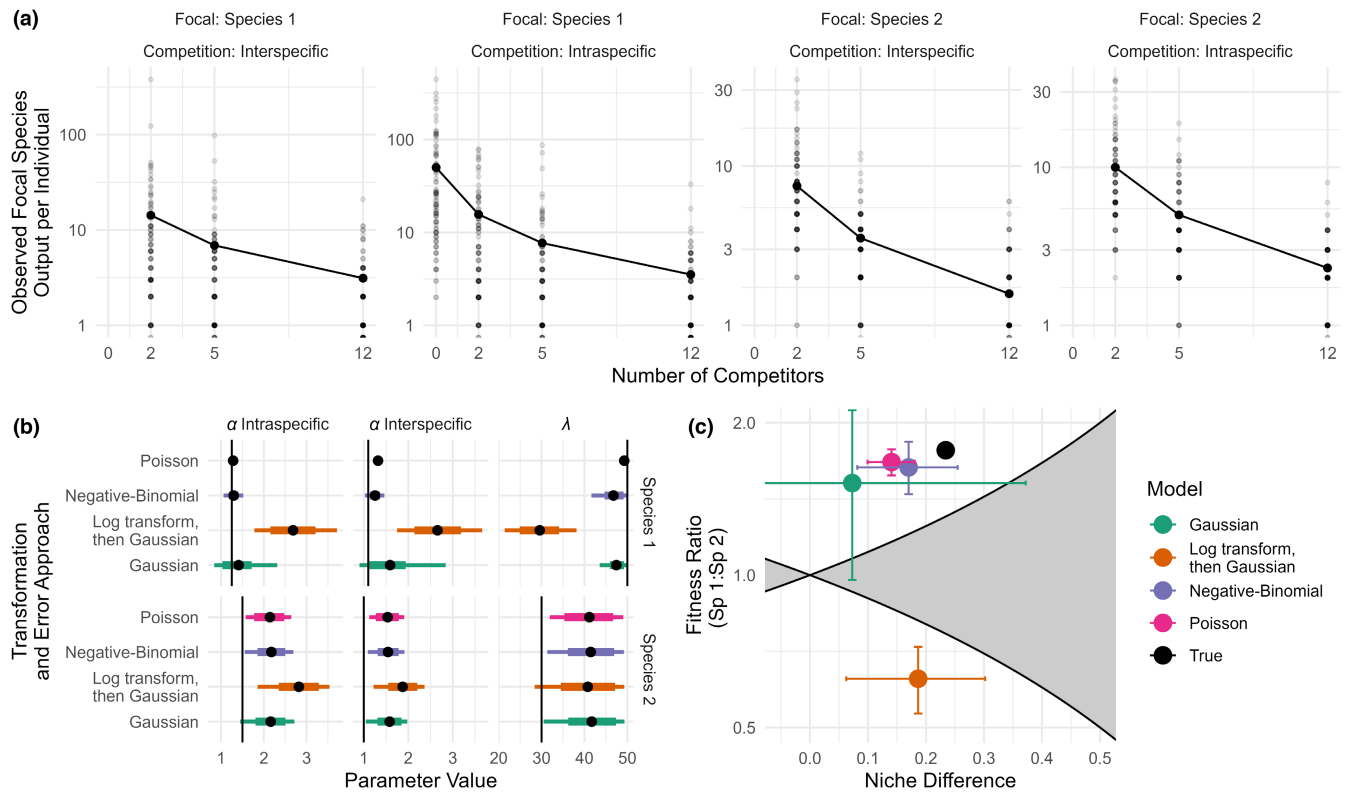


FIGURE 4 The impact of data transformation and error structure on coexistence predictions. Here observation data of two species (1 and 2) are simulated from a Beverton–Holt model (parameters $\lambda_1=50, \lambda_2=30, \alpha_{11}=1.1, \alpha_{22}=1, \alpha_{21}=1.5, \alpha_{12}=1.25$). (a) Observations (100 per level of competitors) were drawn from a discretised log-normal distribution ($\log \sigma_{sp1}=1.3, \log \sigma_{sp2}=0.6$) with an expected mean value equal to that predicted by the competition model (black lines). (b) Four different error approaches were used to fit the ‘correct’ Beverton–Holt competition model using the *brms* R package to these data: (i) directly fitting a Gaussian error model to raw observations, (ii) log-transforming ($\log(x+1)$) both sides of the competition model and then assuming Gaussian errors, (iii) fitting a Poisson error model to raw observations and (iv) fitting a negative-binomial error model to raw observations (fitting separate shape parameters for each species). Parameter estimates are shown with means, central 90% and 60% intervals. The ‘true’ values used to generate the data are shown with the vertical black line. (c) Coexistence plane diagram showing predictions of the four models, and the ‘true’ location. Error bars show central 90% intervals. Fits of all models to the raw data are shown in Figure S3.

implementation differs in a Bayesian framework, for example using LOOIC (leave-one-out-information criteria, Vehtari et al., 2017), ELPD (expected log pointwise predictive density, Vehtari et al., 2017) or WAIC/WBIC (Watanabe–Akaike/Bayesian information criterion, Watanabe, 2013), the same fundamental considerations about how to weigh prediction versus inference remain. Once model comparison has been carried out, we recommend authors include a table reporting the range of models examined, the criteria used to identify the best supported models and (where there is not a clear model) robustness of the conclusions to the use of reasonable alternative models.

For a typical experimental coexistence study (Figure 1), the overall predictive ability of the underlying competition model is likely to be less relevant for most questions than accurately describing the functional form of competition, which could imply a BIC-family measure would be preferred. Within a classical maximum likelihood framework, BIC is as easy to implement as AIC. However, at the time of writing for Bayesian models there is considerably better support for model selection on the basis of

predictive performance (such as WAIC and ELPD), in particular through the ‘loo’ R package (Vehtari et al., 2022) for STAN models. A more direct approach to model fitting simply uses cross-validation either among experimental replicates or generations to identify the best competition model. However, because most coexistence experiments suffer from low replication and are rarely conducted over more than one generation, cross-validation is rarely performed. Despite this abundance of somewhat arbitrary methods to choose from, engaging closely with and reporting the process of model selection is more important than the specific methods used.

Where unique, non-nested models are equally supported, repeating the full analysis using multiple model structures can be informative. If the end conclusions are consistent, then this is strong additional support for the conclusions (e.g. Hess et al., 2022). In contrast, if the conclusions are contingent on a particular model, when alternative functions also fit the data well (Armitage, 2022), then additional data (or justification) may be required to support one over the others. When dealing with nested models, candidate models

sharing equivalent parameters can be used on a species-specific basis, since their derived coexistence metrics have the same formulae. More flexible forms of niche and fitness difference metrics also exist (Spaak & De Laender, 2020) that can directly compare invasion growth rates estimated from non-nested models, sidestepping the need for the same model or even the same model family to be assumed for all species.

Formulaic use of model selection criteria is not without its own risks. Model selection algorithms cannot replace the visual assessment of different models across and beyond the range of data. In particular, inspection of the pattern of residuals across the data is helpful to identify cases where predictions could be dependent on the sampling design of the raw data. Where simple models are favoured by information criteria, consideration should be given to the likelihood that this is a symptom of noisy data, rather than the generality of parsimonious models. In particular, inference based on parameter estimation after model selection can be a problem (Yates et al., 2023) as estimates of significance can be positively biased. Model selection issues have been identified as causing systematic bias in the analysis of predation functional response models which share many of the same fundamental difficulties as competition models (Novak & Stouffer, 2021b).

Whichever model selection path is followed, for clarity and to catch errors, it is helpful to plot the model fits and explicitly report both the fitted log-likelihood (or its equivalent) and number of fitted parameters for each model. More heavily parameterised models are prone to identifiability issues and can struggle to converge with noisy data (Hess et al., 2022; Levine & HilleRisLambers, 2009). For truly nested models, increased model complexity should only increase the log-likelihood. Hence for each parameter removed, it would be expected that the maximum improvement in AIC is 2. Results that do not follow this pattern strongly suggest that the model fitting has not been successful.

A key complement to the statistical approaches described above is the transferability of fitted models to new, independent data. Although the stated goals of most coexistence studies are to understand, rather than predict, nearly every coexistence study bases its conclusions on predictions. But how well do the common suite of models actually predict real competitive dynamics? It is extremely rare for coexistence studies of the classic type described in Figure 1 to challenge their fitted models to predict growth rates and coexistence in a subsequent growing seasons, let alone in a different geographic location or under non-experimental conditions (although see Adler et al., 2013). Rather, where it has been conducted, external validation has primarily relied on mapping qualitative coexistence predictions onto either georeferenced co-occurrence records (Armitage & Jones, 2020; Narwani et al., 2013) or observations from paired short-term experiments (Godoy & Levine, 2014). Although a significant challenge, quantitative independent validation of model predictions is a critical missing link that could help to resolve the common prediction that species pairs which appear to stably co-occur in

nature will deterministically exclude one another under a parameterised competition model.

4 | PARAMETER UNCERTAINTY

Point estimates of the parameters of competition models are never exact owing to the statistical uncertainty derived from measurement error and real intraspecific variation due to differences in micro-environment, genetic, epigenetic and parental effects. This intraspecific variation can in itself have significant consequences for species coexistence (Hart et al., 2016; Stump et al., 2022). Theoretical studies have investigated how perturbations to key parameters can influence coexistence (Barabás et al., 2014). However, these sensitivity analyses are applicable to mathematically small perturbations—that is, those that do not shift populations into alternative dynamical regimes such as from persistence to extinction.

Beyond the mechanistic effects of variation in growth parameters, the model fitting process itself returns variation around point estimates of model parameters. Appraisal of this variation is a critical but often-missing aspect of many coexistence experiments. In practice, the number of competitive trials per free parameter is frequently relatively moderate, typically ranging from 10 to 20 across studies (although sample sizes were not always reported). Nonetheless, some degree of residual uncertainty is expected given the limitations of time- and labour-intensive experiments—for many ecological questions it is often advantageous to test more species pairs reasonably well rather than exhaustively characterise competition between a single pair. The question is how to appropriately evaluate and work with this residual uncertainty to generate reliable and actionable predictions of coexistence.

4.1 | Check 3a—Report uncertainty in model coefficients

Our literature review revealed that a majority of studies included some reporting of the uncertainty in the estimates of key competition model parameters—most frequently as standard errors. This is helpful to allow both readers and reviewers to assess the confidence affordable to the results, but there is still room for improvement. The likelihood profiles of competition (α) terms are often non-symmetric, and are poorly summarised by a single value. Where feasible, it is advantageous to either provide 95% intervals or ideally plots of the posteriors (Johnson et al., 2022). However, an important caveat is that difficulties separating the parameters in competition models causes the uncertainty associated with a particular parameter to be dependent on others. In particular, λ_i and α_{ij} estimates are frequently correlated, with consequences for MCT predictions that are not easily summarised in a list of parameter values (see next section). Furthermore, standard confidence intervals for parameters implicitly rely on the underlying model's suitability, and so can give artificially high confidence when alternative models have not been

compared. For instance, in the example in Figure 4, the Poisson error model gave very precise but inaccurate estimates for the competition coefficients.

4.2 | Check 3b—Report support for effect of treatments on model coefficients

An objective of many MCT studies is to identify the effect of an experimental treatment on coexistence through its actions on the metrics of niche and fitness differences. In most cases, these derived metrics are calculated from ratios of demographic parameters fit from the competition models. Some go further, for example Van Dyke et al. (2022) compare the effect of a water treatment on the two component parts of the relative fitness differences: the demographic ratio $\left(\frac{\lambda_i - 1}{\lambda_j - 1}\right)$ and the competitive response ratio $\sqrt{\frac{\alpha_{ji}\alpha_{ij}}{\alpha_{ij}\alpha_{ji}}}$. However, it is important to assess whether the experimental data statistically support the assumption of treatment effects on parameters in the first place, to avoid conclusions being unduly influenced by noisy data.

A key challenge is that interaction coefficients between species (α terms) are more difficult to accurately estimate than intrinsic reproductive rates (λ). This is usually caused by fewer numbers of data points being available to estimate the competition parameters. Experimenters can easily include relatively large numbers of isolated individuals to directly assay λ terms whereas α terms rely on multispecies experiments across a gradient of relative frequencies. In designs with more than two species, every trial with the focal species can contribute to the estimation of λ , while only those between specific pairs can assess α terms. Consequently, the influence of individual, noisy data points are more likely to influence α terms compared to λ terms, particularly when separate parameters are being fit for each experimental treatment (Terry, 2023b). In light of this, how should a practitioner assess whether the assumption of treatment differences in model parameters is warranted?

One approach is to compare models that fit separate parameters between treatments with models where the parameter values are shared between treatments. However, in larger experiments with more than a single pair of species, the appropriate scale at which to test treatment effects can be hard to define, particularly for the competition coefficients. For instance, it is not always clear if assessment of treatment effects should be conducted for each α separately (as per Johnson et al., 2022), for all α 's belonging to each focal species (Terry, 2023b), or over all terms (both α 's and λ 's) simultaneously (Terry et al., 2021)? Comparing models in this manner can highlight cases where individual species experience treatment-driven shifts in growth or competition, but practitioners must be cautious of the increasing false positive risk with increased numbers of species.

Ultimately, the approach taken and choices made concerning whether to accept or reject specific treatment effects hinges on a pluralistic assessment of statistical evidence for or against these effects combined with expert system-based judgement. If, for example, treatment effects show no increase in a model's predictive

performance but large confidence margins, then one might be better served by using all of the available data to fit a single competition parameter with higher confidence. Alternatively, simulation analysis can identify the sensitivity of the analytic pipeline and end results to noise in the empirical data. Where appropriate, simulated data that do not include a treatment effect can be used as a null model against which to test the effect of hypothesised shifts (Terry, 2023b). This approach is particularly useful in translating uncertainty across levels from population growth model parameters through to final results, where direct analytical solutions to uncertainty propagation are challenging.

5 | HANDLING AND PROPAGATION OF UNCERTAINTY

Errors around parameter estimates do not disappear after being used to calculate coexistence metrics. Rather, propagation of this error through to the final coexistence predictions is critical for a fair assessment of the evidence in favour or against coexistence. While error around individual model parameters is frequently reported, only a small number of papers fully propagate this uncertainty into their coexistence predictions. Rather, many studies simply take the mean or median of this probability cloud and interpret it as an error-free observed outcome. Handling model uncertainty is a significant challenge across scientific fields (Simmonds et al., 2022). In an MCT context, error propagation into final coexistence predictions has been used to inform the results (Cervantes-Loreto et al., 2023; Hart et al., 2019; Hess et al., 2022; Mordecai, 2013; Terry et al., 2021). Bowler et al. (2022) demonstrates particularly well how a failure to propagate uncertainty can result in misleading conclusions about coexistence.

When fitting competition models, the parameters are often not fully separable (i.e. they are non-identifiable), resulting in highly correlated estimates of growth and competition parameters. Since most of the key quantities of MCT are ratios of parameters, these covariances can influence the overall error. While propagating error is best achieved by using Bayesian posterior draws of the parameters to fit distributions of coexistence metrics, it can also be done in a frequentist framework through methods such as moment approximations via Taylor expansions or Monte Carlo simulation (Dietze, 2017). In most cases, the additional work is relatively straightforward. For example, when using the R language, propagating error is often as simple as plugging in a randomly sampled set of posterior draws for each parameter in place of a single value. In the frequentist case, when only a point estimate and associated standard error is returned, error is most easily propagated into niche and fitness difference equations using packages such as *propagate* (Spiess, 2018). The principal challenge is often in reporting the results in concise and readable format.

As a further incentive, propagating multivariate uncertainty when parameters covary can often tighten the confidence in the final predictions. Where variables are divided, positive covariance decreases overall variance. Hence, the most likely covariance arising

between parameters in competition experiments—between λ_i and α_{ij} or α_{ji} —will tend to decrease uncertainty in the relative fitness ratio comprising the fitness differences coexistence metric.

5.1 | Check 4a—Propagation of uncertainty to coexistence plane graphs

The central results of many coexistence studies are presented on cartesian coordinates depicting the predicted outcome of competition between a pair of species in terms of the niche and fitness differences. Here, zones of 'coexistence' and 'exclusion' are demarcated based on the coexistence inequality discussed in the introduction. Plotting outcomes on this plane provides an accessible visual summary of the predicted consequences of experimental treatments on coexistence. However, there are currently no conventions for the appropriate visualisation and interpretation of error on such graphs.

Figure 5 illustrates the range of approaches that have been used to present results on a coexistence plane. While two orthogonal error bars around niche and fitness metrics (Figure 5b) are preferable to nothing, they have multiple deficiencies hindering their interpretation. First, rather than being circular, the true distribution of uncertainty is typically highly irregularly shaped. This arises from covariances between underlying parameters as well as the fact that the α terms contribute to both coexistence metrics, making the two axes non-independent. Second, interpretation of this 2D error requires evaluating the extent to which uncertainty is distributed across a nonlinear boundary defining the predicted outcomes, which can be difficult to assess with only two orthogonal error bars (Figure 5b). Full posterior distributions of coexistence metrics display the most information on uncertainty (Figure 5d), but can be overwhelming when many species pairs or treatments are compared on the same plane (Figure 5c). Summaries of the relative support of each coexistence outcome for each result can be presented in tables or as stacked barplots, depending on the numbers needed. A further option is to use small pie charts (Yan et al., 2022) summarising the relative number of outcomes of coexistence, exclusion or priority effects predicted by the posterior draws (Figure 5e). Importantly, in most cases these plots can only be expected to convey a visual summary of the result—specific statistical tests and summary statistics are also needed. In many cases, there are considerable advantages and very little cost to including the posterior draws as a supplementary .csv file so that future users of the data do not need to refit the models.

5.2 | Check 4b—Propagation of uncertainty to final overall results

We note with concern that highly uncertain niche and fitness differences have often been summarised as single, qualitative binary outcomes, even after the large uncertainties in their underlying parameters have been reported. Most papers presenting uncertainty do so with orthogonal error bars, but do not explicitly mention or

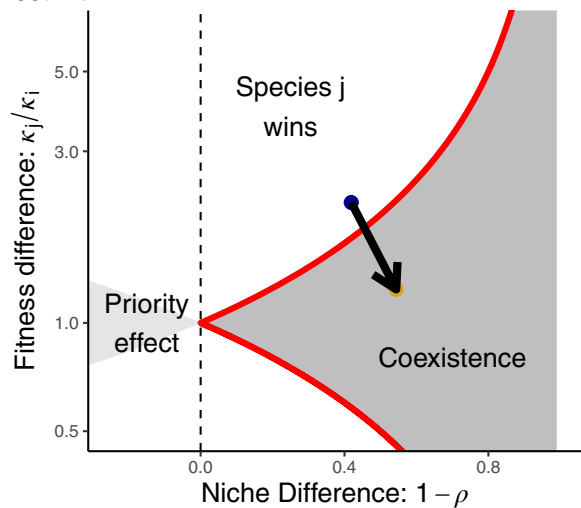
interpret this uncertainty in their reported conclusions. In most cases this uncertainty is a formal expression of the uncertainty in the model, rather than a prediction for the frequency by which any given coexistence treatment or species pairing is likely to result in a particular categorical outcome. However, these interpretations are related—a pair of species with high intrinsic variability in pairwise outcomes will also be difficult to place in a binary coexist or exclude category.

Assessing the evidence for or against coexistence in an error-conscious manner allows the researcher, reviewers and readers to fairly assess the degree of evidence for or against these qualitative, categorical predictions. Bowler et al. (2022) provide a clear application of this approach, wherein joint posterior distributions that have nontrivial densities on either side of the coexistence boundary cannot—under common definitions of statistical evidence—be reliably assigned to either category and must therefore be scored as an ambiguous outcome. While such a conclusion clearly detracts from the apparent cleanliness of the results, we argue that as coexistence outcomes are being increasingly discussed in applied contexts such as restoration (Aoyama et al., 2022) and invasion biology (Epstein et al., 2019), presenting results without corresponding uncertainty is of limited value to managers or other decision-makers. Equally, approaches to quantifying coexistence mechanisms based on partitioning invasion growth rates (Ellner et al., 2019) can also include propagation of uncertainty in parameter estimates (Aoyama et al., 2022).

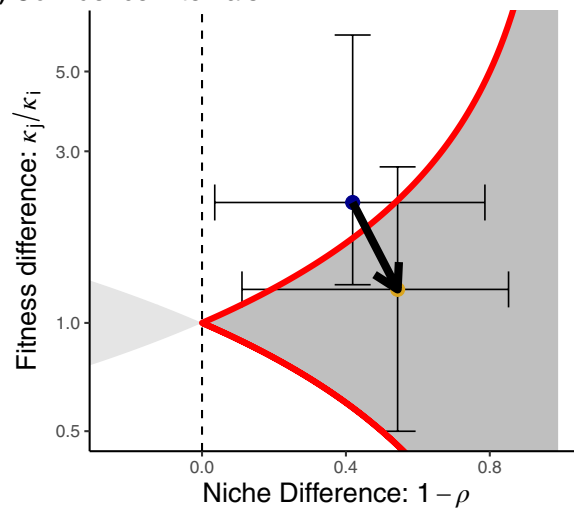
Beyond the issue of assigning coexistence outcomes in a single species pair or treatment, summarising the dynamic effect of experimental treatments on coexistence outcomes is challenging. While it is clear that illustrating treatment effects using only single summary points for each treatment can over-exaggerate these effects (Armitage, 2022; Terry, 2023b), there is no statistical convention for assigning significance to treatment effects on predicted coexistence outcomes even when uncertainty has been quantified. To this end, rather than suggesting a hard-and-fast rule for identifying treatment-driven changes to coexistence (for instance a <5% overlap in the two treatments' joint posterior densities), we advocate practitioners to transparently justify their chosen reasoning for or against such shifts. Options to support such statements include Bayes factors, which quantify relative support for one hypothesis over another, given the data, without arbitrary significance thresholds. In the end, a researcher's decision to weight effect sizes versus the statistical significance of their treatment effects depends on the goal of the study. In cases where errors in predicting coexistence can be costly, such as in invasive species management (Ocampo-Ariza et al., 2018) or in potential future medical settings (Huntly et al., 2021; Letten et al., 2021), effect sizes and probability statements are integral to subsequent intervention decisions, but we recommend they always be presented alongside *p*-values to improve transparency regardless of a study's goal.

While the presentation of probabilistic evidence can require large, ungainly tables (e.g. Terry et al., 2021), a study's main

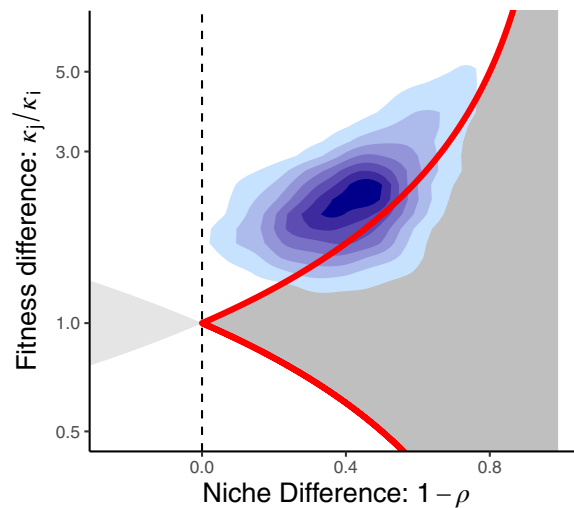
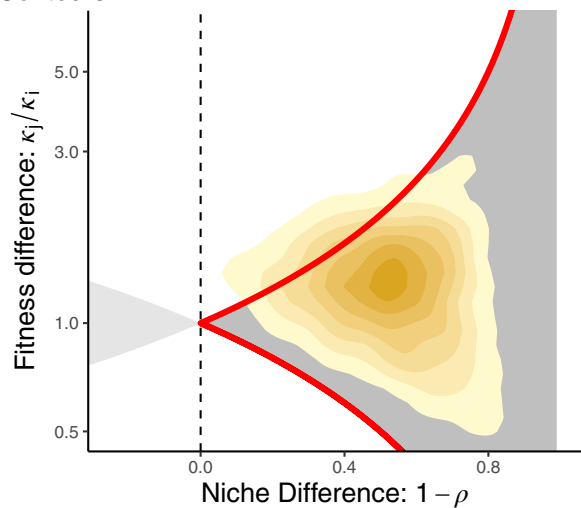
(a) Best Fit



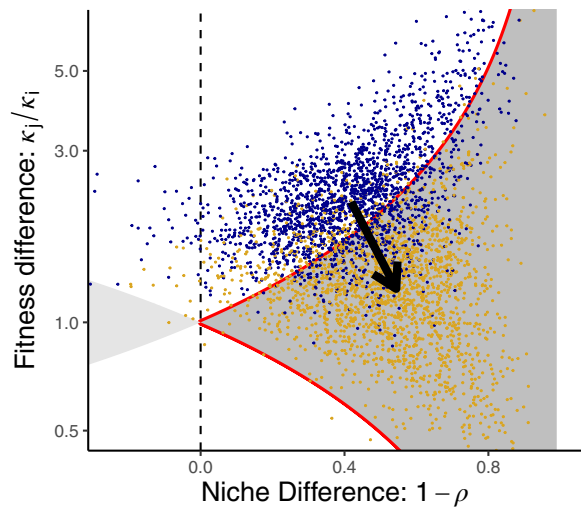
(b) Confidence Intervals



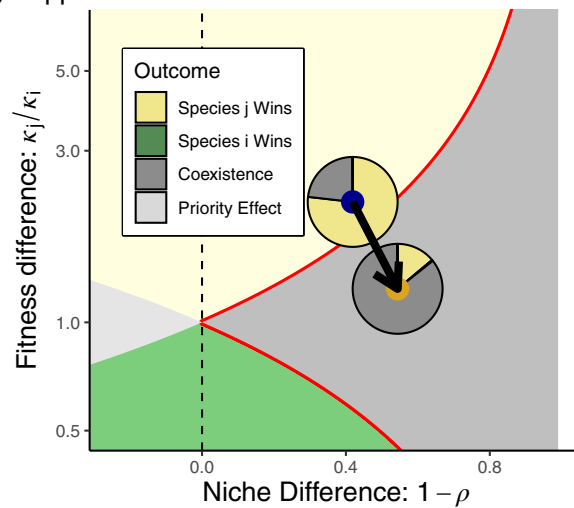
(c) Contours



(d) Posterior Draws



(e) Support Summaries



Trial: ● Treatment ● Control

FIGURE 5 Options for representing uncertainty on coexistence diagrams where the possible potential predicted outcomes are coexistence (both species can invade each other), the exclusion of either species by the other, or a situation where both species can exclude the other (referred to as a priority effect case). (a) Plotting point estimates do not permit assessment of the statistical support for or against coexistence predictions. (b) Orthogonal error bars (here 95% intervals) indicate a high uncertainty in predicted outcome but do not account for the underlying covariance between niche overlap and fitness differences. (c) Graphical summaries of the posterior distribution, such as contour plots, can be more digestible than raw posterior draws but remain hard to quantitatively interpret and usually require faceting of subplots, increasing the space required. (d) Plotting posterior draws directly display uncertainty while permitting comparisons between treatments but are less statistically interpretable than other options. (e) Support summaries offer one way to both maintain the visual simplicity of changes in best fit while including the fraction of posterior support for each of the four potential predictions.

error-conscious conclusions can still be concise. For example, rather than stating that a treatment 'qualitatively altered predicted competitive outcomes for x of the n species pairs' based on the relative position of pairs of mean or median points, uncertainty can be directly incorporated into the conclusions with a rephrasing such as 'within our community, our results suggest for a random pair of species a $p\%$ probability that coexistence outcomes differed among treatments', or an alternative that most closely aligns with the study's objectives.

In non-treatment designs, continuous predictors such as trait or phylogenetic differences have been used to predict niche or fitness differences (and hence coexistence, e.g. Kraft et al., 2015; Pérez-Ramos et al., 2019). However, use of coexistence metrics as predictor or response variables in the standard regression framework ignores their underlying uncertainty. Instead, modern regression techniques such as error-in-variables models (Carroll et al., 2006) widely used in meta-analyses can incorporate previously quantified uncertainty and more accurately identify statistical trends.

6 | PRE-STUDY CONSIDERATIONS

While our proposed statistical checks all take place after experimental data have been gathered, we also recommend some steps be taken when planning experiments in order to avoid some of the aforementioned pitfalls. As a gold-standard, we recommend pre-registration of coexistence experiments. Pre-registration entails drafting and archiving a report outlining the hypothesis, experimental plan, sample sizes and analytical roadmap of one's planned experiment (Parker et al., 2019). This process offers guardrails against selective reporting that is increasingly recognised as a problem in ecology (Kimmel et al., 2023). Such practices are becoming commonplace in many fields, with high-profile journals now offering provisional publishing agreements following adherence to peer-reviewed registered reports (Gya et al., 2023; 'Rolling out registered reports,' 2023). At minimum, we advocate for practitioners to present more well-defined a priori hypotheses concerning treatment-mediated effects on coexistence. Rather than hypothesising that a treatment will qualitatively affect coexistence in some unspecified direction, a stronger hypothesis would predict the direction of the effect and outline a clear accept/reject criterion for how the effect is to be measured and supported.

A second design consideration is sample size since it influences the accuracy and precision of model parameters. Logistical and financial resources can constrain sample sizes, which risk exaggerating responses via type I errors, especially when combined with longstanding publication biases in favour of positive results (Yang et al., 2022). As the scope of most coexistence studies are multi-species assemblages, there is a trade-off between the generality of the results and precisely estimating individual pairwise interactions. An assessment of the statistical power required to detect effects is a recommended approach to informing sample size requirements. To do so, one must already have an approximate guess of the target parameter values, in particular for the expectation of variability and effect sizes for treatments. When calculating power is not feasible, we suggest a minimum of at least 10–40 individual data points per parameter as a starting point for fitting competition models. For communities of moderate diversity (~5–10 species) and two experimental treatments, this can require thousands of replicates to parameterise all pairwise interactions (Terry et al., 2021; Van Dyke et al., 2022). Efforts to reduce the dimensionality of the problem such that the number of experiments required no longer increases with the square of the number of species will be central to scaling-up empirical coexistence studies to the natural diversity of a community (Stouffer et al., 2021; Weiss-Lehman et al., 2022).

A final design consideration is the study's spatiotemporal extent. In our survey of the literature, more than half of all experiments were conducted in a single generation at one location. This excludes the quantification of the hallmark fluctuation-dependent coexistence mechanisms of the original theory which can fundamentally change the way environmental treatments are predicted to act on coexistence (Armitage & Jones, 2020; Letten et al., 2018). As discussed earlier, single-generation experiments also preclude the independent validation of model predictions on new data, limiting the scope of a particular study to a specific time and place. While acknowledging the logistical difficulties and time commitments in replicating these experiments the benefits of spatiotemporal replication are substantial and we urge researchers to conduct their competition experiments over more than one generation if at all possible.

7 | CONCLUSIONS

Based on our literature survey, it appears that a significant number of published empirical MCT papers possess noticeable deficiencies in the statistical treatment of their data, which weakens the

strength of their conclusions. Additionally, many foundational coexistence studies do not provide access to their raw data, hindering their re-analysis. Unfortunately, our case studies reveal that many seemingly innocent omissions, such as insufficient model comparisons or justifications, can relatively easily influence key ecological conclusions.

A key next frontier in the development of coexistence theory will be to improve the empirical validation of predictions. To date, there have been few attempts to directly test how well coexistence predictions made from parameterised models perform in natural settings or even replicated experiments. While the nature of predictions and forecasts will mean direct validation will not be possible in all (or indeed most) scenarios, even qualitatively comparing model predictions with what data are available can enhance the confidence in the approach, for example equilibrium abundances (Hart et al., 2019) or geographic distribution (Armitage & Jones, 2020).

A recognised strength of MCT is the framework's generalisability across a wide range of taxa. In principle, a similar experimental design could generate the data required to measure niche and fitness differences for any tractable group of organisms. This sets up the exciting potential for meta-analyses to investigate generalities in the relative importance of coexistence mechanisms (Buche et al., 2022; Yan et al., 2022), notwithstanding the challenges in separating these fundamentally interlinked processes (Song et al., 2019). However, without individual studies reporting uncertainty in their coexistence predictions, their findings are not effectively usable by meta-analysts. Currently, it is difficult to move beyond the concern that the diverse range of conclusions regarding the relative importance of niche or fitness differences in facilitating coexistence may be influenced, at least partially, by researcher degrees of freedom.

Coexistence theory is on the precipice of moving from relatively abstract theory (Chesson, 2000), through pioneering initial demonstrations (Levine & HilleRisLambers, 2009) to hope for widespread adoption in empirical applications (Grainger, Levine, et al., 2019; Terry et al., 2022). In order to fully harness its potential benefits, it is essential that empirical studies stand on replicable and solid statistical foundations. Although this article offers many critical insights, we hold a strong sense of optimism regarding the future prospects of MCT in empirical research. We aim for this guide to contribute to charting a statistically rigorous course towards a productive and insightful future for the coexistence programme.

AUTHOR CONTRIBUTIONS

Both authors conceptualised the framework of the article. J. Christopher D. Terry led the writing of the first draft of the manuscript, initiated the literature review and made the examples. David W. Armitage contributed to the literature review and both authors contributed substantially to revisions and development of the manuscript.

ACKNOWLEDGEMENTS

We thank the two anonymous reviewers for their constructive comments. Chris Terry was supported by a fellowship from the Leverhulme Trust (ECF-2022-666).

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14227>.

DATA AVAILABILITY STATEMENT

This manuscript uses no new experimental data. All R code and data used to generate the figures are available at https://github.com/jcdterry/MCT_Review_public and archived on Zenodo at <https://doi.org/10.5281/zenodo.8113512> (Terry, 2023a).

ORCID

J. Christopher D. Terry  <https://orcid.org/0000-0002-0626-9938>

David W. Armitage  <https://orcid.org/0000-0002-5677-0501>

REFERENCES

- Abrams, P. A. (2022). *Competition theory in ecology*. Oxford University Press.
- Adler, P. B., Byrne, K. M., & Leiker, J. (2013). Can the past predict the future? Experimental tests of historically based population models. *Global Change Biology*, 19(6), 1793–1803. <https://doi.org/10.1111/gcb.12168>
- Adler, P. B., HilleRisLambers, J., Kyriakidis, P. C., Guan, Q., & Levine, J. M. (2006). Climate variability has a stabilizing effect on the coexistence of prairie grasses. *Proceedings of the National Academy of Sciences of the United States of America*, 103(34), 12793–12798. <https://doi.org/10.1073/pnas.0600599103>
- Adler, P. B., HilleRisLambers, J., & Levine, J. M. (2007). A niche for neutrality. *Ecology Letters*, 10(2), 95–104. <https://doi.org/10.1111/j.1461-0248.2006.00996.x>
- Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, 95(3), 631–636. <https://doi.org/10.1890/13-1452.1>
- Amarasekare, P. (2019). The evolution of coexistence theory. *Theoretical Population Biology*, 133, 49–51. <https://doi.org/10.1016/j.tpb.2019.09.005>
- Aoyama, L., Shoemaker, L. G., Gilbert, B., Collinge, S. K., Faist, A. M., Shackelford, N., Temperton, V. M., Barabás, G., Larios, L., Ladouceur, E., Godoy, O., Bowler, C., & Hallett, L. M. (2022). Application of modern coexistence theory to rare plant restoration provides early indication of restoration trajectories. *Ecological Applications*, 32(7), e2649. <https://doi.org/10.1002/eap.2649>
- Armitage, D. W. (2022). *To remain modern, the coexistence program requires modern statistical rigor*. *bioRxiv*. <https://doi.org/10.1101/2022.12.28.522056>
- Armitage, D. W., & Jones, S. E. (2020). Coexistence barriers confine the poleward range of a globally distributed plant. *Ecology Letters*, 23(12), 1838–1848. <https://doi.org/10.1111/ele.13612>
- Ayala, F. J., Gilpin, M. E., & Ehrenfeld, J. G. (1973). Competition between species: Theoretical models and experimental tests. *Theoretical Population Biology*, 4(3), 331–356. [https://doi.org/10.1016/0040-5809\(73\)90014-2](https://doi.org/10.1016/0040-5809(73)90014-2)
- Barabás, G., D'Andrea, R., & Stump, S. M. (2018). Chesson's coexistence theory. *Ecological Monographs*, 88(3), 277–303. <https://doi.org/10.1002/ecm.1302>
- Barabás, G., Michalska-Smith, J. M., & Allesina, S. (2016). The effect of intra- and interspecific competition on coexistence in multispecies communities. *The American Naturalist*, 188(1), E1–E12. <https://doi.org/10.1086/686901>

- Barabás, G., Pásztor, L., Meszéna, G., & Ostling, A. (2014). Sensitivity analysis of coexistence in ecological communities: Theory and application. *Ecology Letters*, 17(12), 1479–1494. <https://doi.org/10.1111/ele.12350>
- Bimler, M. D., Stouffer, D. B., Lai, H. R., & Mayfield, M. M. (2018). Accurate predictions of coexistence in natural systems require the inclusion of facilitative interactions and environmental dependency. *Journal of Ecology*, 106(5), 1839–1852. <https://doi.org/10.1111/1365-2745.13030>
- Blackford, C., Germain, R. M., & Gilbert, B. (2020). Species differences in phenology shape coexistence. *The American Naturalist*, 195(6), E168–E180. <https://doi.org/10.1086/708719>
- Bowler, C. H., Weiss-Lehman, C., Towers, I. R., Mayfield, M. M., & Shoemaker, L. G. (2022). Accounting for demographic uncertainty increases predictions for species coexistence: A case study with annual plants. *Ecology Letters*, 25(7), 1618–1628. <https://doi.org/10.1111/ele.14011>
- Buche, L., Spaak, J. W., Jarillo, J., & De Laender, F. (2022). Niche differences, not fitness differences, explain predicted coexistence across ecological groups. *Journal of Ecology*, 110(11), 2785–2796. <https://doi.org/10.1111/1365-2745.13992>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411.
- Cardinaux, A., Hart, S. P., & Alexander, J. M. (2018). Do soil biota influence the outcome of novel interactions between plant competitors? *Journal of Ecology*, 106(5), 1853–1863. <https://doi.org/10.1111/1365-2745.13029>
- Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781420010138>
- Carroll, I. T., Cardinale, B. J., & Nisbet, R. M. (2011). Niche and fitness differences relate the maintenance of diversity to ecosystem function. *Ecology*, 92(5), 1157–1165.
- Carroll, R. J., & Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79(386), 321–328. <https://doi.org/10.2307/2288271>
- Cervantes-Loreto, A., Pastore, A. I., Brown, C. R. P., Maraffini, M. L., Aldebert, C., Mayfield, M. M., & Stouffer, D. B. (2023). Environmental context, parameter sensitivity, and structural sensitivity impact predictions of annual-plant coexistence. *Ecological Monographs*, e1592. <https://doi.org/10.1002/ecm.1592>
- Chase, J. M., & Leibold, M. A. (2003). *Ecological niches: Linking classical and contemporary approaches*. University of Chicago Press.
- Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annual Review of Ecology, Evolution, and Systematics*, 31, 343–366.
- Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhout, S., & Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. *Nature Ecology & Evolution*, 2(3), 420–426. <https://doi.org/10.1038/s41559-017-0458-2>
- Dietze, M. C. (2017). *Ecological forecasting*. PUP. <https://press.princeton.edu/books/hardcover/9780691160573/ecological-forecasting>
- Ellner, S. P., Snyder, R. E., Adler, P. B., & Hooker, G. (2019). An expanded modern coexistence theory for empirical applications. *Ecology Letters*, 22(1), 3–18. <https://doi.org/10.1111/ele.13159>
- Epstein, G., Hawkins, S. J., & Smale, D. A. (2019). Identifying niche and fitness dissimilarities in invaded marine macroalgal canopies within the context of contemporary coexistence theory. *Scientific Reports*, 9(1), 8816. <https://doi.org/10.1038/s41598-019-45388-5>
- Fragata, I., Costa-Pereira, R., Kozak, M., Majer, A., Godoy, O., & Magalhães, S. (2022). Specific sequence of arrival promotes coexistence via spatial niche pre-emption by the weak competitor. *Ecology Letters*, 25(7), 1629–1639. <https://doi.org/10.1111/ele.14021>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460. <https://doi.org/10.1511/2014.111.460>
- Germain, R. M., Weir, J. T., & Gilbert, B. (2016). Species coexistence: Macroevolutionary relationships and the contingency of historical interactions. *Proceedings of the Royal Society B: Biological Sciences*, 283(1827), 20160047. <https://doi.org/10.1098/rspb.2016.0047>
- Godoy, O., Gómez-Aparicio, L., Matías, L., Pérez-Ramos, I. M., & Allan, E. (2020). An excess of niche differences maximizes ecosystem functioning. *Nature Communications*, 11(1), 4180. <https://doi.org/10.1038/s41467-020-17960-5>
- Godoy, O., & Levine, J. M. (2014). Phenology effects on invasion success: Insights from coupling field experiments to coexistence theory. *Ecology*, 95(3), 726–736. <https://doi.org/10.1890/13-1157.1>
- Godwin, C. M., Chang, F.-H., & Cardinale, B. J. (2020). An empiricist's guide to modern coexistence theory for competitive communities. *Oikos*, 129(8), 1109–1127. <https://doi.org/10.1111/oik.06957>
- Grainger, T. N., Letten, A. D., Gilbert, B., & Fukami, T. (2019). Applying modern coexistence theory to priority effects. *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), 6205–6210. <https://doi.org/10.1073/pnas.1803122116>
- Grainger, T. N., Levine, J. M., & Gilbert, B. (2019). The invasion criterion: A common currency for ecological research. *Trends in Ecology & Evolution*, 34(10), 925–935. <https://doi.org/10.1016/j.tree.2019.05.007>
- Grainger, T. N., Senthilnathan, A., Ke, P.-J., Barbour, M. A., Jones, N. T., DeLong, J. P., Otto, S. P., O'Connor, M. I., Coblenz, K. E., Goel, N., Sakarchi, J., Szojka, M. C., Levine, J., & Germain, R. M. (2021). An empiricist's guide to using ecological theory. *The American Naturalist*, 199, 1–20. <https://doi.org/10.1086/717206>
- Gya, R., Birkeli, K., Dahle, I. J., Foote, C. G., Geange, S. R., Lynn, J. S., Töpper, J. P., Vandvik, V., Zernichow, C., & Jenkins, G. B. (2023). Registered reports: A new chapter at Ecology & Evolution. *Ecology and Evolution*, 13(4), e10023. <https://doi.org/10.1002/ece3.10023>
- Hallett, L. M., Shoemaker, L. G., White, C. T., & Suding, K. N. (2019). Rainfall variability maintains grass-forb species coexistence. *Ecology Letters*, 22(10), 1658–1667. <https://doi.org/10.1111/ele.13341>
- Hart, S. P., Freckleton, R. P., & Levine, J. M. (2018). How to quantify competitive ability. *Journal of Ecology*, 106(5), 1902–1909. <https://doi.org/10.1111/1365-2745.12954>
- Hart, S. P., Schreiber, S. J., & Levine, J. M. (2016). How variation between individuals affects species coexistence. *Ecology Letters*, 19(8), 825–838. <https://doi.org/10.1111/ele.12618>
- Hart, S. P., Turcotte, M. M., & Levine, J. M. (2019). Effects of rapid evolution on species coexistence. *Proceedings of the National Academy of Sciences of the United States of America*, 116(6), 2112–2117. <https://doi.org/10.1073/pnas.1816298116>
- Hess, C., Levine, J. M., Turcotte, M. M., & Hart, S. P. (2022). Phenotypic plasticity promotes species coexistence. *Nature Ecology & Evolution*, 6(9), 1256–1261. <https://doi.org/10.1038/s41559-022-01826-8>
- Huntly, N., Freischel, A. R., Miller, A. K., Lloyd, M. C., Basanta, D., & Brown, J. S. (2021). Coexistence of “cream skimmer” and “crumb picker” phenotypes in nature and in cancer. *Frontiers in Ecology and Evolution*, 9, 697618. <https://doi.org/10.3389/fevo.2021.697618>
- Hutchinson, G. E. (1959). Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist*, 93(870), 145–159.
- Inouye, B. D. (2001). Response surface experimental designs for investigating interspecific competition. *Ecology*, 82(10), 2696–2706. <https://doi.org/10.1890/0012-9658>
- Jenkins, G. B., Beckerman, A. P., Bellard, C., Benítez-López, A., Ellison, A. M., Foote, C. G., Hufton, A. L., Lashley, M. A., Lortie, C. J., Ma, Z., Moore, A. J., Narum, S. R., Nilsson, J., O'Boyle, B., Provete, D. B., Razgour, O., Rieseberg, L., Riginos, C., Santini, L., ... Peres-Neto, P. R. (2023). Reproducibility in ecology and evolution: Minimum standards for data and code. *Ecology and Evolution*, 13(5), e9961. <https://doi.org/10.1002/ece3.9961>
- Johnson, C. A., Dutt, P., & Levine, J. M. (2022). Competition for pollinators destabilizes plant coexistence. *Nature*, 607, 721–725. <https://doi.org/10.1038/s41586-022-04973-x>

- Kandlikar, G. S., Yan, X., Levine, J. M., & Kraft, N. J. (2021). Soil microbes generate stronger fitness differences than stabilization among California annual plants. *The American Naturalist*, 197(1), E30–E39.
- Kimmel, K., Avolio, M. L., & Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, 1–12, 1525–1536. <https://doi.org/10.1038/s41559-023-02144-3>
- Kraft, N. J. B., Godoy, O., & Levine, J. M. (2015). Plant functional traits and the multidimensional nature of species coexistence. *Proceedings of the National Academy of Sciences of the United States of America*, 112(3), 797–802. <https://doi.org/10.1073/pnas.1413650112>
- Lanuza, J. B., Bartomeus, I., & Godoy, O. (2018). Opposing effects of floral visitors and soil conditions on the determinants of competitive outcomes maintain species diversity in heterogeneous landscapes. *Ecology Letters*, 21(6), 865–874. <https://doi.org/10.1111/ele.12954>
- Law, R., & Watkinson, A. R. (1987). Response-surface analysis of two-species competition: An experiment on *Phleum arenarium* and *Vulpia fasciculata*. *Journal of Ecology*, 75(3), 871–886. <https://doi.org/10.2307/2260211>
- Letten, A. D., Dhami, M. K., Ke, P.-J., & Fukami, T. (2018). Species coexistence through simultaneous fluctuation-dependent mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 115(26), 6745–6750. <https://doi.org/10.1073/pnas.1801846115>
- Letten, A. D., Hall, A. R., & Levine, J. M. (2021). Using ecological coexistence theory to understand antibiotic resistance and microbial competition. *Nature Ecology & Evolution*, 5, 431–441. <https://doi.org/10.1038/s41559-020-01385-w>
- Levine, J. M., Bascompte, J., Adler, P. B., & Allesina, S. (2017). Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, 546(7656), 56–64. <https://doi.org/10.1038/nature22898>
- Levine, J. M., & HilleRisLambers, J. (2009). The importance of niches for the maintenance of species diversity. *Nature*, 461(7261), 254–257. <https://doi.org/10.1038/nature08251>
- Martyn, T. E., Stouffer, D. B., Godoy, O., Bartomeus, I., Pastore, A. I., & Mayfield, M. M. (2021). Identifying “useful” fitness models: Balancing the benefits of added complexity with realistic data requirements in models of individual plant fitness. *The American Naturalist*, 197(4), 415–433. <https://doi.org/10.1086/713082>
- Matías, L., Godoy, O., Gómez-Aparicio, L., & Pérez-Ramos, I. M. (2018). An experimental extreme drought reduces the likelihood of species to coexist despite increasing intransitivity in competitive networks. *Journal of Ecology*, 106(3), 826–837. <https://doi.org/10.1111/1365-2745.12962>
- Milner-Gulland, E. J., & Shea, K. (2017). Embracing uncertainty in applied ecology. *Journal of Applied Ecology*, 54, 2063–2068. <https://doi.org/10.1111/1365-2664.12887>
- Mordecia, E. A. (2013). Despite spillover, a shared pathogen promotes native plant persistence in a cheatgrass-invaded grassland. *Ecology*, 94(12), 2744–2753. <https://doi.org/10.1890/13-0086.1>
- Narwani, A., Alexandrou, M. A., Oakley, T. H., Carroll, I. T., & Cardinale, B. J. (2013). Experimental evidence that evolutionary relatedness does not affect the ecological mechanisms of coexistence in freshwater green algae. *Ecology Letters*, 16(11), 1373–1381. <https://doi.org/10.1111/ele.12182>
- Nature Ecology & Evolution Editors. (2023). Rolling out registered reports. *Nature Ecology & Evolution*, 7(5), 625. <https://doi.org/10.1038/s41559-023-02076-y>
- Novak, M., & Stouffer, D. B. (2021a). Geometric complexity and the information-theoretic comparison of functional-response models. *Frontiers in Ecology and Evolution*, 9, 740362. <https://doi.org/10.3389/fevo.2021.740362>
- Novak, M., & Stouffer, D. B. (2021b). Systematic bias in studies of consumer functional responses. *Ecology Letters*, 24(3), 580–593. <https://doi.org/10.1111/ele.13660>
- Ocampo-Ariza, C., Bufford, J. L., Hulme, P. E., Champion, P. D., & Godsoe, W. (2018). Strong fitness differences impede coexistence between an alien water fern (*Azolla pinnata* R. Br.) and its native congener (*Azolla rubra* R. Br.) in New Zealand. *Biological Invasions*, 20(10), 2889–2897. <https://doi.org/10.1007/s10530-018-1740-1>
- Pande, J., Fung, T., Chisholm, R., & Shnerb, N. M. (2020). Mean growth rate when rare is not a reliable metric for persistence of species. *Ecology Letters*, 23(2), 282. <https://doi.org/10.1111/ele.13430>
- Parker, T. H., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, 33(4), 747–750. <https://doi.org/10.1111/cobi.13342>
- Parker, T. H., Griffith, S. C., Bronstein, J. L., Fidler, F., Foster, S., Fraser, H., Forstmeier, W., Gurevitch, J., Koricheva, J., Seppelt, R., Tingley, M. W., & Nakagawa, S. (2018). Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution*, 2(6), 929–935. <https://doi.org/10.1038/s41559-018-0545-z>
- Pascual, M. A., & Kareiva, P. (1996). Predicting the outcome of competition using experimental data: Maximum likelihood and Bayesian approaches. *Ecology*, 77(2), 337–349. <https://doi.org/10.2307/2265613>
- Pérez-Ramos, I. M., Matías, L., Gómez-Aparicio, L., & Godoy, Ó. (2019). Functional traits and phenotypic plasticity modulate species coexistence across contrasting climatic conditions. *Nature Communications*, 10(1), 2555. <https://doi.org/10.1038/s41467-019-10453-0>
- Petry, W. K., Kandlikar, G. S., Kraft, N. J. B. B., Godoy, O., & Levine, J. M. (2018). A competition-defence trade-off both promotes and weakens coexistence in an annual plant community. *Journal of Ecology*, 106(5), 1806–1818. <https://doi.org/10.1111/1365-2745.13028>
- Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, 331(6018), 703–705. <https://doi.org/10.1126/science.1197962>
- Rey, P. J., Manzaneda, A. J., & Alcántara, J. M. (2017). The interplay between aridity and competition determines colonization ability, exclusion and ecological segregation in the heteroploid *Brachypodium distachyon* species complex. *New Phytologist*, 215(1), 85–96. <https://doi.org/10.1111/nph.14574>
- Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1), 218–227. <https://doi.org/10.1111/j.1365-2664.2007.01377.x>
- Ruel, J. J., & Ayres, M. P. (1999). Jensen's inequality predicts effects of environmental variation. *Trends in Ecology and Evolution*, 14(9), 361–366.
- Saavedra, S., Rohr, R. P., Bascompte, J., Godoy, O., Kraft, N. J. B., & Levine, J. M. (2017). A structural approach for understanding multispecies coexistence. *Ecological Monographs*, 87(3), 470–486. <https://doi.org/10.1002/ecm.1263>
- Schmolke, A., Thorbek, P., DeAngelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: A strategy for the future. *Trends in Ecology & Evolution*, 25(8), 479–486. <https://doi.org/10.1016/j.tree.2010.05.001>
- Schreiber, S. J., Levine, J. M., Godoy, O., Kraft, N. J. B., & Hart, S. P. (2023). Does deterministic coexistence theory matter in a finite world? *Ecology*, 104(1), e3838. <https://doi.org/10.1002/ecy.3838>
- Sears, A. L. W., & Chesson, P. (2007). New methods for quantifying the spatial storage effect: An illustration with desert annuals. *Ecology*, 88(9), 2240–2247.
- Siefert, A., Zillig, K. W., Friesen, M. L., & Strauss, S. Y. (2019). Mutualists stabilize the coexistence of congeneric legumes. *The American Naturalist*, 193(2), 200–212. <https://doi.org/10.1086/701056>
- Simmonds, E. G., Adjei, K. P., Andersen, C. W., Hetle Aspheim, J. C., Battistin, C., Bulso, N., Christensen, H. M., Cretois, B., Cubero, R., Davidovich, I. A., Dickel, L., Dunn, B., Dunn-Sigouin, E., Dyrstad, K., Einum, S., Giglio, D., Gjerløw, H., Godefroidt, A., González-Gil, R., ... O'Hara, R. B. (2022). Insights into the quantification and reporting

- of model-related uncertainty across different disciplines. *iScience*, 25(12), 105512. <https://doi.org/10.1016/j.isci.2022.105512>
- Song, C., Barabás, G., & Saavedra, S. (2019). On the consequences of the interdependence of stabilizing and equalizing mechanisms. *The American Naturalist*, 194(5), 627–639. <https://doi.org/10.1086/705347>
- Spaak, J. W., & De Laender, F. (2020). Intuitive and broadly applicable definitions of niche and fitness differences. *Ecology Letters*, 23(7), 1117–1128. <https://doi.org/10.1111/ele.13511>
- Spaak, J. W., Ke, P.-J., Letten, A. D., & De Laender, F. (2023). Different measures of niche and fitness differences tell different tales. *Oikos*, 2023, e09573. <https://doi.org/10.1111/oik.09573>
- Spies, A.-N. (2018). *propagate: Propagation of uncertainty*. <https://CRAN.R-project.org/package=propagate>
- Stan Development Team. (2022). *Stan modeling language users guide and reference manual* (2.32) [Computer software]. <https://mc-stan.org>
- Stouffer, D. B. (2022). A critical examination of models of annual-plant population dynamics and density-dependent fecundity. *Methods in Ecology and Evolution*, 13(11), 2516–2530. <https://doi.org/10.1111/2041-210X.13965>
- Stouffer, D. B., Godoy, O., Riva, G. V. D., & Mayfield, M. M. (2021). The dimensionality of plant-plant competition. *bioRxiv*. <https://doi.org/10.1101/2021.11.10.467010>
- Stump, S. M., Song, C., Saavedra, S., Levine, J. M., & Vasseur, D. A. (2022). Synthesizing the effects of individual-level variation on coexistence. *Ecological Monographs*, 92(1), e01493. <https://doi.org/10.1002/ecm.1493>
- Terry, J. C. D. (2023a). Jcdterry/MCT_Review_public: V1.1.0. *Zenodo*. <https://doi.org/10.5281/zenodo.8113512>
- Terry, J. C. D. (2023b). *Uncertain competition coefficients undermine inferences about coexistence*.
- Terry, J. C. D., Chen, J., & Lewis, O. T. (2021). Natural enemies have inconsistent impacts on the coexistence of competing species. *Journal of Animal Ecology*, 90(10), 2277–2288. <https://doi.org/10.1111/1365-2656.13534>
- Terry, J. C. D., O'Sullivan, J. D., & Rossberg, A. G. (2022). Synthesising the multiple impacts of climatic variability on community responses to climate change. *Ecography*, 2022(5), e06123. <https://doi.org/10.1111/ecog.06123>
- Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6), e03336. <https://doi.org/10.1002/ecy.3336>
- Van Dyke, M. N., Levine, J. M., & Kraft, N. J. B. (2022). Small rainfall changes drive substantial changes in plant coexistence. *Nature*, 611(7936), 507–511. <https://doi.org/10.1038/s41586-022-05391-9>
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (R package Version 2.5.1) [Computer software]. <https://mc-stan.org/loo/>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Wainwright, C. E., HilleRisLambers, J., Lai, H. R., Loy, X., & Mayfield, M. M. (2019). Distinct responses of niche and fitness differences to water availability underlie variable coexistence outcomes in semi-arid annual plant communities. *Journal of Ecology*, 107(1), 293–306. <https://doi.org/10.1111/1365-2745.13056>
- Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, 7(8), 882–890. <https://doi.org/10.1111/2041-210X.12552>
- Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867–897.
- Weiss-Lehman, C. P., Werner, C. M., Bowler, C. H., Hallett, L. M., Mayfield, M. M., Godoy, O., Aoyama, L., Barabás, G., Chu, C., Ladouceur, E., Larios, L., & Shoemaker, L. G. (2022). Disentangling key species interactions in diverse and heterogeneous communities: A Bayesian sparse modelling approach. *Ecology Letters*, 25(5), 1263–1276. <https://doi.org/10.1111/ele.13977>
- Yan, X., Levine, J. M., & Kandlikar, G. S. (2022). A quantitative synthesis of soil microbial effects on plant species coexistence. *Proceedings of the National Academy of Sciences of the United States of America*, 119(22), e2122088119. <https://doi.org/10.1073/pnas.2122088119>
- Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology*, 28(3), 969–989. <https://doi.org/10.1111/gcb.15972>
- Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93(1), e1557. <https://doi.org/10.1002/ecm.1557>
- Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6), 636–645. [10.1111/2041-210X.12577](https://doi.org/10.1111/2041-210X.12577)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Impact of different approaches to modelling error on fitted parameters and predicted coexistence using real data.

Figure S2. Comparison of the fits of the five competition models described in Figure 3.

Figure S3. Fits to simulated data with different error distribution and transformation choices used in the main text Figure 4.

How to cite this article: Terry, J. C. D., & Armitage, D. W. (2024). Widespread analytical pitfalls in empirical coexistence studies and a checklist for improving their statistical robustness. *Methods in Ecology and Evolution*, 15, 594–611. <https://doi.org/10.1111/2041-210X.14227>