Received: 18 January 2019 | Accepted: 17 October 2019

DOI: 10.1111/1365-2745.13316

RESEARCH ARTICLE

Journal of Ecology BRITISH ECOLOGICAL SOCIETY

Predicting species abundances in a grassland biodiversity experiment: Trade-offs between model complexity and generality

Adam Thomas Clark[1,2,3] | Lindsay Ann Turnbull[4] | Andrew Tredennick[5] | Eric Allan[6] | W. Stanley Harpole[1,2,7] | Margaret M. Mayfield[8] | Santiago Soliveres[9] | Kathryn Barry[2,10] | Nico Eisenhauer[2,10] | Hans de Kroon[11] | Benjamin Rosenbaum[2,12] | Cameron Wagg[13,14] | Alexandra Weigelt[2,10] | Yanhao Feng[1,2] | Christiane Roscher[1,2] | Bernhard Schmid[14]

[1]Department of Physiological Diversity, Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany; [2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany; [3]Synthesis Centre for Biodiversity Sciences (sDiv), Leipzig, Germany; [4]Department of Plant Sciences, University of Oxford, Oxford, UK; [5]Odum School of Ecology and the Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA, USA; [6]Institute of Plant Sciences, University of Bern, Bern, Switzerland; [7]Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany; [8]School of Biological Sciences, The University of Queensland, Brisbane, Queensland, Australia; [9]Department of Ecology, University of Alicante, San Vicente del Raspeig, Spain; [10]Institute of Biology, Leipzig University, Leipzig, Germany; [11]Department of Experimental Plant Ecology, Institute for Water and Wetland Research, Radboud University, Nijmegen, the Netherlands; [12]Institute of Biodiversity, Friedrich Schiller University Jena, Jena, Germany; [13]Fredericton Research and Development Centre, Fredericton, New Brunswick, Canada and [14]Department of Geography, Zurich University, Zurich, Switzerland

Correspondence
Adam Thomas Clark
Email: adam.tclark@gmail.com

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: FOR456/1451 and RO2397/8; Friedrich Schiller University of Jena; Ramón y Cajal Programme (Spanish Government), Grant/Award Number: RYC-2016-20604; Max-Planck-Gesellschaft

Handling Editor: Caroline Brophy

Abstract

1. Models of natural processes necessarily sacrifice some realism for the sake of tractability. Detailed, parameter-rich models often provide accurate estimates of system behaviour but can be data-hungry and difficult to operationalize. Moreover, complexity increases the danger of 'over-fitting', which leads to poor performance when models are applied to novel conditions. This challenge is typically described in terms of a trade-off between bias and variance (i.e. low accuracy vs. low precision).
2. In studies of ecological communities, this trade-off often leads to an argument about the level of detail needed to describe interactions among species. Here, we used data from a grassland biodiversity experiment containing nine locally abundant plant species (the Jena 'dominance experiment') to parameterize models representing six increasingly complex hypotheses about interactions. For each model, we calculated goodness-of-fit across different subsets of the data based on sown species richness levels, and tested how performance changed depending on whether or not the same data were used to parameterize and test the model (i.e. within vs. out-of-sample), and whether the range of diversity treatments being predicted fell inside or outside of the range used for parameterization.

3. As expected, goodness-of-fit improved as a function of model complexity for all within-sample tests. In contrast, the best out-of-sample performance generally resulted from models of intermediate complexity (i.e. with only two interaction coefficients per species—an intraspecific effect and a single pooled interspecific effect), especially for predictions that fell outside the range of diversity treatments used for parameterization. In accordance with other studies, our results also demonstrate that commonly used selection methods based on AIC of models fitted to the full dataset correspond more closely to within-sample than out-of-sample performance.
4. *Synthesis*. Our results demonstrate that models which include only general intra and interspecific interaction coefficients can be sufficient for estimating species-level abundances across a wide range of contexts and may provide better out-of-sample performance than do more complex models. These findings serve as a reminder that simpler models may often provide a better trade-off between bias and variance in ecological systems, particularly when applying models beyond the conditions used to parameterize them.

KEYWORDS
bias-variance trade-off, cross-validation, Gompertz population model, grasslands, interspecific competition, Jena experiment, over-fitting, plant population and community dynamics

1 | INTRODUCTION

> *“What a useful thing a pocket-map is!" I remarked.*
>
> *"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"*
>
> *"About six inches to the mile."*
>
> *"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"*
>
> *"Have you used it much?" I enquired.*
>
> *"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*

-from Lewis Carroll, Sylvie and Bruno Concluded, Chapter XI, London, 1895.

Ecological communities are complex systems, often with many interacting species and variable environmental conditions. As large datasets are increasingly available and computational methods continue to advance, we can fit more complex and parameter-rich models to describe community dynamics (Evans, Merow, Record, McMahon, & Enquist, 2016; Kearney & Porter, 2009; Perretti, Sugihara, & Munch, 2012). In many cases, these advances represent an exciting opportunity to re-examine old questions and gain new insights into the detailed workings of ecological communities (Grimm, Ayllón, & Railsback, 2016; Grubb, 1992; Judson, 1994). It remains unclear, however, at which point increased model complexity yields additional insights that are generalizable beyond the data used to parameterize them (Allen & Starr, 2017; Coelho, Diniz-Filho, & Rangel, 2018; Evans et al., 2013; Lawton, 1999; Levins, 1968; Schaffer, 1981; Wenger & Olden, 2012). Understanding potential trade-offs between model complexity and generality is therefore increasingly important if we are to make accurate predictions across ecological communities.

As an example, consider how we might predict the abundance of a species based on community interactions within a trophic level. A basic model might only contain a single interaction coefficient: a general term to describe the average effect of all other individuals in the community, regardless of their identity (Hubbell, 2001; May, Huth, & Wiegand, 2015). This model could be made more complex by including two coefficients: one to specify the effect of intraspecific interactions (i.e. self-limitation or self-enhancement), and a second to specify the effect of all interspecific interactions (i.e. competition or facilitation) (Adler et al., 2018; Tuck, Porter, Rees, & Turnbull, 2018). A more comprehensive model might unpack the generalized interspecific term into pairwise interspecific interactions between all species (Carrara, Giometto, Seymour, Rinaldo, & Altermatt, 2015; Fort, 2018; Halty, Valdés, Tejera, Picasso, & Fort, 2017; Vandermeer, 1969). Finally, more complex models might assume that species engage in 'higher-order interactions', for example, where the strength or direction of interactions among a subset of species are modified as a function of the abundance of other species in the community (Letten & Stouffer, 2019; Mayfield & Stouffer, 2017).

Each of these increasingly complex models has some empirical support. For example, neutral models with a single interaction coefficient have adequately explained some patterns in ecological communities (Hubbell, 2001). In other cases, observations are better matched by models with separate intra- versus interspecific terms, reflecting that intraspecific effects often far outweigh interspecific effects (Adler et al., 2018; Broekman et al., in press). Alternatively, in some microbial and plant systems, models that include pairwise interactions among species produce more accurate estimates of species abundances (Carrara et al., 2015; Clark, Lehman, & Tilman, 2018; Vandermeer, 1969), potentially indicating species-specific impacts on resource availability (Tilman, 1982). Lastly, there is evidence from several systems that interaction strengths differ depending on community composition, suggesting the existence of 'higher-order' interactions (Bairey, Kelsic, & Kishony, 2016; Grilli, Barabás, Michalska-Smith, & Allesina, 2017; Levine, Bascompte, Adler, & Allesina, 2017; Mayfield & Stouffer, 2017; Wilbur, 1972). These differences can be mediated either directly, e.g. by changes in species densities or foraging strategies (Letten & Stouffer, 2019; Tilman, 1982; Tuck et al., 2018), or indirectly, e.g. via pathogens or herbivores (Kulmatiski, Beard, Grenzer, Forero, & Heavilin, 2016; Michalet et al., 2015; Weigelt et al., 2007). Nevertheless, the relative predictive abilities of these different models are rarely compared within a single system, let alone between systems, and the degree to which results can be extrapolated beyond the data used to fit them remains unclear.

The decision about whether to include complex species interactions in an ecological model is a particular example of the 'bias-variance trade-off' (Hastie, Tibshirani, & Friedman, 2017). Any given dataset will include both general phenomena that we might expect to see elsewhere, and particularities that occur in that dataset alone (Levins, 1968). When model performance is tested using the same data that were used to parameterize it, increased complexity can always reduce uncertainty (i.e., variance) by tuning the model to match observations. However, when different subsets of data are used to parameterize the model versus to test model performance, increased complexity can also lead to poorer performance for the testing subset. This phenomenon is known as 'over-fitting', and occurs when parameter tuning during the fitting process draws predictions towards these peculiarities (i.e., bias) (Wenger & Olden, 2012).

The degree to which a model is able to capture general versus particular phenomena can be assessed by dividing data into two or more (ideally independent) subsets, and testing whether models parameterized with one subset of data can make good predictions in the others (i.e. 'cross-validation') (Brewer, Butler, & Cooksley, 2016; Roberts et al., 2017; Wenger & Olden, 2012). Due to data scarcity, cross-validation is typically applied to randomly chosen subsets of data (e.g. 'k-fold cross validation') (Roberts et al., 2017). Overfitting is indicated when predictions for the subset of data that was not used for parameterization are substantially worse than those for the subset that was (i.e. when 'out-of-sample' goodness-of-fit is much lower than 'within-sample' goodness-of-fit). Given sufficient data availability, cross-validation can also be used to test predictive power under specific novel conditions (i.e. 'model transferability') (Wenger & Olden, 2012). By carefully choosing different subsets of data, it is possible to rigorously assess the circumstances under which a particular model is likely to provide accurate extrapolations. For example, to test how general a model is across space and environmental conditions, we might fit the model using data from one site, and test how well it works at another.

A more common approach for deciding whether to accept a more complex model is to compare within-sample goodness-of-fit of different models, with a term to penalize increases in model complexity. In ecology, such tests are often applied using AIC, or other similar information criteria (Aho, Derryberry, & Peterson, 2014; Brewer et al., 2016; Burnham, Anderson, & Huyvaert, 2011; Stone, 1977). Note that AIC is conceptually and theoretically related to cross-validation at the limit where a single observation is retained for testing, and the remainder of the data are used to parameterize a model ('leave-one-out cross-validation') (Stone, 1977). Thus, these methods are very efficient from the perspective of data requirements, but are often less useful for assessing model transferability (Brewer et al., 2016; Wenger & Olden, 2012).

Here, we test relationships between model precision and generality using a dataset collected as part of the Jena Experiment ('dominance experiment'; Roscher et al., 2004; Weisser et al., 2017). In this experiment, a diversity gradient was established using nine locally abundant plant species, including replicates of all monocultures and two-species mixtures, and a selection of higher-diversity treatments. In the present analysis, our goal is to characterize the trade-off between good within-sample estimates (low variance) and good out-of-sample predictions (low bias), and to identify models that provide a reasonable compromise between these properties. Because we are interested in community dynamics, we focus on predictions of species-level above-ground biomass (in oven-dried g/m^2, hereafter 'abundance') rather than total biomass summed across species, which is the subject of several other studies of species interactions in biodiversity experiments (Connolly et al., 2011; Kirwan et al., 2009).

We fitted a series of increasingly complex models, and compared their performance using three types of tests. First, to test for general effects of over-fitting, we fit models using data from monocultures, two-species mixtures, and nine-species mixtures. Thus, all communities within the dominance experiment fall within the range of the observed diversity treatments used to fit these models. Second, to identify models that might be suitable for extrapolation to higher-diversity communities, we fit the same series of models using only data from the minimum number of diversity treatments needed for parameterization (i.e. monocultures or two-species mixtures). Thus, higher diversity communities fall outside of the range of conditions used for parameterization. Finally, we compare these results with the ranking of models that would have been obtained had we followed a more typical model selection procedure based on the AIC of models fitted to the full dataset (i.e. all diversity levels). In accordance with the bias-variance trade-off, we expected that the most complex models would yield the best within-sample estimates, especially within the range of diversity levels used to parameterize them, whereas models of lower complexity would provide more generalisable out-of-sample predictions. Furthermore, we expected that AIC would not necessarily be an effective indicator of out-of-sample model performance, due to its strong correspondence to within-sample estimates.

2 | MATERIALS AND METHODS

2.1 | Overview

We designed six models which predicted species abundances in each year as a function of increasingly complex combinations of intra- and interspecific interactions. Details about models and corresponding biological hypotheses are discussed in Section 2.2. We then fit these models to field data, first using plots sown with one, two or nine species and second using plots containing one or two-species. Details about the empirical data used to fit and test models are described in Section 2.3. Next, we used these models to predict species abundance in all plots, and compared within-sample and out-of-sample errors among models. We then tested model performance across sown diversity treatments, species, and years. Based on these tests, we identified models that provided the best out-of-sample predictions across these subsets of data. Details about model comparison methods are described in Section 2.4.

2.2 | Model structure

The six models represent increasingly complex hypotheses about intra- and interspecific interactions. All six were adapted from the same underlying model in Equations (1a)–(1b), and estimate the abundance of each species, in each year, as a function of its own abundance in the previous year (i.e. an autoregressive model), and the abundance of other species in the community in the previous year.

First, the *intra-only* model, Equation (2a), hypothesises that a species is only influenced by its own abundance in the previous year. This model largely serves as a 'baseline' and characterizes the predictive power of intraspecific density dependence and autoregressive processes. Second, the *intra = inter* model, Equation (2b), hypothesises that species are equally influenced by their own abundance and by the abundance of any other species, such that interspecific and intraspecific interactions are equivalent. Third, the *intra + inter* model, Equation (2c), hypothesizes that species are differentially influenced by interspecific and intraspecific interactions, but assumes that the per-capita interspecific effects are identical. Fourth, the *intra*inter* model, Equation (2d), is identical to the *intra + inter* model, except that it hypothesizes that interspecific effects vary depending on the abundance of the focal species (i.e. when a resident species is abundant vs. rare, it experiences differential per-capita effects of interspecific interactions). Fifth, the *pairwise* model, Equation (2e), hypothesizes that each species has distinct interspecific effects on other members of the community. Note that this model is similar to the classical form of Lotka–Volterra competition. Lastly, the *intra*pairwise* model, Equation (2f), is identical to the *pairwise* model, but hypothesizes that the pairwise interspecific effects vary depending on the density of the focal species.

Interspecific and intraspecific interactions can be either negative (net competition) or positive (net facilitation) in all of our models. Interactions between interspecific parameters and the abundance of the focal species (as included in the *intra*inter* and *intra*pairwise* models) represent rough approximations of potential higher-order interactions among species (Tuck et al., 2018). We structured the higher-order interactions in this way because the models can then be parameterized using only data from communities sown with one and two species, making them comparable to the other models in extrapolation tests.

The specific functional forms of these six models are adapted from a discrete-time autoregressive Gompertz model. This model is commonly used in population ecology to characterize species dynamics, and can be fit using time-series data and standard linear regression methods (Ives, Dennis, Cottingham, & Carpenter, 2003; Tredennick, Hooten, & Adler, 2017). In its basic form, this model expresses $A_{i,q}(t)$, abundance of species i in plot q at time t as follows:

$$A_{i,q}(t) = A_{i,q}(t-1) \times \exp\left[\beta_0 + (\beta_1 - 1) \times \log\left(A_{i,q}(t-1)\right)\right], \qquad (1a)$$

where β_0 is the intrinsic growth rate and β_1 is the dependence of this year's abundance on last year's abundance (i.e., density dependence). Log-transforming Equation (1a) yields a simple linear model:

$$\log\left(A_{i,q}(t)\right) = \beta_0 + \beta_1 \times \log\left(A_{i,q}(t-1)\right) \qquad (1b)$$

to which additive effects of additional covariates (in log space), such as abundances of other species, can be incorporated and fitted using standard regression techniques.

Our six regression models characterize the log abundance of species i in plot q at time t, as a function of up to three types of covariates:

intra only

$$\log\left(A_{i,q}(t)+1\right) = \beta_0 + \beta_t + \beta_1 \log\left(A_{i,q}(t-1)+1\right) \qquad (2a)$$

intra = inter

$$\begin{aligned}\log\left(A_{i,q}(t)+1\right) = {} & \beta_0 + \beta_t + \log\left(A_{i,q}(t-1)+1\right) \\ & + \beta_1 \log\left(\Sigma_k\left[A_{k,q}(t-1)+1\right]\right)\end{aligned} \qquad (2b)$$

intra + inter

$$\begin{aligned}\log\left(A_{i,q}(t)+1\right) = {} & \beta_0 + \beta_t + \beta_1 \log\left(A_{i,q}(t-1)+1\right) \\ & + \beta_2 \log\left(\Sigma_{k \neq i}\left[A_{k,q}(t-1)\right]+1\right)\end{aligned} \qquad (2c)$$

intra*inter

pairwise

$$\log\left(A_{i,q}\,(t)+1\right)=\beta_0+\beta_t+\beta_1\log\left(A_{i,q}\,(t-1)+1\right) + \left[\beta_2+\beta_3\log\left(A_{i,q}\,(t-1)+1\right)\right]\log\left(\Sigma_{k\neq i}\left[A_{k,q}\,(t-1)\right]+1\right) \tag{2d}$$

intra*pairwise

$$\log\left(A_{i,q}\,(t)+1\right)=\beta_0+\beta_t+\beta_1\log\left(A_{i,q}\,(t-1)+1\right) + \Sigma_{k\neq i}\left[\beta_{2,k}\log\left(A_{k,q}\,(t-1)+1\right)\right] \tag{2e}$$

$$\log\left(A_{i,q}\,(t)+1\right)=\beta_0+\beta_t+\beta_1\log\left(A_{i,q}\,(t-1)+1\right) + \Sigma_{k\neq i}\left[\left[\beta_{2,k}+\beta_{3,k}\log\left(A_{k,q}\,(t-1)+1\right)\right] \log\left(A_{k,q}\,(t-1)+1\right)\right] \tag{2f}$$

In all of these models, β_0 is the intercept, β_t is a random (categorical) year effect, and β_1 describes intraspecific density dependence. The number of fitted fixed effect parameters per species for each model is therefore: 2 for intra only and intra = inter; 3 for intra + inter; 4 for intra*inter; 10 for pairwise; and 18 for intra*pairwise.

Because we considered abundances for every species in every plot in which they were sown, even if their abundance was zero, we added 1/gm^2 to all abundances before log-transforming. We did not include spatial random effects (i.e. plot, block, or spatial coordinates) for two reasons. First, because treatments were randomly assigned to plots, there was no a priori reason to assume spatial autocorrelation that was related to community composition (when included in models, nested plot and block random effects explained, on average, about 2% of variance). Second, because our subsequent statistical tests are based on cross-validation of total model goodness-of-fit, rather than on significance tests of individual model parameters, adjustments in the degrees of freedom related to the random effects structure did not influence our results.

In Equation (2b), β_1 describes the generalized effect of all intraspecific and interspecific interactions, and log ($A_{i,q}$(t – 1) + 1) is also included as an offset (i.e. without a fitted covariate) so that the function can still be interpreted as a growth rate. Consequently, for models fit using only data from monocultures, fitted values for β_1 are identical to those in Equation (2a) (after adjusting for the offset). In Equation (2c) (*intra + inter*) and Equation (2d) (*intra*inter*), β_2 is the generalized effect of interspecific interactions from all non-focal species combined, whereas β_3 describes how this effect varies as a function of the abundance of the focal species. In Equation (2e) (*pairwise*) and Equation (2f) (*intra*pairwise*), $\beta_{2,k}$ is a vector of species-specific interspecific interaction strengths (e.g. the effect of species *k* on the focal species *i*), $A_{k,q}$ is the observed abundance of species *k* in plot *q*, and $\beta_{3,k}$ describes how $\beta_{2,k}$ changes as a function of the focal species' abundance. Note that the fitted β values differ among species and models.

We fitted all models using the LMER function from the LME4 package (Bates, Mächler, Bolker, & Walker, 2015) in R version 3.4.2 (R. Development Core Team, 2017). We fitted separate regressions for each species and model type. As discussed above, we used two subsets of data to parameterize each model. First, all models were fit using data from all the one, two, and nine species plots (we chose these because they span the full range of diversity treatments, and because all possible species combinations of these diversity levels were represented in the plots). Second, we fit all models using data from communities sown with one and two species, except for the *intra only* and *intra = inter* models, for which we fit regressions using only data from monocultures (i.e. we chose the fewest number of diversity levels required to fit the model, based on the types of species interactions that they hypothesized). We then used these fitted models, following Equations (2a)–(2f), to predict biomass of each species in each community and year (i.e. across all sown richness levels). The full, annotated code used for our analyses is available in the file 'Clark_etal_JE_Dominance.R' in the Supporting Information.

2.3 | Study site

The 'dominance experiment' is part of the Jena Experiment (Roscher et al., 2004; Weisser et al., 2017), which was established on a former agricultural field in the floodplain of the river Saale close to the city of Jena (Germany, 50°55'N, 11°35'E, 130 m) in spring 2002. The region has a mean annual temperature of 9.9°C, and mean annual precipitation is 610 mm (1980–2010; Hoffmann, Bivour, Früh, Koßmann, & Voß, 2014). The dominance experiment is based on a pool of nine grassland species, which often reach high abundances and relative dominance in Central European mesophilic grasslands of the Arrhenatherion type (Ellenberg, 1988). These include five grass species (*Alopecurus pratensis* L., *Arrhenatherum elatius* (L.) J. Presl et C. Presl, *Dactylis glomerata* L., *Phleum pratense* L., *Poa trivialis* L.), two non-legume forb species (*Anthriscus sylvestris* (L.) Hoffm., *Geranium pratense* L.) and two legume species (*Trifolium pratense* L., *T. repens* L.). See Table S1 in the Supporting Information for more information about species, and Table S2 for specific data on treatments. Data are available in Weigelt et al. (2016), and in the *Jena Experiment Information System* (www.the-jena-experiment.de/data). In our study, we used data from the full 13-year duration of the experiment (i.e. 2002–2015).

Sown species richness levels were 1, 2, 3, 4, 6, or 9 species. Each species was equally represented at each richness level, and all possible two-species combinations were present with the same frequency at each richness level of the multi-species mixtures (i.e. 2–9 species). Each combination of species was replicated twice. The design included 2 × 9 monoculture plots, 2 × 36 two-species mixtures (all possible combinations), 2 × 24 three-species mixtures, 2 × 18 four-species mixtures, 2 × 12 six-species mixtures, and 8 nine-species mixtures, resulting in a total of 100 distinct communities (different species compositions) and 206 plots (Roscher et al., 2004). The experiment was set up in four blocks perpendicular to the Saale River following a gradient in soil texture. Mixtures were randomly assigned to the blocks, ensuring that each block contained the same number of plots per species-richness level. Plots were established by sowing with a constant density of 1,000 germinable seeds per m^2 (adjusted for germination rates from standard laboratory tests), which were equally distributed among species in the mixtures. Plots were initially sown across 3.5 × 3.5 m areas, but only maintained in the central 1 × 1 m portion of the plot from 2010 onwards because of the high expense and difficulty of maintaining the larger areas.

The sown species combinations were maintained by weeding all other species not sown into a particular plot (i.e. weeds) two to three times per year. Plots were mowed two times per year (June, September), and mown biomass was removed, as is typical for extensively used meadows in the study region. Plots were not fertilized. Annual above-ground plant biomass production was derived from the sum of two harvests per year, taken at estimated peak biomass (May and August) shortly before mowing. Plant biomass was harvested in two randomly allocated 0.5 × 0.2 m quadrats from 2003 to 2009, and one quadrat (of the same size) in the plot centre after the reduction in plot size in 2010. Harvested biomass was sorted to species after removal of detached dead plant material. Samples were weighed after drying at 70°C for 48 hr. For full species-level dynamics across treatments, see Figure S1 in the Supporting Information.

2.4 | Testing model performance

To quantify the goodness-of-fit of each model, we calculated the second-order 'coefficient of efficiency' (Legates & McCabe, 1999; Li, 2017; Willmott et al., 1985):

$$E_j = 1 - \Sigma_i \left(|O_i - P_i|^j \,/\, |O_i - \text{mean}(O)|^j \right), \tag{3}$$

where j is the order of the coefficient, O_i is observation i, P_i is prediction i, and mean(O) is the average taken across observations. Note that E_2 is similar to the classical R^2 metric, except that it measures scatter around the 1–1 line rather than a fitted regression line. Because E_2 can overweight outliers in some cases (Legates & McCabe, 1999), we also repeated all analyses using E_1, which corresponds to absolute error rather than squared error. These results are almost identical to those for E_2, but are presented in Figures S2–S5 for reference.

After fitting models, we tested for differences in goodness-of-fit by aggregating predictions by three different sets of grouping variables—sown richness level, species identity, and year. We chose these groupings following preliminary analyses, which showed significant between-group variability in model goodness-of-fit (see Tables S3a–b for details). These groupings are also consistent with well-supported hypotheses that species interactions vary as a function of species richness (i.e. higher-order interactions), species identity (i.e. non-neutral interactions), and time (i.e. successional status or temporal feedbacks) (Adler et al., 2018; Deyle, May, Munch, & Sugihara, 2016; Mayfield & Stouffer, 2017; Tuck et al., 2018).

To quantify variation in goodness-of-fit, we applied a simple bootstrapping algorithm. For each grouping described above, we resampled from the full pool of observations and model predictions 20,000 times with replacement, and calculated goodness-of-fit for each iteration. Because of differences in sample size among the grouping levels described above, we performed bootstrapping separately for each of the groupings (i.e. we conducted separate stratified sampling within each level of sown richness, species, or year). We then used the resulting distribution of E_2 and E_1 values observed across iterations to calculate a mean, standard error of the mean, and p-values comparing differences in goodness-of-fit among factors.

Finally, to compare our findings to those that would arise from more classic methods of model comparison, we also calculated AIC for each of the fitted models. For these comparisons, we fitted models to the full set of data from all plots (i.e. plots sown with 1–9 species) and included a random intercept for plot nested within block. Following 'best practices' for AIC comparison of models that include different fixed effects, all models were fitted using maximum likelihood (i.e. rather than REML) (Bates et al., 2015). In cases where the random effects structure led to convergence issues, we removed either the plot, or the plot and block random effects (11/54 and 24/54 cases, respectively). Jointly, these approaches were intended to mirror a more typical statistical analysis of ecological data.

3 | RESULTS

Considered across all diversity levels, species identities, and years, all six model forms did a similar job of estimating average within-sample abundances, albeit with a slightly increased goodness-of-fit for more complex models, especially for predictions that fell within the range of diversity conditions used for parameterization (Figure 1a). For out-of-sample extrapolations outside of the range of diversity conditions used for parameterization (i.e. those parameterized with richness levels of 1–2 species, hereafter '1–2 species models'), models of intermediate complexity (*intra + inter* and *intra***inter*) performed significantly better than the other models (Figure 1b, vertical axis). For out-of-sample predictions made within the range of diversity conditions used for parameterization (i.e. those parameterized with richness levels of 1, 2, and 9 species, hereafter '1–9 species models'), most models performed similarly, except for *intra***pairwise* and *intra = inter*, which performed significantly worse (Figure 1b, horizontal axis). See Tables S4a–k for all regression coefficients, and Tables S5a–h for p-values corresponding to differences in goodness-of-fit among models. For plots of observed versus predicted values for each model and species, see Figs. S6a–b.

Results were similar within individual diversity levels. For 1–2 species models, out-of-sample extrapolations for *intra + inter* and *intra***inter* again had the highest goodness-of-fit, followed closely by *intra only* and *pairwise* (Figure 2c–f, vertical axis). For all models, goodness-of-fit declined somewhat as a function of diversity, but the reductions were steeper for *intra = inter* and *pairwise*, and were especially steep for *intra***pairwise*. Thus, despite the increasing number of potential competitors, models of increasing complexity did not improve out-of-sample in more diverse communities, even for 1–9 species models.

When considering each species individually, goodness-of-fit was also consistently highest for the *intra only*, *intra + inter*, *intra***inter*, and *pairwise* models, whereas *intra = inter* and *intra***pairwise* typically

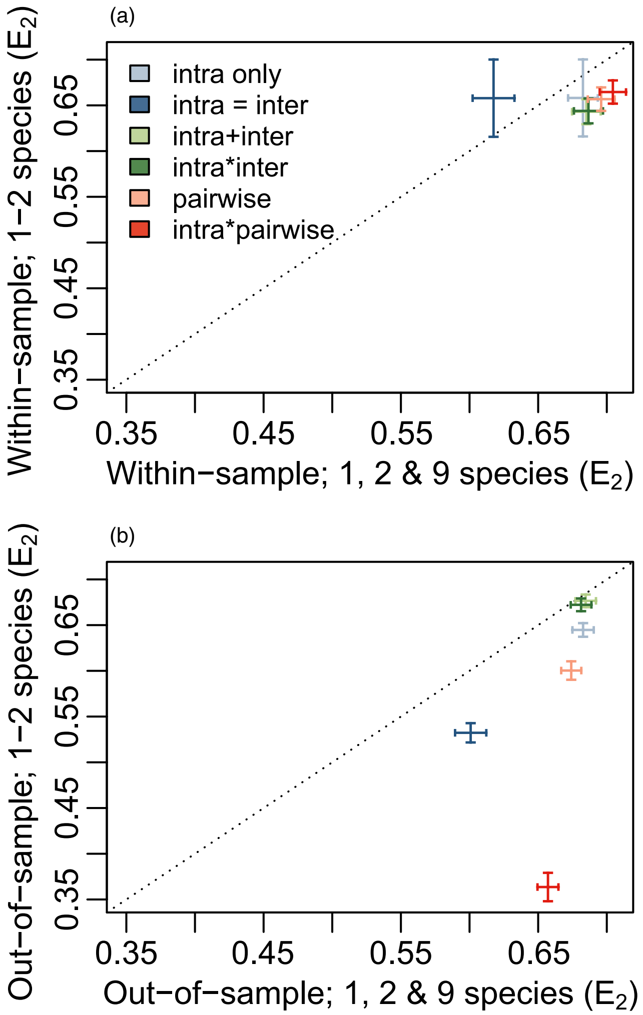

FIGURE 1 Goodness-of-fit for regression models, showing predictive ability when considered across all years, species, and sown richness levels. E_2 describes squared error between observations and predictions, relative to the total sum of squares — see Equation (3) in the main text for details. Models correspond to Equations (2a–2f) in the main text. Panel (a) shows results for data that were used to fit the regression (i.e. 'within-sample'), whereas (b) shows results for data that were not used to fit the regression (i.e. 'out-of-sample'). Vertical axes show results for models that were parameterized with data that spanned the full range of sown diversity levels (i.e. plots sown with 1, 2, and 9 species), whereas horizontal axes show results for models that were parameterized using only data from low-diversity plots (i.e. 1 or 2 species). Dotted line shows 1-1 relationship. Intervals show mean ± one standard error of the mean, based on the bootstrapping routine described in the main text. As a rough rule of thumb, cases where standard errors overlap imply that the two means do not differ significantly at p = .05. See Table S5a in the Supporting Information for p-values summarising differences among model fits [Colour figure can be viewed at wileyonlinelibrary.com]

performed poorly, especially for the 1–2 species models (Figure 3). There were, however, some exceptions to this pattern. All models performed well for *A. pratensis*, and the 1–9 species models provided relatively good predictions for *A. sylvestris*, *D. glomerata*, and *G. pratense*. For both types of parameterization, *intra*pairwise* had especially high goodness-of-fit for *A. elatius*, and for 1–9 species models it also provided good predictions for *P. trivialis*, *T. repens*, and *T. pratense*. Goodness-of-fit for *intra = inter* was especially high for *P. pratense*, and for 1–2 species models of *T. repens*.

Across years, there was a strong decline in out-of-sample predictive ability around 2007 (Figure 4), corresponding to a major decline in legume abundance (especially for *T. pratense*—see Figure S1). However, predictive power recovered rapidly, and by 2009 was roughly equal to that before 2007. All 1–9 species models performed similarly, except *intra = inter* for which goodness-of-fit was slightly lower across all years. For 1–2 species models, all followed the same general trend, but *intra = inter*, *pairwise*, and *intra*pairwise* had particularly low goodness-of-fit, especially around 2006–2008. To account for the potential influence of these outlier years on our analyses, we repeated all of our model fitting and comparisons with years 2006–2008 omitted from the dataset, but found no major changes in our results other than a slight increase in goodness-of-fit for models of the two legumes (not shown).

For our more typical analysis based on the AIC of models fitted to the full dataset, we found that the index typically identified models of higher complexity as the 'best fitting' (Table 1). When compared across all species, the *intra*pairwise* model had the lowest AIC, which differed from the next best model (*pairwise*) by more than 150 units. Likewise, at the level of individual species, AIC never indicated that models of low complexity provided the best fit, and selected *intra + inter* as the best model for one species (*T. pratense*), *pairwise* for four species (*A. pratensis*, *A. elatius*, *G. pratense*, and *P. trivialis*), and *intra*pairwise* for four species (*A. sylvestris*, *D. glomerata*, *P. pratense*, and *T. repens*).

4 | DISCUSSION

Our primary result indicates that models of intermediate complexity—which only separate the effects of intra- and interspecific competition—can provide the best extrapolative predictions of species-level abundances in both simple and higher-diversity communities. In particular, our findings support the hypothesis that models of intermediate complexity should usually produce the best out-of-sample predictions, especially for extrapolations (Allen & Starr, 2017; Evans et al., 2013; Wenger & Olden, 2012). In accordance with other studies, we also find that classical methods of AIC comparison, based on the full dataset, selected more complex models than those selected using out-of-sample extrapolations from the models that were fitted using only data from plots sown with 1–2 species. Taken together, these results demonstrate that simple models can yield broadly generalizable predictions in complex systems, but also that commonly used methods of model selection may not always be effective at identifying these models.

In general, we can divide our models into three groups. First, those treating intraspecific and interspecific interactions as equivalent (*intra = inter*) almost always yielded poor out-of-sample predictions. In accordance with theory and a recent meta-analysis, this result suggests that differences between intraspecific and

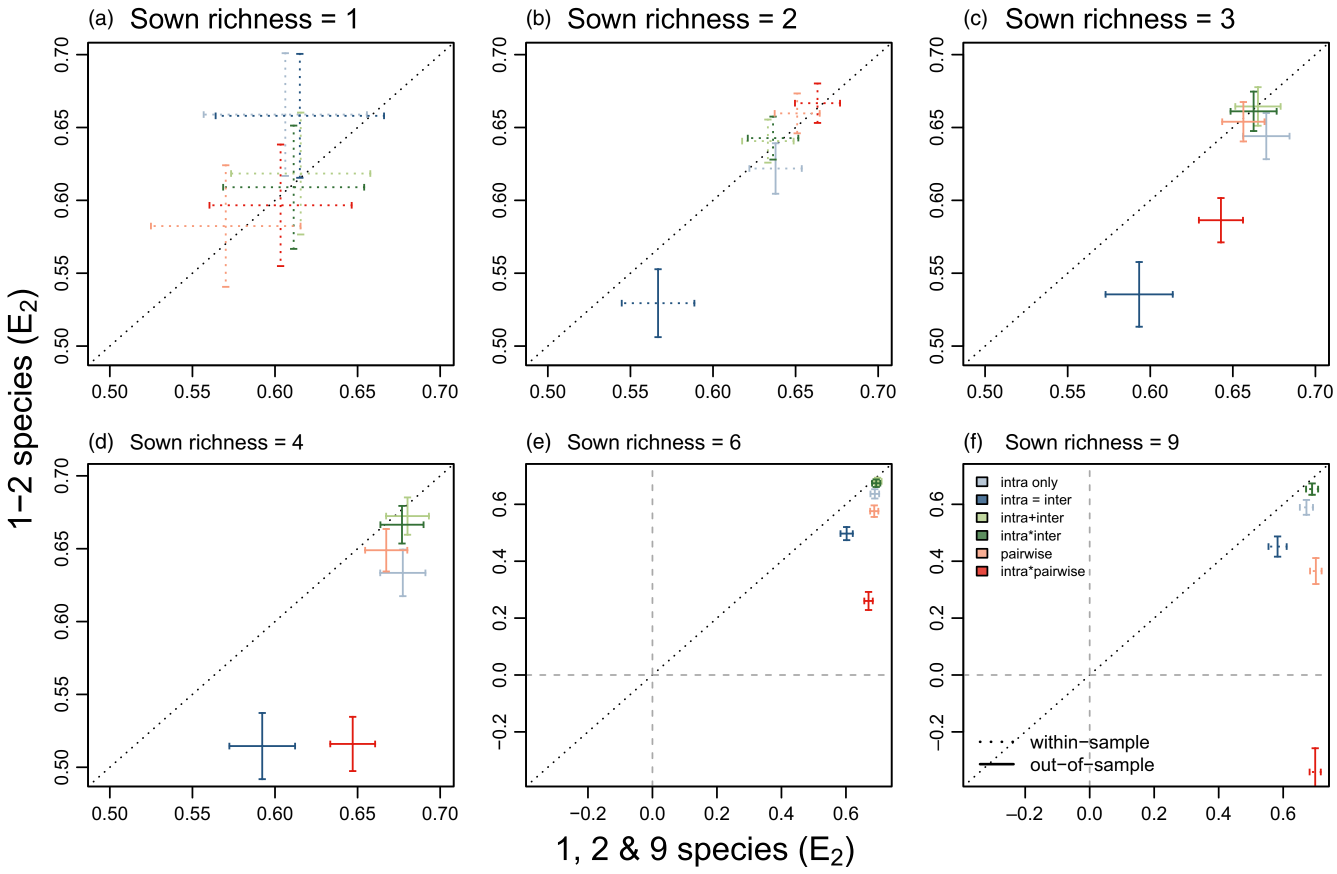

FIGURE 2 Goodness-of-fit for regression models at each sown richness level, calculated across all species and years. Dashed intervals show estimates for data that were used to fit the regression (i.e. 'within-sample'), whereas solid lines show fits for data that were not used to fit the regression (i.e. 'out-of-sample'). Intervals, axis labels, and indices are as described in the legend to Figure 1. See Table S5b in the Supporting Information for *p*-values summarizing differences among model fits. Note that the axes are expanded for 6 and 9 species mixtures [Colour figure can be viewed at wileyonlinelibrary.com]

interspecific interactions are important for explaining species-level abundances (Adler et al., 2018; Broekman et al., in press). Second, models that allowed for differential effects of interspecific and intraspecific interactions (*intra + inter* and *intra*inter*) provided accurate predictions across most combinations of richness treatments, species, and years. Their strong out-of-sample goodness-of-fit suggests that these models contain sufficient complexity to explain broad trends, but not enough to cause over-fitting (Wenger & Olden, 2012). Finally, models with pairwise interactions among species (*pairwise* and *intra*pairwise*) appear, at least in some cases, to be over-fit. In particular, when considered across out-of-sample extrapolations from the 1–2 species models, *pairwise* and *intra*pairwise* performed significantly worse than all other tested models except *intra = inter*, and performance was especially poor for extrapolations early in the experiment, and in diverse mixtures. Jointly, these results suggest that the added complexity in the pairwise models may be more reflective of peculiarities in the training dataset than of general phenomena (Hastie et al., 2017).

Critically, our findings should not be taken to suggest that some models are more *correct* than others, but rather that they are more *practical from the perspective of prediction* given our system and available data. An important caveat for all of our models is that they do not consider any specific mechanisms of competition (e.g. allelopathy, shared resources). Thus, the poorer out-of-sample and extrapolative performance that we find for complex models may be indicative of important omitted mechanisms, rather than of the unimportance of species-specific interactions. For example, in mechanistic resource competition models, pairwise interaction strengths can change dramatically as a result of small shifts in community composition (Letten & Stouffer, 2019; Tilman, 1982). These changes would not be captured by any of the models that we test here, and it may be that simple phenomenological models are better able to average across these different conditions than are more parameter-rich phenomenological models.

4.1 | Biological interpretation of models

The baseline model, which included only intraspecific interactions (*intra only*), often performed similarly to the models with both intraspecific and interspecific interactions (*intra + inter* and *intra*inter*), however, predictions were usually significantly worse for *intra only*, especially for extrapolations from the 1 to 2 species models. The strong performance of the *intra-only* model indicates that species

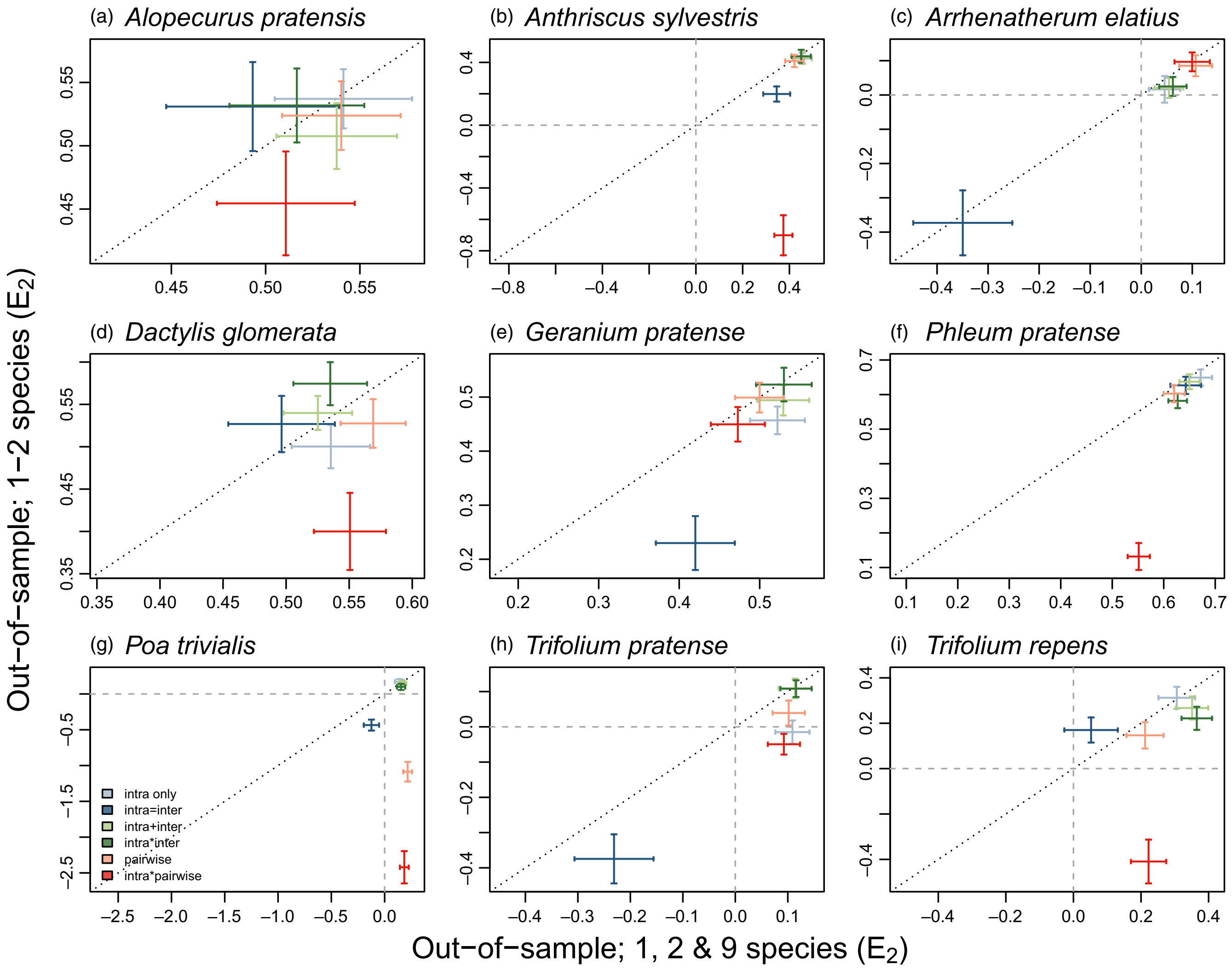

FIGURE 3 Goodness-of-fit for fitted regression models of each species, calculated across all years and sown richness levels. Results are shown only for out-of-sample goodness-of-fit. Intervals, axis labels, and indices are as described in the legend to Figure 1. See methods, or Table S1 in the appendix, for further details about species. See Table S5c in the Supporting Information for *p*-values summarizing differences among model fits. Because values vary greatly among species, note that axes are on different scales in each panel [Colour figure can be viewed at wileyonlinelibrary.com]

abundances were strongly regulated by density dependence and temporal autocorrelation, such that observed species abundances were largely explained by their abundance in the previous time step (Petchey et al., 2015). Note, however, that intraspecific density dependence and autocorrelation alone were not sufficient to guarantee accurate out-of-sample predictions. For example, the *intra = inter* also included these components, but performed poorly, indicating that differences between intraspecific and interspecific interactions play an important role in determining abundance in our system.

For the two intermediately complex models (*intra + inter* and *intra*inter*), interspecific interaction coefficients were typically negative, which accords with expectations of strong net competition among communities of locally dominant species in mesic grasslands (see Tables S4a–k) (Weigelt et al., 2007). Similarly, the effects of intraspecific interactions were predicted to be about 5–10 times stronger than interspecific effects, which is consistent with theoretical and empirical results from many ecological systems (Adler et al., 2018; Barabás, Michalska-Smith, & Allesina, 2017; Tuck et al., 2018). The more complex model (*intra*inter*) generally indicated significant reductions in the strength of interspecific competition with increasing focal species abundance, which suggests that competition is asymmetric and size-mediated, as might be expected, for example, light competition (Schwinning & Weiner, 1998). However, given that *intra*inter* rarely performed significantly better than *intra + inter*, it appears that any effects of these size-mediated biological mechanisms were either not very influential, or could be abstracted into the generic interspecific terms in the simpler model.

For the two pairwise models (*pairwise* and *intra*pairwise*), the decline in out-of-sample extrapolative performance early in the experiment and at higher diversity levels could indicate that species interactions vary in ways that we do not account for in Equation (2e–2f)—for example, changes in pairwise competitive interactions over the course of community assembly, or as a function of diversity (Mayfield & Stouffer, 2017; Tilman, 1982). If complex higher-order

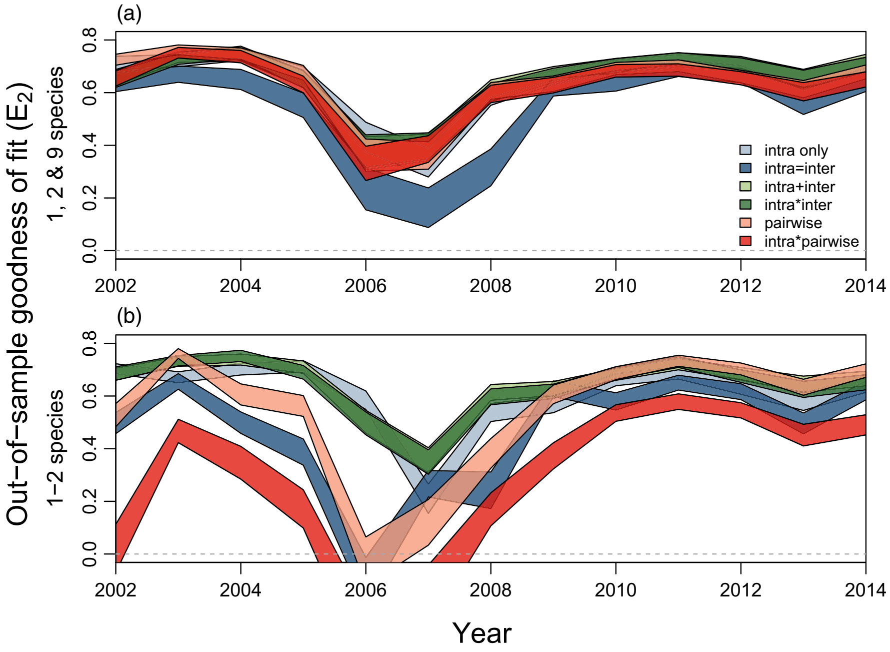

FIGURE 4 Goodness-of-fit of fitted regression models for each year, calculated across all species and sown richness levels. Results are shown only for out-of-sample goodness-of-fit. Intervals show mean ± one standard error of the mean. Axis labels and indices are as described in the legend to Figure 1. Note that the *intra + inter* and *intra*inter* trajectories are overlapping. See Table S5d in the Supporting Information for *p*-values summarising differences among model fits [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 ΔAIC values for the six models presented in the main text. Columns show different, models, and rows show whether AIC was calculated across all species, sown diversity treatments, and years ('Total'), or separated by species. Recall that lower AIC values indicate better performance. Bold numbers mark cases where ΔAIC < 2, which is commonly used as a rule of thumb to identify meaningful differences in model performance. Regressions are fit to all observations across all diversity treatments (i.e. 1–9 sown species richness treatments). See main text for details about model fitting and random effects structure

	intra only	intra = inter	intra + inter	intra*inter	pairwise	intra*pairwise
Total	748.46	2,354.15	484.01	406.83	154.9	**0**
Alopecurus pratensis	58.58	96.22	13.15	3.01	**0**	7.75
Anthriscus sylvestris	66.13	204.25	52.18	34.75	2.81	**0**
Arrhenatheum elatius	47.49	344.6	23.34	25.09	**0**	**1.5**
Dactylis glomerata	102.65	153.5	63.55	48.13	**1.01**	**0**
Geranium pratense	63.12	209.77	50.56	39.93	**0**	7.39
Phleum pratense	175	367.42	143.05	120.95	107.22	**0**
Poa trivialis	102.77	351.98	66.27	75.66	**0**	5.97
Trifolium pratense	11.27	270.47	**0**	**1.89**	12.76	17.4
T. repens	161.46	395.96	111.93	97.45	71.12	**0**

processes are an important determinant of community dynamics in our system, then accurate extrapolations from complex models may require that we consider a more diverse—and potentially more mechanistic—set of models (Letten & Stouffer, 2019). However, such a result may make generalizing findings to other systems difficult, especially if they differ in the kinds of higher-order processes that are locally important. Alternatively, the relatively good performance of the 1–2 species pairwise models could suggest that the extrapolative models were simply not fitted to enough data, in which case any apparent 'over-fitting' might be alleviated by increasing the number of experimental replicates. It would be interesting to test this possibility with larger datasets.

Although many models performed well under some circumstances, none provided significantly better out-of-sample predictions than *intra + inter* and *intra*inter*, neither for 1–2 nor for 1–9 species models (with one exception—see Figure 3c). This result accords with a recent meta-analysis of competition experiments, which found that average intraspecific interaction strengths observed across species can be a good proxy for unknown interactions (Fort, 2018). These findings imply that, at least on average, the relative strength of intraspecific versus interspecific interactions is a stronger determinant of species abundances than are individual pairwise effects. Mechanisms that act directly on density dependence, such as species-specific pathogens, may therefore be more important for maintaining diversity in our system than are mechanisms related to pairwise competitive abilities, such as competition for a small number of limiting resources (Hubbell, 2001, 2006), which accords with existing hypotheses about coexistence in the Jena experiment (Curtois et al. 2016; Weisser et al., 2017).

4.2 | Species-level results

Although average predictive accuracy varied greatly across species, the relative performance of the six models usually matched the general order observed for the total dataset. One exception to this pattern was *Arrhenatherum elatius*, for which goodness-of-fit was low for all models, but with significantly higher goodness-of-fit for the pairwise models. Because its abundance was consistently high, and varied little across plots, treatments, or years, the 'null expectation' that we used to predict goodness-of-fit in Equation (3)—that is, mean observed biomass—performed particularly well, which left relatively little room for improvement. A potential explanation for the consistently high abundance of *A. elatius* is that it is an especially effective competitor for light and nitrogen: it is the tallest species in the dominance experiment, invests a larger fraction of biomass in supporting tissues (Lorentzen, Roscher, Schumacher, Schulze, & Schmid, 2008; Roscher, Schumacher, Weisser, Schmid, & Schulze, 2007), and produces the most biomass per unit nitrogen (Roscher, Thein, Schmid, & Scherer-Lorenzen, 2008). Thus, *A. elatius* may be buffered from the effects of interspecific competition, making models of generic interspecific interactions less effective.

Three other species also have tall stature and produce high biomass per unit nitrogen—*Alopecurus pratensis*, *Dactylis glomerata*, and *Phleum pratense* (Lorentzen et al., 2008; Roscher et al., 2008). Goodness-of-fit for these species was high for all models, except for some poor extrapolations from the 1–2 species *intra*pairwise* model. As with *A. elatius*, these species generally had high abundance, and dynamics were consistent across plots and diversity treatments (Figure S1) potentially because their traits shielded them from competitive interactions with most other species. However, unlike *A. elatius*, their abundances varied over time, such that mean observed biomass alone was not a good predictor. Thus, high goodness-of-fit may indicate that their dynamics were sufficiently complex to differ from the null expectation, but not so complex as to confound our models. That said, high abundance might actually contribute to the strong predictive power of these models by reducing the impact of observation error on predictive outcomes, or increasing temporal autocorrelation (Hastie et al., 2017).

Goodness-of-fit was also relatively high across models for two species that did not have traits associated with strong light or nitrogen competitive ability—*Anthriscus sylvestris* and *Geranium pratense* (Lorentzen et al., 2008; Roscher et al., 2008). Potentially because of poorer competitive abilities, both species were slow to establish, and abundances varied greatly among plots and treatments (Figure S1) (Lorentzen et al., 2008). Although strong effects of interspecific competition may explain the relatively poor performance of *intra = inter*, the high goodness-of-fit for *intra only* suggests that within-plot autocorrelation alone was sufficient for predicting species dynamics.

Lastly, for three species—*Poa trivialis*, *Trifolium pratense*, and *T. repens*—goodness-of-fit was relatively low for all models. A partial explanation is that over a short period of about 2006–2008 there were especially large declines in abundance for *T. pratense* and *T. repens*, and increases in abundance for *P. trivialis* (see Figure S1). When these years were removed from the dataset, goodness-of-fit increased for the two legumes, but not for *P. trivialis*. Curiously, this time period did not correspond to any major weather events or changes in experiment management. Instead, it is possible that the initially sown cohort of shorter-lived perennial species (e.g. *T. pratense* and *T. repens*) began to die back around this time and failed to establish successful reproductive populations, leading to rapid changes in species relative abundance (Roeder, Schweingruber, Fischer, & Roscher, 2017; Roscher et al., 2011).

4.3 | Methodological implications of results

There have been many rallying calls for better, and more general, predictive models in ecology (Coelho et al., 2018; Dietze et al., 2018; Evans et al., 2013; Houlahan, McKinney, Anderson, & McGill, 2017; Lawton, 1999; Wenger & Olden, 2012). Not all of these favour reductions in model complexity. In particular, the advent of individual-based models, which can be formulated around rules and behaviours, has been suggested as a potential way to simultaneously satisfy the needs for nuance and generality (Allen & Starr, 2017; Evans et al., 2013; Grimm et al., 2016; Judson, 1994). Nevertheless, to our knowledge, the majority of studies that have empirically tested model performance across a gradient of complexity accord with our findings—that is, that intermediately complex models provide the best extrapolations—including models of bacterial growth (Buchanan, Whiting, & Damert, 1997), fish populations (Wenger & Olden, 2012), and plant communities (Petitpierre, Broennimann, Kueffer, Daehler, & Guisan, 2017; Rüger, Wirth, Wright, & Condit, 2012).

Despite the relatively broad support for using extrapolative ability as an indicator of model generality, most ecologists still conduct model selection, often by fitting models to their full dataset and applying various information criteria (e.g. AIC, AICc, WAIC, BIC, etc.; Burnham et al., 2011; Aho et al., 2014; Brewer et al., 2016; Houlahan et al., 2017; Coelho et al., 2018). While there is nothing inherently wrong with this approach, it is important to remember that these tests are primarily designed to identify models that perform well within the general range of conditions used for parameterization, rather than to estimate how models are likely to perform under novel conditions (Brewer et al., 2016; Stone, 1977). Nevertheless, assuming that such models will also yield good extrapolations can be problematic. Because most ecological systems are inherently complex and interconnected, it is likely that given enough data, significant interactions can be detected among any ecological variables, at least via indirect routes (Levin, 1998; Sugihara et al., 2012). But, much of this complexity is likely to be non-transferable across systems (Lawton, 1999). Hence, if our goal is to obtain generality, we should not necessarily emphasize the particular, but rather should focus on identifying models that perform well across a wide range of contexts (Wenger & Olden, 2012). For example, simpler models associated with AIC values that are > 2 units higher than the best models are unlikely to be explored or considered further, yet they may well provide similar within-sample goodness of fit, and better out-of-sample predictions and extrapolations.

Similarly, it is rare to see a thorough analyses of the real impact of higher-order parameters. Figure 1a illustrates this point: including additional terms does indeed improve the within-sample estimates from the 1–9 species models, but not by much. A more detailed analysis of how well simpler models perform—and where, and by how much, they actually get things wrong—is therefore probably more informative than simply accepting some subset of models based on their goodness-of-fit within a single set of conditions (Brewer et al., 2016; Mayfield & Stouffer, 2017; Wenger & Olden, 2012). We do not mean to imply that complexity will *never* be needed to understand particular aspects of ecological systems. Rather, it may be wise to address complexity judiciously, only after we have convinced ourselves that simpler models fail to provide sufficient precision or generality.

5 | CONCLUSIONS

Because of their complexity, ecological systems will always pose a particular challenge for modellers. Though complexity and nuance may well be important for explaining the abundance and distribution of species in many sites and systems, our results demonstrate that in some cases, simpler models can provide more general predictions. More broadly, our analysis serves as a reminder that the best way to test the generalizability of a fitted model is to use it to make predictions in a new context—and that commonly used model selection tools are not designed to test this property.

ACKNOWLEDGEMENTS

This paper arose as part of the 'BEF-Coexist' workshop, organized by Y.F., C.R., and W.S.H., with funding from the German Research Foundation (RO2397/8). A.T.C. was supported by a 'catalyst' postdoctoral fellowship from sDiv. S.S. was supported by the Spanish Government under a Ramón y Cajal contract (RYC-2016- 20604). We thank C. de Mazancourt, M. Fischer, H. Hillebrand, S. Meyer, and R. Richter for insightful comments on earlier versions of this manuscript, and are grateful to the associate editor Caroline Brophy and two anonymous reviewers for helpful feedback. Funding and support for the Jena Experiment were provided by the German Research Foundation (FOR456/1451), Friedrich Schiller University of Jena, and Max Planck Society. We are grateful to the gardeners, students, and staff who have worked to maintain the Jena Experiment, and to Anne Ebeling for coordinating maintenance and data collection.

AUTHORS' CONTRIBUTIONS

This paper arose as part of the 'BEF-Coexist' workshop, co-led by Y.F., C.R. and W.S.H., and attended by A.T.C., L.A.T., A.T., E.A., M.M.M., S.S., K.B., H.K., B.R., C.W., and B.S. A.T.C., A.T., and L.A.T. conceived of the theoretical aspects of the study. A.T.C. and L.A.T. wrote the first draft of the manuscript with input from A.T., E.A., W.S.H., M.M.M., and S.S. A.T.C. and A.T. developed the first version of the models. A.T.C., A.T., L.A.T., E.A., W.S.H., M.M.M., S.S., B.R. and C.W. helped revise and further develop the models. C.R., A.W. and C.W. contributed biomass data. All authors contributed to revisions.

DATA AVAILABILITY STATEMENT

All data are publicly available through the PANGEA database (https://doi.org/10.1594/PANGAEA.866358), and can also be accessed through the *Jenaexperiment Information System* (www.the-jena-experiment.de/data).

ORCID

Adam Thomas Clark https://orcid.org/0000-0002-8843-3278
Andrew Tredennick https://orcid.org/0000-0003-1254-3339
W. Stanley Harpole https://orcid.org/0000-0002-3404-9174
Margaret M. Mayfield https://orcid.org/0000-0002-5101-6542
Kathryn Barry https://orcid.org/0000-0001-6893-6479
Benjamin Rosenbaum https://orcid.org/0000-0002-2815-0874
Cameron Wagg https://orcid.org/0000-0002-9738-6901
Yanhao Feng https://orcid.org/0000-0003-0460-4883
Christiane Roscher https://orcid.org/0000-0001-9301-7909
Bernhard Schmid https://orcid.org/0000-0002-8430-3214

REFERENCES

Adler, P. B., Smull, D., Beard, K. H., Choi, R. T., Furniss, T., Kulmatiski, A., … Veblen, K. E. (2018). Competition and coexistence in plant communities: Intraspecific competition is stronger than interspecific competition. *Ecology Letters*, *21*, 1319–1329. https://doi.org/10.1111/ele.13098

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*, 631–636. https://doi.org/10.1890/13-1452.1

Allen, T. F. H., & Starr, T. B. (2017). *Hierarchy: Perspectives for ecological complexity*. Chicago, IL: University of Chicago Press.

Bairey, E., Kelsic, E. D., & Kishony, R. (2016). High-order species interactions shape ecosystem diversity. *Nature Communications*, *7*, 12285. https://doi.org/10.1038/ncomms12285

Barabás, G., Michalska-Smith, M. J., & Allesina, S. (2017). Self-regulation and the stability of large ecological networks. *Nature Ecology & Evolution*, *1*, 1870–1875. https://doi.org/10.1038/s41559-017-0357-6

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICc, and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*, 679–692.

Broekman, M. J. E., Muller-Landau, H. C., Visser, M. D., Jongejans, E., Wright, S. J., & de Kroon, H. (2019). Signs of stabilization and stable coexistence. *Ecology Letters*, *22*, 1957–1975.

Buchanan, R. L., Whiting, R. C., & Damert, W. C. (1997). When is simple good enough: A comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves. *Food Microbiology*, *14*, 313–326. https://doi.org/10.1006/fmic.1997.0125

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*, 23–35. https://doi.org/10.1007/s00265-010-1029-6

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*, 895–906. https://doi.org/10.1111/2041-210X.12363

Clark, A. T., Lehman, C., & Tilman, D. (2018). Identifying mechanisms that structure ecological communities by snapping model parameters to empirically observed tradeoffs. *Ecology Letters*, *21*, 494–505. https://doi.org/10.1111/ele.12910

Coelho, M. T. P., Diniz-Filho, J. A., & Rangel, T. F. (2018). A parsimonious view of the parsimony principle in ecology and evolution. *Ecography*, *42*, 968–976.

Connolly, J., Cadotte, M. W., Brophy, C., Dooley, Á., Finn, J., Kirwan, L., … Weigelt, A. (2011). Phylogenetically diverse grasslands are associated with pairwise interspecific processes that increase biomass. *Ecology*, *92*, 1385–1392. https://doi.org/10.1890/10-2270.1

Cortois, R., Schröder-Georgi, T., Weigelt, A., van der Putten, W. H., & De Deyn, G. B. (2016). Plant-soil feedbacks: Role of plant functional group and plant traits. *Journal of Ecology*, *104*, 1608–1617. https://doi.org/10.1111/1365-2745.12643

Deyle, E. R., May, R. M., Munch, S. B., & Sugihara, G. (2016). Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20152258.

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., … White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 1424–1432. https://doi.org/10.1073/pnas.1710231115

Ellenberg, H. (1988). *Vegetation ecology of Central Europe*. Cambridge, UK: Cambridge University Press.

Evans, M. R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C. M., … Benton, T. G. (2013). Do simple models lead to generality in ecology? *Trends in Ecology & Evolution*, *28*, 578–583. https://doi.org/10.1016/j.tree.2013.05.022

Evans, M. E. K., Merow, C., Record, S., McMahon, S. M., & Enquist, B. J. (2016). Towards process-based range modeling of many species. *Trends in Ecology & Evolution*, *31*, 860–871. https://doi.org/10.1016/j.tree.2016.08.005

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. https://doi.org/10.1016/j.ecolmodel.2018.09.009

Grilli, J., Barabás, G., Michalska-Smith, M. J., & Allesina, S. (2017). Higher-order interactions stabilize dynamics in competitive network models. *Nature*, *548*, 210–213. https://doi.org/10.1038/nature23273

Grimm, V., Ayllón, D., & Railsback, S. F. (2016). Next-generation individual-based models integrate biodiversity and ecosystems: yes we can, and yes we must. *Ecosystems*, *20*, 229–236. https://doi.org/10.1007/s10021-016-0071-2

Grubb, P. J. (1992). Presidential address: A positive distrust in simplicity – lessons from plant defences and from competition among plants and among animals. *Journal of Ecology*, *80*, 585–610. https://doi.org/10.2307/2260852

Halty, V., Valdés, M., Tejera, M., Picasso, V., & Fort, H. (2017). Modeling plant interspecific interactions from experiments with perennial crop mixtures to predict optimal combinations. *Ecological Applications*, *27*, 2277–2289. https://doi.org/10.1002/eap.1605

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd edition). Springer series in statistics. New York, NY: Springer.

Hoffmann, K., Bivour, W., Früh, B., Koßmann, M., & Voß, P.-H. (2014). Klimauntersuchungen in Jena für die Anpassung an den Klimawandel und seine erwarteten Folgen. Deutscher Wetterdienst: Berichte des Deutschen Wetterdienstes.

Houlahan, J. E., McKinney, S. T., Anderson, T. M., & McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*, 1–7. https://doi.org/10.1111/oik.03726

Hubbell, S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton Monographs in Population Biology. Princeton, NJ: Princeton University Press.

Hubbell, S. (2006). Neutral theory and the evolution of ecological equivalence. *Ecology*, *87*, 1387–1398. https://doi.org/10.1890/0012-9658(2006)87[1387:NTATEO]2.0.CO;2

Ives, A. R., Dennis, B., Cottingham, K. L., & Carpenter, S. R. (2003). Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, *73*, 301–330. https://doi.org/10.1890/0012-9615(2003)073[0301:ECSAEI]2.0.CO;2

Judson, O. P. (1994). The rise of the individual-based model in ecology. *Trends in Ecology & Evolution*, *9*, 9–14. https://doi.org/10.1016/0169-5347(94)90225-9

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*, 334–350. https://doi.org/10.1111/j.1461-0248.2008.01277.x

Kirwan, L., Connolly, J., Finn, J. A., Brophy, C., Lüscher, A., Nyfeler, D., & Sebastià, M.-T. (2009). Diversity–interaction modeling: Estimating contributions of species identities and interactions to ecosystem function. *Ecology*, *90*, 2032–2038. https://doi.org/10.1890/08-1684.1

Kulmatiski, A., Beard, K. H., Grenzer, J., Forero, L., & Heavilin, J. (2016). Using plant-soil feedbacks to predict plant biomass in diverse communities. *Ecology*, *97*, 2064–2073. https://doi.org/10.1890/15-2037.1

Lawton, J. H. (1999). Are there general laws in ecology? *Oikos*, *84*, 177. https://doi.org/10.2307/3546712

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*, 233–241. https://doi.org/10.1029/1998WR900018

Letten, A. D., & Stouffer, D. B. (2019). The mechanistic basis for higher-order interactions and non-additivity in competitive communities. *Ecology Letters*, *22*, 423–436. https://doi.org/10.1111/ele.13211

Levin, S. A. (1998). Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, *1*, 431–436. https://doi.org/10.1007/s100219900037

Levine, J. M., Bascompte, J., Adler, P. B., & Allesina, S. (2017). Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, *546*, 56–64. https://doi.org/10.1038/nature22898

Levins, R. A. (1968). *Evolution in changing environments: Some theoretical explorations*. Monographs in population biology. Princeton, NJ: Princeton University Press.

Li, J. (2017). Assessing the accuracy of predictive models for numerical data. *PLoS ONE*, *12*, e0183250.

Lorentzen, S., Roscher, C., Schumacher, J., Schulze, E.-D., & Schmid, B. (2008). Species richness and identity affect the use of aboveground space in experimental grasslands. *Perspectives in Plant Ecology, Evolution and Systematics*, *10*, 73–87. https://doi.org/10.1016/j.ppees.2007.12.001

May, F., Huth, A., & Wiegand, T. (2015). Moving beyond abundance distributions: Neutral theory and spatial patterns in a tropical forest. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20141657–20141657.

Mayfield, M. M., & Stouffer, D. B. (2017). Higher-order interactions capture unexplained complexity in diverse communities. *Nature Ecology & Evolution*, *1*, 0062. https://doi.org/10.1038/s41559-016-0062

Michalet, R., Chen, S.-Y., An, L.-Z., Wang, X.-T., Wang, Y.-X., Guo, P., … Xiao, S. A. (2015). Communities: Are they groups of hidden interactions? *Journal of Vegetation Science*, *26*, 207–218. https://doi.org/10.1111/jvs.12226

Perretti, C. T., Sugihara, G., & Munch, S. B. (2012). Nonparametric forecasting outperforms parametric methods for a simulated multi-species system. *Ecology*, *94*, 794–800.

Petchey, O. L., Pontarp, M., Massie, T. M., Kéfi, S., Ozgul, A., Weilenmann, M., … Pearse, I. S. (2015). The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters*, *18*, 597–611. https://doi.org/10.1111/ele.12443

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions:

Which predictors increase the transferability of SDMs? *Global Ecology and Biogeography*, *26*, 275–287. https://doi.org/10.1111/geb.12530

R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Roeder, A., Schweingruber, F. H., Fischer, M., & Roscher, C. (2017). Growth ring analysis of multiple dicotyledonous herb species—A novel community-wide approach. *Basic and Applied Ecology*, *21*, 23–33. https://doi.org/10.1016/j.baae.2017.05.001

Roscher, C., Schumacher, J., Baade, J., Wilcke, W., Gleixner, G., Weisser, W. W., … Schulze, E.-D. (2004). The role of biodiversity for element cycling and trophic interactions: An experimental approach in a grassland community. *Basic and Applied Ecology*, *5*, 107–121. https://doi.org/10.1078/1439-1791-00216

Roscher, C., Schumacher, J., Weisser, W. W., Schmid, B., & Schulze, E.-D. (2007). Detecting the role of individual species for overyielding in experimental grassland communities composed of potentially dominant species. *Oecologia*, *154*, 535–549. https://doi.org/10.1007/s00442-007-0846-4

Roscher, C., Thein, S., Schmid, B., & Scherer-Lorenzen, M. (2008). Complementary nitrogen use among potentially dominant species in a biodiversity experiment varies between two years: Complementary nitrogen use. *Journal of Ecology*, *96*, 477–488. https://doi.org/10.1111/j.1365-2745.2008.01353.x

Roscher, C., Thein, S., Weigelt, A., Temperton, V. M., Buchmann, N., & Schulze, E.-D. (2011). N2 fixation and performance of 12 legume species in a 6-year grassland biodiversity experiment. *Plant and Soil*, *341*, 333–348. https://doi.org/10.1007/s11104-010-0647-0

Rüger, N., Wirth, C., Wright, S. J., & Condit, R. (2012). Functional traits explain light and size response of growth rates in tropical tree species. *Ecology*, *93*, 2626–2636. https://doi.org/10.1890/12-0622.1

Schaffer, W. M. (1981). Ecological abstraction: The consequences of reduced dimensionality in ecological models. *Ecological Monographs*, *51*, 383–401. https://doi.org/10.2307/2937321

Schwinning, S., & Weiner, J. (1998). Mechanisms determining the degree of size asymmetry in competition among plants. *Oecologia*, *113*, 447–455. https://doi.org/10.1007/s004420050397

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society B*, *39*, 44–47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*, 496–500. https://doi.org/10.1126/science.1227079

Tilman, D. (1982). A comparison with classical theory. In *Resource competition and community structure* (pp. 190–204). Princeton, NJ: Princeton University Press.

Tredennick, A. T., Hooten, M. B., & Adler, P. B. (2017). Do we need demographic data to forecast plant population dynamics? *Methods in Ecology and Evolution*, *8*, 541–551. https://doi.org/10.1111/2041-210X.12686

Tuck, S. L., Porter, J., Rees, M., & Turnbull, L. A. (2018). Strong responses from weakly interacting species. *Ecology Letters*, *21*, 1845–1852. https://doi.org/10.1111/ele.13163

Vandermeer, J. H. (1969). The competitive structure of communities: An experimental approach with protozoa. *Ecology*, *50*, 362–371. https://doi.org/10.2307/1933884

Weigelt, A., de Luca, E., Roscher, C., Temperton, V., Buchmann, N., Fischer, M., … Meyer, S. T. (2016). Data from: Collection of aboveground community and species-specific plant biomass from the Jena Experiment (time series since 2002). PANGAEA, https://doi.org/10.1594/PANGAEA.866358

Weigelt, A., Schumacher, J., Walther, T., Bartelheimer, M., Steinlein, T., & Beyschlag, W. (2007). Identifying mechanisms of competition in multi-species communities. *Journal of Ecology*, *95*, 53–64. https://doi.org/10.1111/j.1365-2745.2006.01198.x

Weisser, W. W., Roscher, C., Meyer, S. T., Ebeling, A., Luo, G., Allan, E., … Eisenhauer, N. (2017). Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions. *Basic and Applied Ecology*, *23*, 1–73. https://doi.org/10.1016/j.baae.2017.06.002

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation: Model transferability. *Methods in Ecology and Evolution*, *3*, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Wilbur, H. M. (1972). Competition, predation, and the structure of the ambystoma-rana sylvatica community. *Ecology*, *53*, 3–21. https://doi.org/10.2307/1935707

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., … Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, *90*, 8995–9005. https://doi.org/10.1029/JC090iC05p08995

SUPPORTING INFORMATION

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

How to cite this article: Clark AT, Ann Turnbull L, Tredennick A, et al. Predicting species abundances in a grassland biodiversity experiment: Trade-offs between model complexity and generality. *J Ecol*. 2020;108:774–787. https://doi.org/10.1111/1365-2745.13316