Received: 13 May 2022 | Accepted: 22 September 2022

DOI: 10.1111/2041-210X.14028

Methods in Ecology and Evolution

RESEARCH ARTICLE

Modelling ecological communities when composition is manipulated experimentally

Abigail Skwara[1,2] | Paula Lemos-Costa[1] | Zachary R. Miller[1] | Stefano Allesina[1,3]

[1]Department of Ecology & Evolution, University of Chicago, Chicago, Illinois, USA

[2]Department of Ecology & Evolutionary Biology, Yale University, New Haven, Connecticut, USA

[3]Northwestern Institute for Complex Systems, Northwestern University, Evanston, Illinois, USA

Correspondence
Abigail Skwara
Email: abby.skwara@yale.edu

Funding information
National Science Foundation, Grant/Award Number: DEB-#2022742 and DGE-1746045

Handling Editor: Hao Ye

Abstract

1. In an experimental setting, the composition of ecological communities can be manipulated directly. Starting from a pool of n species, it is possible to co-culture species in different combinations, ranging from monocultures, to pairs, and all the way up to the full species pool. Leveraging datasets with this experimental design, we advance methods to infer species interactions using density measurements taken at a single time point across a variety of distinct community compositions.
2. First, we introduce a fast and robust algorithm to estimate parameters for simple statistical models describing these data, which can be combined with likelihood maximization approaches. Second, we derive from consumer–resource dynamics a family of statistical models with few parameters, which can be applied to study systems where only a small fraction of the potential community compositions have been observed. Third, we show how a Weighted Least Squares framework can be used to account for the fact that species abundances often display a strong relationship between means and variances.
3. To illustrate our approach, we analyse datasets spanning plant, bacteria and phytoplankton communities, as well as simulations, consistently recovering a good fit to the data and demonstrating the ability of our methods to predict equilibrium densities in out-of-sample communities.
4. By combining more robust model structures and fitting procedures along with a more flexible error model, we greatly extend the applicability of recently proposed methods to model community composition from experimental data, opening the door for the analysis of larger pools of species using sparser and noisier datasets than was previously possible.

KEYWORDS
generalized linear model, generalized Lotka–Volterra model, species coexistence, species interactions, weighted least squares

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

1 | INTRODUCTION

Efforts to relate empirical measurements of population growth to dynamical models can be traced back to the origins of ecology. For example, in his famous study proposing the logistic growth model, Verhulst (1838) presented the fitted time-series for the human population growth of France and Belgium. Interest in contrasting empirical data with models for population dynamics has only grown through the years (Gilpin, 1973; Powell & Steele, 2012), with the development of sophisticated approaches (Ellner et al., 2002) accounting for different sources of errors (Carpenter et al., 1994; De Valpine & Hastings, 2002), and important applications such as forecasting population trajectories (Clark et al., 2001; Sugihara et al., 2012), modelling the evolution of disease epidemics (Du et al., 2017), and detecting chaos in natural systems (Perry et al., 2012; Sugihara & May, 1990).

Because of the intrinsic difficulty of manipulating natural systems, much of the literature on these issues has historically focused on inferring the parameters of dynamical models from time-series observations (Downing et al., 2020; Ives et al., 2003; Vandermeer, 1969). With the advent of high-throughput laboratory techniques, however, a different approach has become viable. Instead of attempting to learn the parameters of a model by analysing the fluctuations of several interacting populations through time, it is possible to use steady-state abundances recorded for a variety of community compositions, each captured at a single time point, to infer model parameters (Ansari et al., 2021; Fort, 2018; Maynard et al., 2020; Voit et al., 2021; Xiao et al., 2017). For a species pool of interest, the initial species composition is manipulated in a series of experiments, and then the resulting set of final community compositions can be used to estimate the parameters of a statistical model. For example, to infer the interactions among *n* species, one could perform a series of experiments where species are grown in isolation, or in pairs, triplets or larger subsets (Dormann & Roxburgh, 2005; Friedman et al., 2017). Once any transient dynamics have elapsed, the densities of all surviving species are recorded. Provided that a sufficiently large and diverse set of sub-systems (i.e. distinct species compositions) have been observed, it is possible to infer the parameters of a statistical model for species interactions from these static measurements (Carrara et al., 2015; Maynard et al., 2020). Such statistical models can be derived from corresponding dynamical models, thereby linking static estimates to models for population dynamics.

We build upon previous work (Carrara et al., 2015; Maynard et al., 2020; Xiao et al., 2017) in which the density of each species is assumed to be linearly related to the densities of the other species in a community. This statistical structure arises naturally from the ubiquitous generalized Lotka–Volterra (GLV) dynamical model (Lotka, 1920; Volterra, 1926)—although it is not necessary for dynamics to obey this simple model to yield a good fit to data (Maynard et al., 2020). While very easy to formulate, this type of statistical model is both difficult and computationally expensive to fit. Here, we extend the approach of Maynard et al. (2020) in three main ways to overcome issues that have limited the application of these methods. First, we introduce a fast iterative algorithm that finds parameters yielding a good fit to data. This algorithm can be used in conjunction with more rigorous—but less efficient—fitting approaches by providing a high-quality starting point for numerical likelihood optimization. We show that this hybrid approach is computationally very efficient and produces better results than current methods. Second, we derive simplified versions of the statistical model by exploiting the parallel between the model structure and Lotka–Volterra dynamics. These versions of the model require the estimation of fewer parameters, while still providing a good fit to data and retaining a clear ecological interpretation. Finally, we show that a more sophisticated error model for the observations results in a Weighted Least Squares (WLS) problem that can be solved efficiently. Extending the error structure offers more flexibility to fit empirical data, especially when the variance of experimental observations changes with the mean, as typically seen in ecological data (Grilli, 2020; Taylor, 1961). To illustrate these improvements, we examine four recently published datasets spanning communities of plants (Kuebbing et al., 2015), phytoplankton (Ghedini et al., 2022) and bacteria (Chu et al., 2021), as well as simulated data.

2 | MATERIALS AND METHODS

2.1 | Experimental setup and data

Given a pool of *n* species, suppose we have performed a large set of experiments in which different combinations of species are co-cultured for a suitably long time. At the end of each experiment (once transient dynamics have elapsed), we measure and record the biomass/density of each extant species. The experiments encompass a variety of initial compositions and initial densities, and are conducted with replication. Provided that our measurements, after any species extinctions, span a sufficient variety of sub-communities, we can fit a simple statistical model to these empirical 'endpoints', which may, in turn, be used to predict the densities of each species in any unobserved subset of species, and in particular whether the given subset will coexist.

To test our models, we use four recently published datasets that have a suitable experimental design. In particular, we consider two datasets from Kuebbing et al. (2015), who selected two pools (natives and non-natives) of four plants each, and grew them in 14 out of 15 possible combinations of species. Similarly, Ghedini et al. (2022) considered five phytoplankton species, and grew them in monoculture, in all possible pairs, and all together. Finally, we consider a subset of the data from Chu et al. (2021), consisting of four bacterial strains co-cultured along with *Pseudomonas fluorescens* in different combinations. A detailed description of each dataset is reported in Supporting Information S1.

2.2 | A simple statistical framework

We start by summarizing the statistical framework presented in Maynard et al. (2020), which we will extend below. The framework

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

assumes that we have measured densities for several of the possible communities of coexisting species we can form from a pool of n species. The approach rests on two main assumptions. First, we take each observed measurement to be a noisy realization of a 'true' value:

$$\tilde{x}_i^{(k)} = x_i^{(k)} + \epsilon_i^{(k)}. \tag{1}$$

That is, the observed density of species i when grown in community k, $\tilde{x}_i^{(k)}$, is given by a 'true' value, $x_i^{(k)}$, modified by an error term, $\epsilon_i^{(k)}$. We can arrange our empirical data in a matrix $\tilde{X}$, with n columns (one for each species in the pool) and as many rows as the number of observed communities. In particular, $\tilde{X}_{ki} = \tilde{x}_i^{(k)}$ if species i is present in community k, and zero otherwise. Similarly, $\mathscr{E}_{ki} = \epsilon_i^{(k)}$ if species i is in community k and zero otherwise. Then, we can rewrite Equation 1 in matrix form:

$$\tilde{X} = X + \mathscr{E}. \tag{2}$$

When several replicates are available, we assume that the density recorded for species i in community k and replicate r follows $\tilde{x}_i^{(k,r)} = x_i^{(k)} + \epsilon_i^{(k,r)}$; that is, all replicate measurements of species i in community k stem from the same true value (to recover the notation in Equation 2, we simply stack the replicates in matrix $\tilde{X}$ and the corresponding true means in matrix X). This assumption amounts to ruling out 'true multi-stability' in the underlying dynamics (Xiao et al., 2017)—if a set of species coexists, we require that it always reaches the same attractor (be it an equilibrium, cycle or chaotic attractor). Notice, however, that this framework is compatible with the fact that experiments initialized with the same set of species may yield distinct sets of coexisting species at the experimental endpoint. This could occur if stochastic dynamics or differences in initial densities drive two experimental systems to different attractors. Because this framework exclusively uses data gathered at the end of each experiment, it is completely blind to initial conditions; we only require that communities that reach the same final composition have also reached the same dynamical attractor.

Second, following Maynard et al. (2020), we assume the true species' densities in a given community are linearly related to each other. For any species i in community k, we can express these densities by writing:

$$X_{ki} = \alpha_i - \sum_{j \neq i} \beta_{ij} X_{kj},$$

where α_i is the density that species i would attain when grown in isolation, and the coefficients β_{ij} model the effects of the other species in community k on the density of species i. Importantly, β_{ij} depends only on the identities of the species, and not on the community we are modelling—this assumption amounts to ruling out higher-order interactions or other nonlinearities that would make the per-capita effect of species j on species i dependent on the state of the system. Rearranging, we obtain:

$$\sum_{j \neq i} \frac{\beta_{ij}}{\alpha_i} X_{kj} + \frac{1}{\alpha_i} X_{ki} = 1$$
$$\sum_j B_{ij} X_{kj} = 1,$$

with $B_{ii} = 1/\alpha_i$ and $B_{ij} = \beta_{ij}/\alpha_i$. Naturally, systems with sufficiently strong higher-order effects (or highly nonlinear systems) would be incompatible with this assumption. Indeed, Maynard et al. (2020) used simulations to show that, while the framework is generally quite robust to model misspecification, strong higher-order interactions result in poor fit and poor inference of true parameters. Conversely, good fit to the data would suggest the absence of strong higher-order interactions or nonlinearities.

Because $X_{kj} = 0$ whenever species j is not present in community k, we can define the sub-matrix $B^{(k)}$ obtained by retaining only the rows and columns of B corresponding to species that are present in community k. We similarly take $X^{(k)}$ to be a vector containing only the densities of the species in k, and $1^{(k)}$ to be a vector of ones with as many elements as $X^{(k)}$. Then the model can be written in matrix form as $B^{(k)}X^{(k)} = 1^{(k)}$ for each community k.

Suppose that we have estimated a matrix B and we want to make a prediction about a set of species k, such as whether these species can coexist. We solve the corresponding equation for $X^{(k)}$,

$$X^{(k)} = \left(B^{(k)}\right)^{-1} 1^{(k)}, \tag{3}$$

with two possible outcomes: (a) all the components in $X^{(k)}$ are positive, in which case we predict that the species can coexist with densities $X^{(k)}$ or (b) some of the components in $X^{(k)}$ are negative, which we interpret as the impossibility of coexistence of this combination of species (Maynard et al., 2020).

Arguably, the simplest version of this statistical model is obtained by assuming that errors are independent, identically distributed random values sampled from a normal distribution, such that $\tilde{X}_{ki} \sim \mathcal{N}\left(X_{ki}, \sigma^2\right)$ whenever species i is in community k. Then, fitting the model requires minimizing the sum of squared deviations between the observed data and model predictions (Ordinary Least Squares, OLS).

This suggests a straightforward method to fit the model: propose a matrix B, compute the predicted densities for all species in all observed communities using Equation 3, and search for the matrix B that minimizes the deviations between the observed data $\tilde{X}$ and the predicted X. Unfortunately, this method is quite inefficient and computationally very expensive, as it requires inverting a sub-matrix of B for each observed community, and there may be up to $2^n - 1$ unique community compositions that can be built from a pool of n species. Moreover, the problem of minimizing deviations is markedly non-convex—starting from different initial estimates, we are likely to converge to different (and thus in general sub-optimal) solutions.

To circumvent this problem, Maynard et al. (2020) proposed a simple analytical approach (dubbed the 'naïve method') to find a rough estimate of B from observed data; this initial estimate of B can then be used as a starting point for more sophisticated fitting routines. However, as discussed in their study, this method suffers from a number of issues (detailed in Supporting Information S3). In particular, the statistical model assumes that observations are noisy, while the naïve method assumes that they have been observed precisely

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and that instead Equation 3 only holds approximately. In practice, this means that this approach does not provide the maximum likelihood estimate for B when the data are noisy. One of the goals of the present work is to build an iterative algorithm that is not only computationally efficient, but that also improves upon the naïve estimate for B by yielding a superior starting point for subsequent parameter search.

3 | RESULTS

3.1 | A fast, iterative algorithm to estimate parameters

We want to find the maximum likelihood estimate for the matrix B, that is, the choice of B minimizing the sum of squared deviations:

$$\mathrm{SSQ} = \sum_k \sum_r \sum_i \left(\tilde{x}_i^{(k,r)} - x_i^{(k)} \right)^2 .$$

The computational bottleneck we face is that determining the predicted abundances $x_i^{(k)}$ from the matrix B (via Equation 3) for all community compositions is very expensive (requiring the inversion of many matrices). We therefore develop an algorithm in which this expensive calculation is performed only seldomly and at defined intervals, and optimization is carried out in between these steps without having to re-calculate the predicted $x_i^{(k)}$. To achieve this goal, we divide the process of optimizing B into two steps: a prediction step (computationally expensive), and a numerical optimization step (computationally cheaper). By alternating between the two steps, we quickly converge to a good draft matrix B.

To construct this algorithm, we derive two 'auxiliary' matrices that are useful for computation. Having arranged our data in the matrix $\tilde{X}$ as detailed above, we take $\tilde{X} = X + \mathscr{E}$ (Equation 2), transpose each side and multiply by B. We obtain the sum of two new matrices:

$$B\tilde{X}^T = BX^T + B\mathscr{E}^T = P^T + S^T. \tag{4}$$

In the remainder of this section, we use P and S simply as a convenient means to build our algorithm—we discuss their ecological interpretation in Supporting Information S4. From Equation 4, we have $B^{-1}P^T = X^T$ and $B^{-1}S^T = \mathscr{E}^T$. Our sum of squared deviations is simply the squared Frobenius norm of $\mathscr{E}$, $\sum_{ij}\mathscr{E}_{ij}^2 = \| \mathscr{E} \|_F^2 = \| \mathscr{E} \|$, and from Equations 2 and 4, we can write:

$$\begin{aligned} \| \mathscr{E}^T \| &= \| B^{-1}S^T \| \\ &= \| \tilde{X}^T - B^{-1}P^T \|. \end{aligned}$$

Our goal is therefore to find a matrix B such that $B^{-1}P^T$ is as close as possible to the observed data $\tilde{X}^T$. However, neither B nor P are known, although P can be calculated from X and B, and hence from B. We therefore attempt to minimize the sum of squared deviations through an iterative algorithm reminiscent of the expectation–maximization approach (Moon, 1996):

1. Propose a candidate matrix B;
2. Consider B to be fixed, and use it to compute X via Equation 3 (prediction step);
3. Compute $P^T = BX^T$;
4. Consider P to be fixed, and find an updated B^{-1} by numerically minimizing $\| \tilde{X}^T - B^{-1}P^T \|$ (optimization step). Invert it to recover an updated B;
5. Repeat steps 2–4 until convergence.

In principle, this iterative algorithm depends on the proposal of the initial matrix; for our numerical experiments, we always start from the identity matrix (i.e. no interactions between species). Finally, we perform an additional numerical optimization to refine the results produced by this algorithm. The algorithm above may return values of $X_{ki} < 0$, because the solution minimizing the sum of squared deviations does not need to contain only non-negative densities. Clearly, such solutions would be biologically unattainable, and qualitatively incompatible with the observed data. Thus, when numerically refining the solution, we only accept proposals that yield non-negative predictions for the observed densities.

While this algorithm is not guaranteed to find the maximum likelihood estimate of B, we observe that in practice, the combination of our iterative algorithm and the final numerical optimization step yields very good solutions. In all cases, we converge on a solution that is superior to the result using the naïve method. As shown in Figure 1 (and Supporting Information S8), the SSQ typically decays rapidly with the number of iterative steps, and the final numerical search provides an additional improvement. While for the data presented here the algorithm converges smoothly, in principle, one could observe the SSQ oscillating as the algorithm progresses. As for gradient descent and similar methods, this problem can be alleviated by introducing a 'momentum' (Polyak, 1964; Qian, 1999): when updating matrix B in Step 4, one could take a weighted average of the matrix used to compute P (the current estimate of B), and the matrix that maximizes the likelihood given P (the new estimate of B). The addition of momentum would help achieve a smooth, monotonic decay of the SSQ, at the cost of having to take a larger number of steps before convergence.

3.2 | Simplified versions of the model

Fitting the n^2 parameters of the interaction matrix B requires having observed a sufficiently varied set of experimental community compositions. It is necessary to observe each of the n species in at least n distinct communities, and each pair of species coexisting in at least one community (for a full derivation of these conditions, see Supporting Information S2). These are very demanding requirements, and therefore many published datasets do not contain a sufficient variety of communities to allow the identification of all parameters. These conditions grow more onerous with the number of species in the pool, making the approach impractical for species pools of even moderate size. To address this issue, we

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

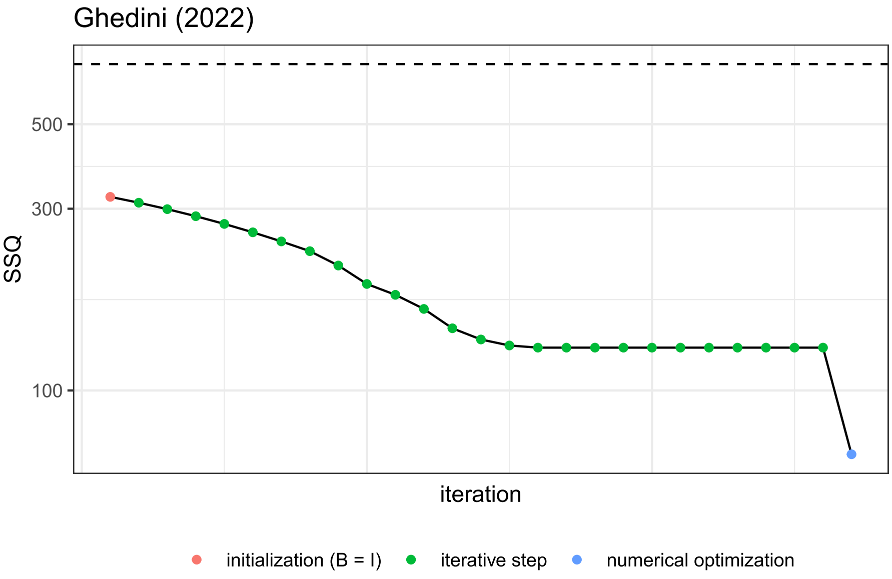

FIGURE 1 Sum of squared deviations (shown in log-scale, y-axis) for the initial matrix $B = I$, at each successive iteration of the algorithm, and after the numerical maximization (colours). The data from Ghedini et al., (2022). Is used as an example. The dashed line marks the SSQ obtained using the naïve method of Maynard et al. (2020).

propose a nested set of simplified versions of the statistical model that use fewer parameters, but retain the basic model structure and a straightforward ecological interpretation. The data requirements are greatly reduced, scaling linearly, rather than quadratically with n (see Supporting Information S7 for detailed data requirements). These simpler models have the added benefit of being extremely efficient to fit from a computational standpoint (Supporting Information S6), and useful for model selection, especially when one suspects that the full model with n^2 parameters may be susceptible to overfitting.

We follow a disciplined approach to develop these simpler models: we consider a version of the MacArthur's consumer–resource model (MacArthur, 1970) in which each species has access to its own private resources, and all species have access to a shared resource (Supporting Information S5). By progressively reducing the number of free parameters in the model, we obtain simpler structures for the matrix B (Table 1).

Given that the matrix B can be interpreted as detailing pairwise interactions (i.e. the effect of the density of species j on species i), the model $B = D(d) + vw^T$ corresponds to the following ecological picture: each species interacts with conspecifics through their private resources (corresponding to the coefficients d_i) and with all species via the shared resource. Interactions arising from the shared resource are given by the product of a resource utilization vector w (i.e. attack rates), and a resource transformation vector v, where each species is characterized by its v_i and w_i values. The interpretation of the simpler models is similar: by considering equal transformation rates one makes $vw^T = vv^T$ symmetric, and by assuming that all species also have the same attack rates, $vw^T = \alpha 1 1^T$. Note that in all these simplified models, the intraspecific interactions (i.e. the diagonal elements of B) are modelled with great flexibility; the reduction in parameters is obtained by assuming that interspecific interactions follow a simple pattern, defined by a few traits for all species.

In Figure 2, we show the fit of these four models when analysing the data reported in Ghedini et al. (2022).

For all datasets, we find the same qualitative results: the full model (n^2 parameters, 25 parameters for the dataset of Ghedini et al. (2022)) and the simplified model in which only two values per species determine all interspecific interactions ($3n - 1$ parameters, 14 for Ghedini et al. (2022)) have very similar performance (Supporting Information S8), while any further simplification results in a marked loss of fit. These trends are evident when contrasting the total SSQ across models and datasets (Figure 3).

3.3 | Allowing the variance to change with the mean

So far, we have assumed that errors are independent and identically distributed for all measurements. In many situations, however, this assumption would be quite unrealistic. For example, some species could systematically grow to much higher density than others—resulting in potentially larger absolute errors in their measurement—or measurements might be made on a small sub-plot or sample volume and then extrapolated to the whole plot or volume through multiplication. In such cases, the data would display marked heteroskedasticity (i.e. the variance would change with species density).

In Figure 4, we show that the variance scales with a power of the mean in the phytoplankton data from Ghedini et al. (2022). This is in fact the expected behaviour of many ecological systems, as posited by Taylor's law (Gaston & McArdle, 1994; Routledge & Swartz, 1991; Taylor, 1961).

A possible approach to dealing with the systematic heteroskedasticity observed in these datasets is to use WLS instead of OLS. To implement this approach, each squared residual is divided by the variance of the corresponding distribution, which is equivalent to measuring residuals as standard deviations of the standardized data. Here for simplicity we use the variance in the observed $\tilde{x}_i^{(k,r)}$ to estimate the variance $\sigma_i^{(k)2}$ (a more sophisticated approach would

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Simpler versions of the model derived by considering a MacArthur's consumer–resource model in which species have access to a private pool of resources as well as a shared pool of resources. See Supporting Information S5 for a derivation

Model	Number of parameters	Interpretation
B arbitrary	n^2	Arbitrary interactions between the species
$B = D(d) + vw^T$	$3n - 1$	Arbitrary intraspecific interactions; interspecific interactions $B_{ij} = v_i w_j$ given by the product of the 'resource transformation' of species i (v_i), and the 'attack rate' of species j (w_j)
$B = D(d) + vv^T$	$2n$	Arbitrary intraspecific interactions; resource transformation is the same for all species
$B = D(d) + \alpha 11^T$	$n + 1$	Arbitrary intraspecific interactions; resource transformation and attack rates are the same for all species

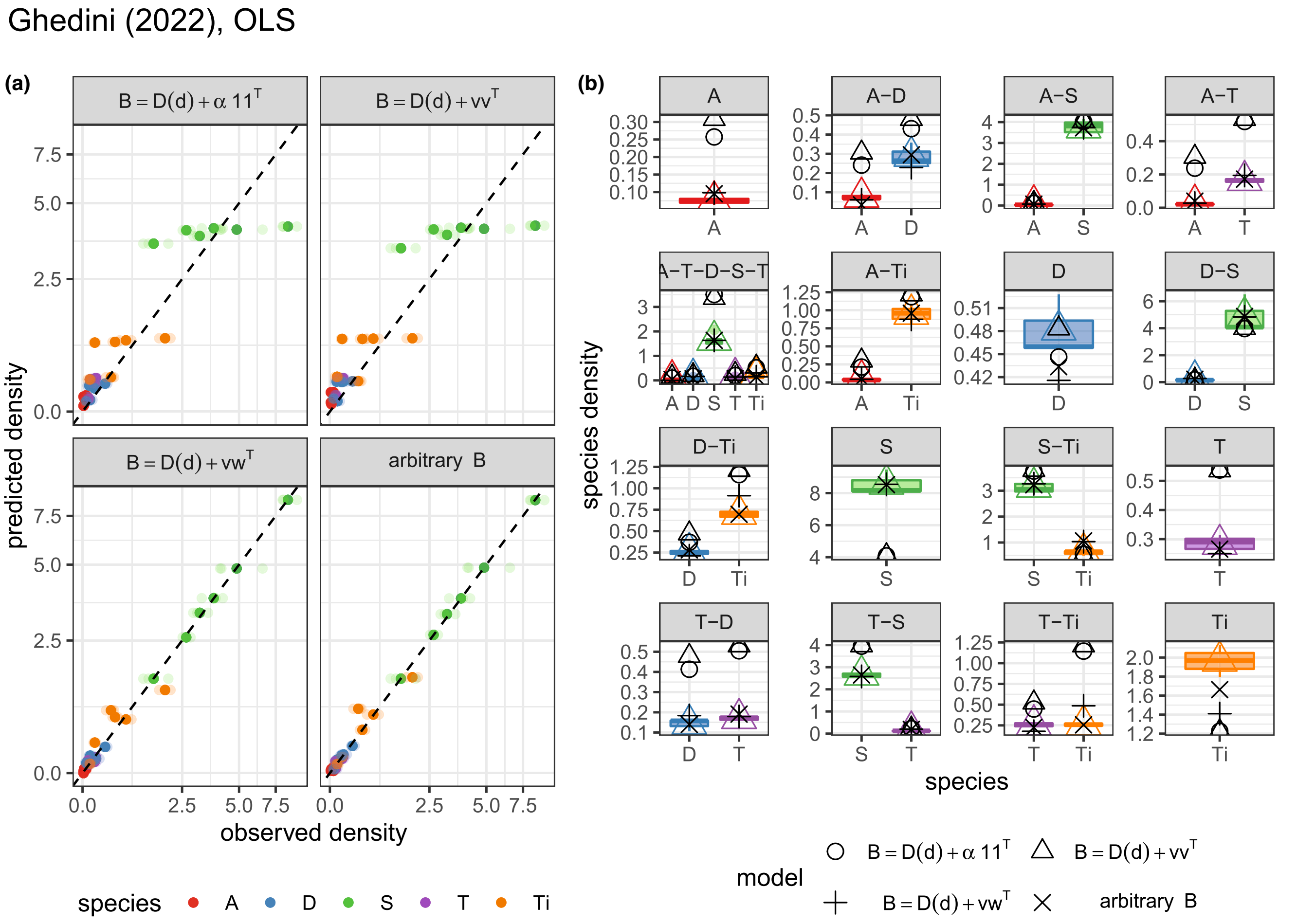

FIGURE 2 (a) Observed species (colour) densities (x-axis) vs. predicted densities (y-axis) for the data from Ghedini et al. (2022), when fitting the four versions of the model (panels). Replicate measurements for each species/community are reported as semi-transparent points and the mean for each species/community combination as a solid point. (b) Observed species (colour, x-axis) densities (y-axis). Boxplots show the distribution of the species densities across replicates, with the median density reported as a solid coloured line; the mean density is represented by the coloured triangle. Predicted values for each species in each community are represented by black open symbols (one for each of the four versions of the model).

call for the simultaneous estimation of means and variances, as in methods based on iteratively reweighted least squares). The WLS approach can be interpreted as reweighting the relative importance of errors made in our predictions, such that small errors made when estimating low species abundances are penalized as much as larger errors made when predicting higher species abundances. This is in contrast to OLS, where a 1% error in estimating the density of a species with high abundance contributes much more to the sum of squared deviations than the same proportional error for a species with low abundance.

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

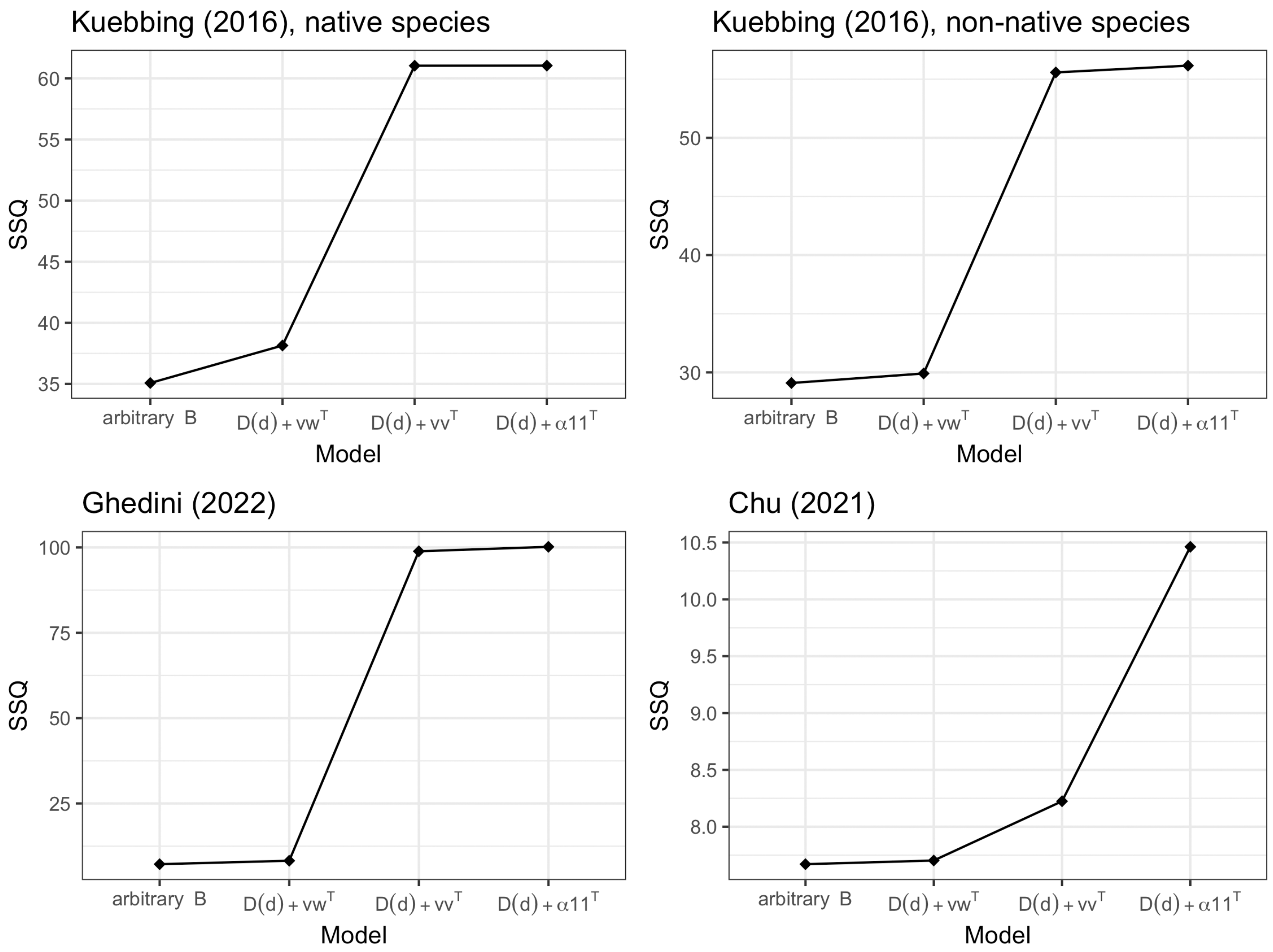

FIGURE 3 Sum of squared deviations between observed and predicted densities for the four datasets considered, fitted using the four models in Table 1 in order of decreasing model complexity.

Since experimental replicates are necessary to estimate these variances, we must have replicates for all communities (as we do for all the datasets considered here) to implement the WLS approach in this straightforward manner.

In Figure 5 (and Supporting Information S8), we compare these two error structures and find that we indeed observe a better fit for species with lower abundances when the WLS approach is used. This has important implications for the qualitative prediction of species' coexistence, because for a species present at low abundance in a community, the numerical difference between a positive abundance and a negative abundance (corresponding to a predicted lack of coexistence) could be quite small.

Finally, in Supporting Information S9, we perform simulations in which observations are taken from distributions with known mean–variance relationship, and fit the data using OLS, WLS or a likelihood-based approach. In all cases, we find that WLS outperforms OLS, allowing us to fit the data well and recover with good confidence the parameters used to generate the data.

3.4 | Predicting experiments out-of-sample

For the empirical datasets considered here, we always find good agreement between the observed and the fitted values when using the full model under either an OLS or WLS scheme. We also consider a leave-one-out (LOO) cross-validation approach, to verify that the models capture real features of the interactions between species in the communities, and are not simply over-fitting the data. For a dataset in which m experimental communities have been measured, we implement the LOO approach by designating one of the communities (along with any replicates) as out-of-sample (sometimes also called 'out-of-fit', as in Maynard et al. (2020)), and fitting our model on the remaining $m-1$ communities. This process can be repeated for each of the m communities in turn, and we then compare the predicted species' abundances with their experimentally observed values, as shown in Figure 6 (and Supporting Information S8). While the quality of the predictions is necessarily worse, in almost all cases we would have correctly predicted the experimental outcome both qualitatively (i.e. possibility of coexistence) and quantitatively.

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

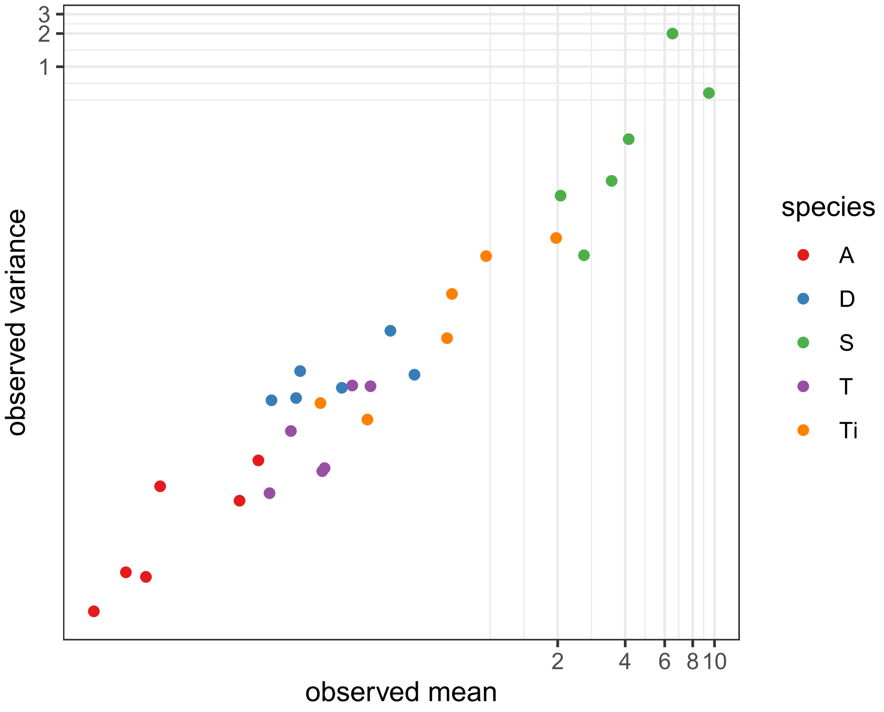

FIGURE 4 Mean (*x*-axis) vs. variance (*y*-axis) for the data published by Ghedini et al. (2022). For each species (colours) and community combination, the mean and variance of the observed species densities are computed across replicates. Note that both axes have a logarithmic scale, and thus the strong linear trend displayed by the data corresponds to a power-law relationship between the mean and the variance.

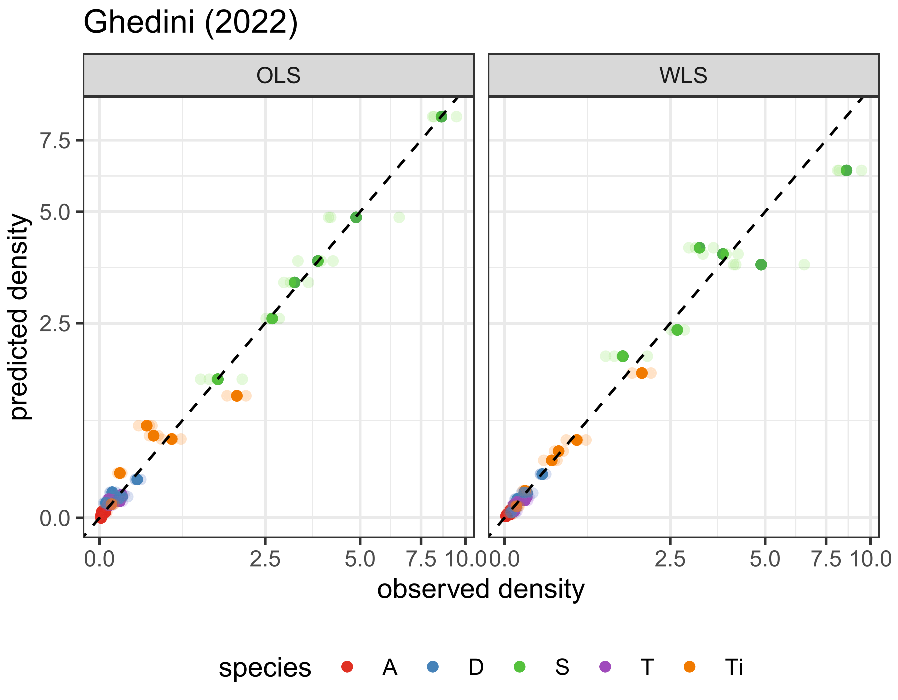

FIGURE 5 Observed (*x*-axis) vs. predicted (*y*-axis) species densities in all communities using the data by Ghedini et al. (2022), when fitting the simplified model $B = D(d) + vw^T$ and either minimizing the sum of squared deviations (Ordinary Least Squares [OLS]) or the sum of standardized squared deviations (Weighted Least Squares [WLS]). Replicate measurements for each species/community are reported as semi-transparent points; the mean for each species/community combination is reported as a solid point. In OLS, small (proportional) deviations of highly abundant species (e.g. *Synechococcus* sp., in green) are penalized more than larger (proportional) deviations of lower-abundance species (e.g. *Tisochrysis lutea*, in yellow). In contrast, when performing WLS each data point is standardized by its corresponding variance, levelling the importance of each measurement.

4 | DISCUSSION

In this study, we have further developed a simple, extensible framework to infer species' interactions from experimental data where community composition has been manipulated, and showcased its potential by testing this approach on datasets spanning plant, phytoplankton and bacterial communities, as well as simulations. We have improved the computational performance of the fitting routines,

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

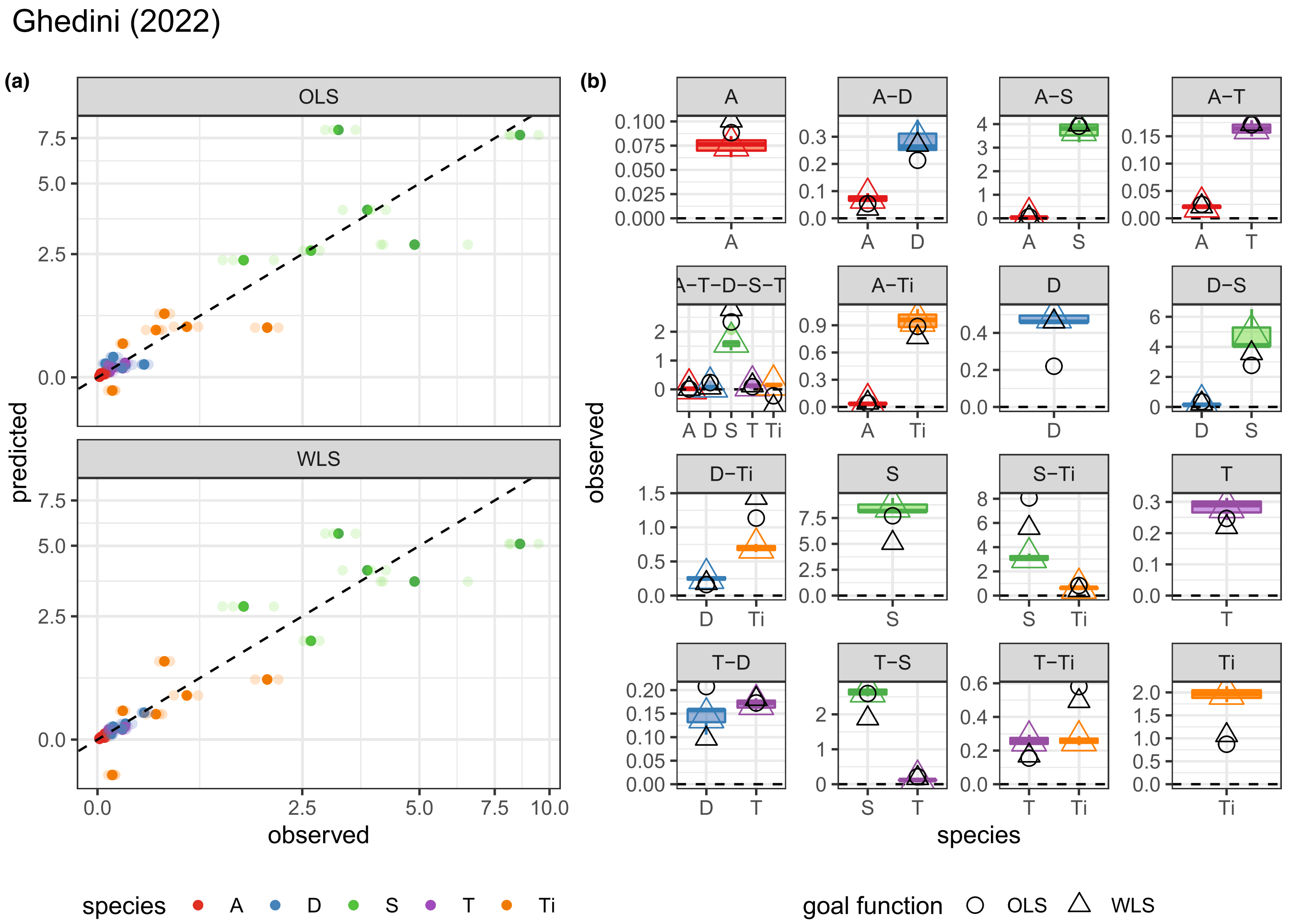

FIGURE 6 As Figure 2, but showing only out-of-sample predictions. Each panel is obtained by excluding the corresponding data, fitting the model using the remaining data and then predicting the data reported in the panel out-of-sample. Results are presented for the model $B = D(d) + vw^T$ under either Ordinary Least Squares (OLS) or Weighted Least Squares (WLS). Note that in the panel displaying the measurements for the full community (A-T-D-S-Ti), *Tisochrysis lutea* is predicted at a negative abundance—Without having observed these data, our models would suggest a lack of coexistence.

derived simplified versions of the model and extended the error structure used to capture the variability in the data.

Previous studies have used a variety of approaches to model this type of experimental data, including bottom-up 'assembly rules', in which the focus is on observing all possible pairs of species and using the outcomes to predict coexistence of larger species assemblages (Dormann & Roxburgh, 2005; Friedman et al., 2017); methods relying on monocultures and LOO communities (Ansari et al., 2021; Venturelli et al., 2018), using the extremes of very simple and highly speciose communities to infer pairwise interactions; and methods based on time-series that use fluctuations to estimate interaction strengths and the effects of environmental variables on ecological dynamics (Downing et al., 2020; Ives et al., 2003).

Our framework bridges the dynamical and statistical perspectives of these approaches. Importantly, though the models presented here are statistical in nature, they have a straightforward connection to the GLV dynamical model. By exploiting this parallel, we have derived a family of simplified models that put a premium on the ecological interpretability of the parameters. While an alternative approach to model regularization would be data-driven, machine learning methods (e.g. enforcing the sparsity or parsimony of B through a penalized regression), we have shown that ecologically motivated model constraints are a viable option. By clearly relating statistical and dynamical models, we retain the ability to probe other properties of these systems, for example in relation to invasibility, assembly and dynamical stability (Maynard et al., 2020).

Using our framework, we are able to obtain quantitative predictions for coexistence and abundances of an arbitrary set of species both in- and out-of-sample, even for datasets that comprise just a fraction of the possible sub-communities that may be formed from a fixed pool of species. Here, our simpler models are used to fit the relatively sparse datasets when fitting the full model is impossible. As the number of combinations that can be formed from a set of n species grows exponentially with n, this ability to predict species abundances out-of-sample is critical for exploring larger systems.

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Additionally, the good performance out-of-sample indicates that the models are capturing meaningful information about species interactions, rather than merely over-fitting the data.

Remarkably, we find that a model where interspecific interactions are approximated by a simplified structure, while intraspecific interactions are modelled more freely, achieves a comparable level of accuracy to the full model for the datasets. These results suggest that interspecific interactions in these systems are not completely idiosyncratic, but rather are largely characterized by a lower-dimensional structure. This observation agrees with the finding that simple rules govern the structure of interactions in microbial communities (Friedman et al., 2017), as well as recent work suggesting that plant–plant competitive interactions are characterized by low dimensionality (Stouffer et al., 2021). Similarly, these patterns are consistent with the success of a sparse-modelling approach demonstrating that, when considering only a few focal species of plants, many heterospecific interactions can be captured by a generic interaction term (Weiss-Lehman et al., 2022). Our results suggest that effective interspecific interactions sit in between extremely low dimensional (i.e. as in the model with identical interspecific interactions, which has a poor performance across all datasets) and fully structured pairwise interactions (arbitrary *B*).

Naturally, to develop these simple models, we need to make certain strong assumptions about the community dynamics. Here we have ruled out 'true-multi-stability', in which a set of species can coexist at distinct configurations of abundances. We have also neglected higher-order or highly nonlinear interactions, which would make the effect of species *i* on *j* context dependent. While we would expect this framework to perform poorly when these assumptions are violated, the good agreement with experimental data suggests that departures from these strict assumptions are modest. However, relaxing these assumptions could further expand the applicability of these methods.

Another area that deserves further exploration is quantifying uncertainty in the point estimates produced by this approach. While one could gauge these effects via bootstrapping of experimental data, we instead advocate a fully Bayesian approach to uncertainty quantification, for example as implemented by Maynard et al. (2020), because deriving a posterior distribution for the matrix *B* would also allow one to determine the probability of coexistence for a set of species, and better characterize the correlations between species abundances. Both bootstrapping and Markov chain Monte Carlo would require evaluating the predicted values for a set of parameters a large number of times. In this respect, the computational gains afforded by our simplified models could be key to making such approaches viable in future studies.

The main outstanding problem with our approach is the assumption that dynamics have reached a steady state, and thus the observed community composition is the 'true' final composition for the system. Violations of this assumption can greatly complicate inference. Suppose, for example, that we have two species, *A* and *B*, and that *A* excludes *B* asymptotically. If we sample this system before the extinction of *B*, we force the model to find parameters consistent with the robust coexistence of *A* and *B*, thereby biasing the results considerably. This problem of 'spurious coexistence' is especially troublesome for microbial communities, where species' presence is frequently determined by sequencing. Sequencing-based methods often detect some species at very low densities (Venturelli et al., 2018), potentially spuriously, making it difficult to discriminate between coexistence of rare species and actual extinctions masked by 'background noise'. A Bayesian approach could make it possible to simultaneously impute the 'true' final composition (i.e. which species are truly coexisting in the sample, taken as a latent parameter) as well as determine the distribution of parameters.

This general framework can be further extended in a number of directions. For example, one could introduce a more sophisticated error model assuming that species abundances are correlated within communities (e.g. overestimating the density of a predator could be associated with an underestimate of the densities of its prey). Similarly, we could assume more generally that observed densities $\tilde{x}_i^{(k,r)}$ are sampled from a particular distribution (e.g. Gamma or Inverse Gaussian), with mean $x_i^{(k)}$ and ancillary parameters controlling the shape of the distribution. In this case, instead of minimizing the sum of squared deviations, we would maximize the likelihood of the parameters for the chosen distribution. Preliminary results shown in Supporting Information S9 demonstrate that indeed we are able to recover the parameters used to generate simulated data with specified error models. The ability to model the data using a variety of distributions would bring this framework one step closer to the flexibility that characterizes Generalized Linear Models while maintaining the connection to ecological dynamics.

Overall, the methods presented here make it easier to contrast experimental data with simple models for population dynamics, returning parameters that have clear ecological interpretation, and allowing us to test predictions in a straightforward manner. The type of ecological data examined here are appearing with increasing frequency in the ecological literature, and these methods provide a complementary (or alternative) approach to model-fitting via time-series analysis. The minimal data needed to fit the simplified models and the fact that each experimental community can be measured just once (possibly destructively) make this framework especially appealing and cost-effective.

AUTHOR CONTRIBUTIONS

All authors contributed to the development of the methods. Abigail Skwara and Stefano Allesina wrote the code, analysed the data and wrote the manuscript. All authors edited the manuscript.

ACKNOWLEDGEMENTS

SA and PL-C were supported by the National Science Foundation, award DEB-#2022742 to SA. ZRM was supported by a National Science Foundation Graduate Research Fellowship (DGE-1746045).

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We thank P. Lechon-Alonso and K. Della Libera for comments. We thank an anonymous reviewer and D. Stouffer for constructive criticism.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14028.

DATA AVAILABILITY STATEMENT

All data and model code necessary to recreate the analyses presented in this manuscript are available at https://github.com/StefanoAllesina/skwara_et_al_2022 and archived in Zenodo at https://doi.org/10.5281/ZENODO.7240345 (Skwara et al., 2022).

ORCID

Abigail Skwara https://orcid.org/0000-0001-8062-5921
Paula Lemos-Costa https://orcid.org/0000-0001-6983-2022
Zachary R. Miller https://orcid.org/0000-0001-9762-8301
Stefano Allesina https://orcid.org/0000-0003-0313-8374

REFERENCES

Ansari, A. F., Reddy, Y., Raut, J., & Dixit, N. M. (2021). An efficient and scalable top-down method for predicting structures of microbial communities. *Nature Computational Science*, *1*(9), 619–628.

Carpenter, S., Cottingham, K., & Stow, C. (1994). Fitting predator-prey models to time series with observation errors. *Ecology*, *75*(5), 1254–1264.

Carrara, F., Giometto, A., Seymour, M., Rinaldo, A., & Altermatt, F. (2015). Inferring species interactions in ecological communities: A comparison of methods at different levels of complexity. *Methods in Ecology and Evolution*, *6*(8), 895–906.

Chu, X.-L., Zhang, Q.-G., Buckling, A., & Castledine, M. (2021). Interspecific niche competition increases morphological diversity in multi-species microbial communities. *Frontiers in Microbiology*, *2103*. https://doi.org/10.3389/fmicb.2021.699190

Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M., Pascual, M., Pielke, R. J., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W. H., Wall, D. H., & Wear, D. (2001). Ecological forecasts: An emerging imperative. *Science*, *293*(5530), 657–660.

De Valpine, P., & Hastings, A. (2002). Fitting population models incorporating process noise and observation error. *Ecological Monographs*, *72*(1), 57–76.

Dormann, C. F., & Roxburgh, S. H. (2005). Experimental evidence rejects pairwise modelling approach to coexistence in plant communities. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1569), 1279–1285.

Downing, A. L., Jackson, C., Plunkett, C., Ackerman Lockhart, J., Schlater, S. M., & Leibold, M. A. (2020). Temporal stability vs. community matrix measures of stability and the role of weak interactions. *Ecology Letters*, *23*(10), 1468–1478.

Du, X., King, A. A., Woods, R. J., & Pascual, M. (2017). Evolution-informed forecasting of seasonal influenza a (H3N2). *Science Translational Medicine*, *9*(413).

Ellner, S. P., Seifu, Y., & Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology*, *83*(8), 2256–2270.

Fort, H. (2018). On predicting species yields in multispecies communities: Quantifying the accuracy of the linear Lotka-Volterra generalized model. *Ecological Modelling*, *387*, 154–162.

Friedman, J., Higgins, L. M., & Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, *1*(5), 1–7.

Gaston, K. J., & McArdle, B. H. (1994). The temporal variability of animal abundances: Measures, methods and patterns. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1314), 335–358.

Ghedini, G., Marshall, D. J., & Loreau, M. (2022). Phytoplankton diversity affects biomass and energy production differently during community development. *Functional Ecology*, *36*, 446–457.

Gilpin, M. E. (1973). Do hares eat lynx? *The American Naturalist*, *107*(957), 727–730.

Grilli, J. (2020). Macroecological laws describe variation and diversity in microbial communities. *Nature Communications*, *11*(1), 1–11.

Ives, A. R., Dennis, B., Cottingham, K., & Carpenter, S. (2003). Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, *73*(2), 301–330.

Kuebbing, S. E., Classen, A. T., Sanders, N. J., & Simberloff, D. (2015). Above-and below-ground effects of plant diversity depend on species origin: An experimental test with multiple invaders. *New Phytologist*, *208*(3), 727–735.

Lotka, A. J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences of the United States of America*, *6*(7), 410–415.

MacArthur, R. (1970). Species packing and competitive equilibrium for many species. *Theoretical Population Biology*, *1*(1), 1–11.

Maynard, D. S., Miller, Z. R., & Allesina, S. (2020). Predicting coexistence in experimental ecological communities. *Nature Ecology & Evolution*, *4*(1), 91–100.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47–60.

Perry, J. N., Smith, R. H., Woiwod, I. P., & Morse, D. R. (2012). *Chaos in real data: The analysis of non-linear dynamics from short ecological time series* (Vol. *27*). Springer Science & Business Media.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, *4*(5), 1–17.

Powell, T. M., & Steele, J. H. (2012). *Ecological time series*. Springer Science & Business Media.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, *12*(1), 145–151.

Routledge, R. D., & Swartz, T. B. (1991). Taylor's power law re-examined. *Oikos*, *60*, 107–112.

Skwara, A., Lemos-Costa, P., Miller, Z. R., & Allesina, S. (2022). StefanoAllesina/skwara_et_al_2022: v0.01. https://doi.org/10.5281/ZENODO.7240345

Stouffer, D. B., Godoy, O., Dalla Riva, G. V., & Mayfield, M. M. (2021). The dimensionality of plant–plant competition. *bioRxiv*. https://doi.org/10.1101/2021.11.10.467010

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500.

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*(6268), 734–741.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*(4766), 732–735.

Vandermeer, J. H. (1969). The competitive structure of communities: An experimental approach with protozoa. *Ecology*, *50*(3), 362–371.

Venturelli, O. S., Carr, A. V., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., Hromada, S., Northen, T., & Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology*, *14*(6), e8157.

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Verhulst, P.-F. (1838). Notice sur la loi que la population poursuit dans son accroissement. *Correspondence Mathematique et Physique*, *10*, 113–121.

Voit, E., Davis, J., & Olivenca, D. (2021). Inference and validation of the structure of Lotka-Volterra models. *bioRxiv*. https://doi.org/10.1101/2021.08.14.456346

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, *118*, 558–560.

Weiss-Lehman, C. P., Werner, C. M., Bowler, C. H., Hallett, L., Mayfield, M. M., Godoy, O., Aoyama, L., Barabás, G., Chu, C., Ladouceur, E., Larios, L., & Shoemaker, L. (2022). Disentangling key species interactions in diverse and heterogeneous communities: A Bayesian sparse modeling approach. *Ecology Letters*, *25*(5), 1263–1276.

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., & Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. *Nature Communications*, *8*(1), 1–12.

SUPPORTING INFORMATION

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

How to cite this article: Skwara, A., Lemos-Costa, P., Miller, Z. R., & Allesina, S. (2023). Modelling ecological communities when composition is manipulated experimentally. *Methods in Ecology and Evolution*, *14*, 696–707. https://doi.org/10.1111/2041-210X.14028

2041210x, 2023, 2, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14028, Wiley Online Library on [08/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License