Scale both confounds and informs characterization of species coexistence in empirical systems

Accepted for Publication in *American Naturalist:* 3 June 2019

Authors:

Adam Clark[1,2,3]*, *adam.tclark@gmail.com*

Helmut Hillebrand[2,4,5], *helmut.hillebrand@uni-oldenburg.de*

W. Stanley Harpole[1,2,6], *stan.harpole@idiv.de*

Author Affiliations:

*Corresponding author

[1] Department of Physiological Diversity, Helmholtz Centre for Environmental Research (UFZ), Permoserstrasse 15, Leipzig 04318, Germany

[2] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

[3] Synthesis Centre for Biodiversity Sciences (sDiv), Deutscher Platz 5e, 04103 Leipzig, Germany

[4] Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University Oldenburg, Schleusenstr. 1, 26382 Wilhelmshaven, Germany

[5] Helmholtz-Institute for Functional Marine Biodiversity at the University Oldenburg (HIFMB), Ammerländer Heerstrasse 231, 26129 Oldenburg, Germany

[6] Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, Halle (Saale) 06108, Germany.

Abstract

Identifying stable coexistence in empirical systems is notoriously difficult. Here, we show how spatiotemporal structure and complex system dynamics can confound two commonly used stability metrics in empirical contexts – response to perturbation, and invasion rate when rare. We use these metrics to characterize stable coexistence across a range of spatial and temporal scales for five simulated models, in which the ability of species to coexist in the long-term is known a priori, and an empirical old-field successional time-series. We term the resulting multivariate distribution of metrics a stability fingerprint. In accordance with a wide range of classic and recent studies, our results demonstrate that no combination of empirically tractable metrics or measurements is guaranteed to "correctly" characterize coexistence. However, we also find that heuristic information from the stability fingerprint can be used to broadly characterize dynamic behavior, and identify circumstances under which particular combinations of species are likely to persist. Moreover, stability fingerprints appear to be particularly well-suited for matching potential theoretical models to observed dynamics. These findings suggest that it may be prudent to shift the focus of empirical stability analysis away from quantifying single measures of stability, and towards more heuristic, multivariate characterizations of community dynamics.

Key words: population stability; community stability; spatial scale; temporal scale; perturbation; mutual invasibility

Introduction

The primary purpose of this paper is to demonstrate how an empiricist might study stable coexistence in real-world settings. Such assessments are of great practical importance, as they are necessary prerequisites for prescriptive management of ecological systems and for testing theory (Levin 1992; Murdoch 1994; Ives and Carpenter 2007). However, the proper application of existing coexistence metrics requires substantial a priori knowledge about community structure, or measurements that are infeasible in empirical contexts (Stommel 1963; Turelli 1986; Donohue et al. 2016). Consequently, predictions about whether a particular combination of species is likely to coexist in the long-term remain rare, particularly in diverse, real-world systems (Lawton 1999; Chesson 2003; Siepielski and McPeek 2010; Adler et al. 2013).

Although there is no universally accepted definition of "stable coexistence", most ecologists would probably agree that a stably coexisting community should retain a particular combination of species, despite minor disturbances, even when measured at some point in the distant future. This property is called "persistence", and encompasses a wide range of dynamic behaviors (Pimm 1984; Anderson et al. 1992; Grimm and Wissel 1997). Persistent states might be centered around a fixed point (Saavedra et al. 2017), an oscillatory cycle (May 1974), a stochastic distribution (Turelli 1986), or a moving trend (Chesson 2017). They might be robust to infinitely small perturbations (May 1973), perturbations that fall within a fixed range (Armstrong and McGehee 1980), or any perturbation of any size (MacArthur 1970). Species might theoretically be able to persist indefinitely, or only up to some finite time horizon (Turelli 1980).

In practice, ecologists typically classify coexistence using one of two methods. Perturbation tests measure species responses to small disturbances (i.e. asymptotic stability; May 1973; Arnoldi et al. 2016). Stable coexistence is indicated if all species return to their pre-

disturbance state (which could be a fixed abundance or a dynamic trajectory). Alternatively, invasion analyses test whether species can re-invade a community after being driven to low abundance (i.e. mutual invasibility; Chesson 2000*a*, 2003). If all species can invade, this is taken to imply stable coexistence. These two metrics are conceptually similar, as they measure, respectively, the behavior of a community as it is pushed away from equilibrium towards a boundary, or away from a boundary towards equilibrium.

When applied correctly, either metric can rigorously codify opportunities for stable coexistence. However, fulfilling the necessary criteria can be onerous. For example, because perturbation tests only assess local stability (i.e. in the vicinity of individual equilibria), every possible equilibrium must be tested separately (Anderson et al. 1992; Saavedra et al. 2017). Similarly, in systems with multiple stable states – as might occur due to Allee effects, trophic structure, or intransitive competition – species may be able to coexist at high abundance even if they cannot invade when rare (Barabás et al. 2018). Thus, invasion analyses typically need to be augmented (Turelli 1981, 1986; Chesson and Ellner 1989), e.g. by introducing species at high enough density to exceed Allee thresholds (Chesson 2000*a* pp. 359-360), jointly introducing predator/prey or mutualist pairs (Chesson and Kuang 2008; Levine et al. 2017), or assessing invasions from every possible boundary point (Chesson 2018 p. 1786).

For several reasons, it seems unlikely that these criteria can be fully satisfied in empirical systems. First, without perfect a priori understanding of system structure, it is unclear which test augmentations might be necessary. For example, if a species fails to invade, this could simply indicate that Allee thresholds were not overcome, or that an obligate mutualist was omitted (Levine et al. 2017; Barabás et al. 2018). Likewise, if a perturbed species does not recover its initial state, this might indicate lack of stability, or that the equilibrium is nonstationary (Chesson

2017). A second challenge is that real-world systems are subject to demographic stochasticity, which can obscure effects of perturbations or invasions (Durrett and Levin 1994). For example, failure to recover after a disturbance, or failure to invade from low density, might be indicative of random chance rather than average system behavior (Turelli 1981, 1986; Tilman 2004). Finally, even in well-understood systems, it may not be feasible to conduct all necessary measurements. For example, diverse communities have enormous numbers of potential equilibria, each of which might need to be tested in separate experiments (Levine et al. 2017; Saavedra et al. 2017; Chesson 2018). This challenge is compounded by effects of spatial and temporal structure, which can cause community responses to perturbations and invasions to vary greatly across scales (Stommel 1963; Levin 1992). Worse still, the scales required for coexistence to manifest may differ across species (Leibold and Chase 2018), and "may be much larger than is considered in most models and field studies" (Chesson 2000*a*, p. 344), e.g. continents or centuries (Lawton 1999; Davis and Shaw 2001; Ricklefs 2008).

These fundamental challenges are potentially discouraging, and, in the words of Robert May, lead to "rather gloomy thoughts as to the extent to which one can, or cannot, hope to give the empiricist some precise, measurable definition of stability" (May 1973 p. 213). In other words, it seems likely that in empirical systems, no feasible combination of additional data or improved methods can conclusively demonstrate stable coexistence. In response, we suggest a shift of focus – away from rigidly interpreted tests, and towards a more heuristic approach. In doing so, we are inspired by classical insights from theory (Turelli 1986; Levin 1992; Lawton 1999), and from a growing consensus in resilience ecology which suggests that stability is best summarized using suites of metrics (Pimm 1984; Grimm and Wissel 1997; Carpenter et al. 2001; Ives and Carpenter 2007; Donohue et al. 2013, 2016; Arnoldi et al. 2016; Hillebrand et al. 2018;

Meyer et al. 2018; Zelnik et al. 2018). Building on these results, we demonstrate: *(i)* that empirically feasible measurements of coexistence reveal complex stability landscapes, which confound attempts to classify stability in a definitive manner; but *(ii)* that by combining information from across multiple metrics and scales of measurement, this complexity can be used to inform studies of general system behavior.

Methods

We approach our analyses from the perspective of an empiricist who has been confronted with a novel ecosystem. Given limited ability to conduct measurements, and no a priori information about its constituent species, we attempt to characterize coexistence. Due to these constraints, our methods are never guaranteed to "correctly" identify stable coexistence. We therefore follow the advice of Turelli (1986, p. 331) who, in reference to applying invasion analyses to complex community dynamics, suggests: "In desperation, one can ignore mathematical rigor, apply a heuristic coexistence criterion, then do simulations to check its accuracy and hope for the best." In other words, in cases where no theoretically sound, empirically tractable metrics exist, one must make do with imperfect, practical alternatives. In this spirit, we optimistically present a heuristic solution that provides useful insights to the problem of characterizing species coexistence in ecological communities.

We proceed in three stages. First, we introduce five spatially and temporally explicit models of ecological communities. It is known a priori from theory that in some of these models, long-term coexistence is possible, whereas in others, it is not. However, we hide this information from the empiricist. Second, we allow the empiricist to conduct perturbation tests and invasion analyses in these systems across a variety of spatial and temporal extents. These do not

necessarily correspond to the "best" scales for measurements, but rather represent subsets of scales to which measurements might be limited due to practical constraints (Fig. 1). We term the resulting multi-scale distributions of metrics "stability fingerprints" because they summarize unique information about system dynamics, and use them to assess stability in each model. Lastly, we apply this fingerprinting procedure to characterize empirical dynamics during old-field succession, using data from a 90-year chronosequence. This worked example serves as a guide for implementing our approach in a system where very few of the theoretical requirements for stability analysis have been met.

Simulating model dynamics

Our models are individual-based stochastic patch-occupancy simulations with similar structures. For all but one model (*PSF* – see below), we implement simulations using Gillespie's method following Lehman *et al*. (2012). This method has several advantages, including that demographic stochasticity arises as an emergent property, and that simulations perfectly match analytically tractable models at large spatial scales (i.e. master equations, *sensu* Black and McKane (2012)), which contributes a priori theoretical knowledge about their stability. Simulations occur on a 100-by-100 site grid, where sites can be occupied by at most a single individual. For each species i, dynamics depend on interspecific interactions and stochastic colonization and mortality events, with average rates c_i and m_i, respectively. For simplicity, we assume uniform dispersal across all patches. See *Appendix A.1-2* in the supplement model derivations and theoretical stability properties, *Appendix B.1-2* for details on Gillespie's method and for results in the absence of demographic stochasticity, and *Appendix E* for source code.

The first model, *Levins*, is a spatially and temporally explicit realization of the metapopulation model of Tilman (1994) and Levins (1969). At large spatial extents (i.e. large grids), average dynamics of species considered across all patches (i.e. the maximum spatial extent) follow Tilman (1994):

$$\frac{dp_i}{dt} = c_i p_i\left(1 - \sum_{j \le i} p_j\right) - m_i p_i - p_i \sum_{j<i} c_j p_j \quad (1)$$

where p_i is the fraction of sites occupied by species i. Competition in *Levins* is perfectly transitive, such that species i=1 is competitively superior to all other species, species i=2 is competitively superior to all species but species 1, etc. The first term in Eq. (1) represents colonization into sites that are empty or occupied by inferior competitors, the second term represents mortality, and the third term represents displacement caused by superior competitors. Following Tilman (1994), species equilibrium abundance p_i^* can be calculated sequentially starting with the best competitor as:

$$p_i^* = 1 - \frac{m_i}{c_i} - \sum_{j<i} p_j^* \left(1 + \frac{c_j}{c_i}\right) \text{ if } p_i^* > 0; = 0 \text{ otherwise} \quad (2)$$

For species i to persist at equilibrium (i.e. $p_i^*>0$), c_i must exceed m_i, and colonization rates of inferior competitors must be sufficiently large to offset competition. Coexistence is globally stable, meaning that species are drawn towards the equilibria in Eq. (2) from any nonzero starting abundance (Fig. 2a-b).

Our second model, *disturbance*, is almost identical to *Levins*, but also includes periodic events every D time steps that destroy fraction d_i individuals of each species. When considered across time (i.e. periods with and without disturbances), disturbances increase average mortality rate by a factor $\log(1 - d_i)/D$. Time-averaged abundances and global stability at large spatial scales can be calculated using these average rates and Eq. (2), just as in the *Levins* model (Chesson and Huntly 1997; Barabás and Ostling 2013) (see *Appendix A.2* and Fig. S1 for more

details). However, the more erratic dynamics caused by disturbances mean that short-term responses to perturbations – and therefore empirical estimates of stability metrics – can differ greatly between *Levins* and *disturbance*.

Our third model, *plant-soil feedback* (*PSF*), follows a previously published model (Suding et al. 2013) in which spatial structure and demography are similar to those in *Levins*, but species alter the environment in patches they occupy. These effects build up over time, such that species either increase their own mortality and decrease that of their competitors, or vice versa (for negative or positive feedbacks, respectively). Because of these feedbacks, we could not implement *PSF* using Gillespie's method (see *Appendix B.1* for details). Given uniform dispersal, long-term coexistence requires negative feedbacks for all species (Suding et al. 2013). This coexistence is locally stable, such that communities can recover from small perturbations, but large perturbations lead to alternate stable states. These alternate states form because negative feedbacks reduce species performance over time, eventually allowing invaders to displace resident species (Fig. 2e-f; see *Appendix A.2* and Figs. S2-S3 for details).

Our fourth model, *rock paper scissors* (*RPS*), has the same structure as *Levins*, except that species follow an intransitive competitive hierarchy, such that in terms of competitive ability, species $i>j$, $j>k$, $k>l$, and $l>i$. All other species pairs can displace one another, but without hierarchical advantages. We consider a community of four species in which species share the same demographic rates m and c, which leads to neutrally stable oscillations around a fixed point (sensu Allesina and Levine 2011; Grilli et al. 2017). Demographic stochasticity therefore causes species to drift toward extinction, meaning that long-term coexistence of all four species is not possible (Fig. 2g-h). However, because oscillations are compensatory across species, total

summed community abundance acts like a single species in the *Levins* model, and is globally stable around:

$$\sum_{i=1}^{n} p_i^* = 1 - \frac{m}{c} \quad (3)$$

Lastly, our fifth model, *neutral*, again follows the same form as *Levins*, except that all species share the same demographic rates *m* and *c*, and no species are competitively dominant (i.e. no species can colonize an occupied site). Population-level dynamics are therefore dominated by ecological drift *sensu* Hubbell (2001), and species cannot coexist in the long-term (Fig. 2i-j). As with *RPS*, equilibrium total community abundance is globally stable and is approximated following Eq. (3).

For simplicity, we simulated communities of two species for all models, except for *RPS* which included four species. Whenever possible, we chose parameters that resulted in long-term persistence of all species (see *Appendix A.3* for parameter values). For *Levins*, we included a fast-dispersing inferior competitor and a slow-dispersing superior competitor, for which globally stable coexistence is predicted by Eq. (2). For *disturbance*, we chose parameter values such that competitive exclusion was predicted in the absence of disturbance, but long-term coexistence was possible due to trade-offs between competitive ability and resistance to disturbance. For *PSF*, we included two species, both with negative feedbacks, which allowed locally stable coexistence. Recall that long-term coexistence is not possible in *RPS* or *neutral*.

Testing for stable coexistence

For each model, we characterized coexistence across a range of spatial and temporal extents (denoted *s* and *t*, respectively) by conducting perturbation tests and invasion analyses. These extents represent contiguous units of space and time, as might be sampled in an empirical

study (e.g. a 1-by-1 m plot, measured over five field seasons). We considered extents ranging from 0.5% to 100% of the maximum spatial extent (i.e. all 100-by-100 sites), and from 1 to 200 simulated time-steps.

We conducted manipulations and measurements only at the focal extents corresponding to each test. These procedures were meant to mimic plot-based ecological observations, where only a fraction of the landscape can be manipulated and observed (e.g. Fig. 1; see *Appendix A.4* for details). We simulated 20,000 iterations of each model, and report the median test result across iterations (see Fig. S4 for distributions). This procedure generated fingerprints for each model that summarized stability across extents.

To conduct perturbation tests, we applied small perturbations to a species (20% reduction in abundance), and compared dynamics to those in simulations without the perturbation (e.g. as might be accomplished by comparing control and treatment plots; see Fig. 2, "perturbation"). We quantified this response as the average rate of return:

$$r_e(s,t) = \max(\text{over all species } i)\left[\frac{1}{t}\log\left(\frac{|p_{i,s,t}-p^0_{i,s,t}|}{|p_{i,s,0}-p^0_{i,s,0}|}\right)\right] \quad (4)$$

where $p_{i,s,t}$ describes abundance of species i, measured at spatial extent s and temporal extent t, and $p^0{}_{i,s,t}$ describes what the abundance of species i would have been had the perturbation not occurred. If the system returns towards its initial state, then $r_e<0$, which is interpreted as stability. Thus, r_e represents the most positive, and therefore the least stable, response observed across all species (see Fig. S5a-c for examples), and is formally an approximation of the Lyapunov exponent (see *Appendix A.5* for details and justification).

To conduct invasion analyses, we removed a species from the community, allowed the resulting community to equilibrate to account for transient dynamics, and then re-introduced the species at low abundance (5% of unoccupied sites) *sensu* Chesson (2000*a*) (e.g. as might be

accomplished through experimental introduction of species into an existing community; see Fig. 2, "removal" and "invasion"). We quantified invasions using the average growth rate:

$$r_0(s,t) = \min(\text{over all species } i)\left[\frac{1}{t}\log\left(\frac{p_{i,s,t}}{p_{i,s,0}}\right)\right] \quad (5)$$

where t describes time since re-invasion, and r_0 describes the minimum rate observed across all species (i.e. the least positive, and therefore the least stable; see Fig. S5d-e for examples).

Note that if systems recover from perturbations or invasions, both r_e and r_0 necessarily approach zero for large t. Ideally, this property should be controlled for by considering different subsets of temporal extents (Sheil and May 1996). However, this may not always be possible in empirical settings (e.g. Fig. 1a, where system dynamics are fast relative to measurements). Thus, particularly for large t, our results represent potential outcomes that might arise due to sampling constraints, rather than ideal tests that are best suited for detecting coexistence.

Empirical example

To demonstrate how our methods might be applied to real-world systems, we analyzed a 90-year old-field successional chronosequence from the Cedar Creek Ecosystem Science Reserve in Minnesota, USA (Inouye et al. 1987; Clark et al. 2019). This chronosequence includes 23 fields, abandoned from agricultural use between 1927 and 2015. In each field, species-level percent-cover of herbaceous plants has been surveyed in 100 permanent 0.5-by-1-m plots, roughly every five years since 1983. We chose this dataset because of its uniquely large range of measured spatial and temporal extents (see Fig. S6). Note, however, that very few theoretical assumptions for coexistence analysis are met in this study: e.g. all species are perturbed to low abundance simultaneously through tilling, and no control observations in

undisturbed fields are available. Data can be accessed on the LTER network data portal: https://doi.org/10.6073/pasta/aa029df8f7a6091ea879ceb5c6673963 (Knops, 2018).

We combined species into functional groups following Clark (2017), and retained the three most abundant categories: annual species, cool-season (C3) perennial grasses, and warm-season (C4) perennial grasses/sedges. These accounted for >80% of total cover. Importantly, successional dynamics and long-term persistence for these groups are well-known: annuals are primarily found early in succession, C3 grasses are typically mid-successional, and C4 grasses dominate late in succession (Clark et al. 2019). See *Appendix C* for more details on the site and analysis methods.

We conducted analyses in three steps. First, to measure stability, we treated succession as an invasion event and approximated r_0 based on changes in percent-cover over successional time following Eq. (5). Similar metrics might be calculated following any large disturbance that reduces species to low abundance – e.g. experimental manipulations or natural events. Although we could not do so here, in other systems it may be possible to approximate r_e based on differences between perturbed and unperturbed replicates, following Eq. (4).

Next, we calculated r_0 across observed spatial and temporal extents to generate an empirical stability fingerprint. We used years since agricultural abandonment as a proxy for t, yielding extents ranging from 1 to 89 years post-disturbance. For spatial extents, we aggregated nested subsets of plots based on proximity (e.g. neighboring plots), and calculated s as total surveyed area, ranging from 0.5 to 406 m^2 (n.b. we excluded some experimentally burned plots from analyses). Similar methods could be employed in any system with repeated sampling through time or spatially replicated observations. Unlike our analyses of simulated data, we present rates scaled by time (i.e. $r_0 t$), as this helps visualize dynamics over long temporal extents

(n.b. this transformation does not influence subsequent analyses of metric distributions or model comparisons, which focus on relative values within individual spatiotemporal extents). In general, we suggest similar transformations for any analysis of long-term data, especially if systems appear to approach an equilibrium.

Finally, we compared empirical fingerprints from the old-fields to those from two models: *(i) Levins-OF*, a realization of *Levins* parameterized with demographic rates following Clark (2017); and *(ii) neutral-OF*, a realization of *neutral* based on species average demographic rates. We test *Levins-OF* because it has been hypothesized as a model of successional dynamics at Cedar Creek (Tilman 1994; Clark et al. 2019). We include *neutral-OF* as a simple alternate model, because it requires little a priori information for parameterization. To emulate the old-fields' disturbance history, we initialized models using abundances observed immediately post-abandonment, and simulated 90 years of succession. We then compared the empirical and simulated systems by calculating the likelihood of empirical fingerprint given simulated fingerprints (see *Appendix D* for details).

Results

Model dynamics

Levins most clearly demonstrated properties that are commonly associated with stable coexistence. At large extents, species recovered rapidly following perturbations and invasions, and equilibrated around analytical expectations from Eq. (2) (Fig. 2a-b). Results were similar for *disturbance*, except that species followed oscillatory cycles (Fig. 2c-d). For *PSF*, species recovered from perturbations, but invasions led to alternate stable states wherein each species could displace its competitor, but the two could not coexist (Fig. 2e-f). For *RPS*, perturbation

tests led to persistent changes in oscillatory dynamics (Fig. 2g; n.b. solid and dashed lines do not converge). In invasion tests, removing one species led to extinction of a second species, which left the removed species' superior competitor unchecked (Fig. 2h). Thus, subsequent re-invasion was only transiently successful. Lastly, as expected for *neutral*, both perturbation and invasion tests changed system dynamics, with no recovery to initial state (Fig. 2i-j).

Stability fingerprints

As expected, neither heuristic metric "correctly" identified stable coexistence across all models and scales (recall, long-term coexistence is possible in *Levins*, *disturbance*, and *PSF*, but not *RPS* and *neutral*). Across models, r_e generally decreased with spatial extent, such that below 1% of the maximum extent, it was always positive (Fig. 3a-e). For r_0, patterns were more variable, though it often increased at larger spatial extents (Fig. 3f-j). Both metrics approached zero for large temporal extents, as expected for time-averaged growth rates (Sheil and May 1996).

For *Levins*, both metrics correctly indicated that long-term coexistence was possible, with negative r_e (i.e. recovery from perturbation) and positive r_0 (i.e. successful invasion), except at large spatial or small temporal extents (Fig. 3a,f). For *disturbance*, metrics gave conflicting results. Although long-term coexistence was possible, immediately after disturbances we found negative r_e (indicating stability – note narrow bands visible at larger spatial scales) and negative r_0 (indicating lack of stability), whereas between disturbance events we detected positive r_e and positive r_0 (Fig. 3b,g). For *PSF*, for which long-term coexistence is locally stable, results were similar to those for *Levins*, despite the existence of alternative stables states (Fig. 3c,h). Finally, for *RPS* and *neutral*, both metrics correctly indicated lack of stability at large spatiotemporal

extents (Fig. 3d-e,i-j). However, for both models, r_0 incorrectly indicated stability at small spatial extents, and for *RPS*, r_0 also indicated stability during periods of transient reinvasion.

Empirical example

Although our empirical analysis of successional dynamics failed to meet many theoretical requirements, results nevertheless accorded with species hypothesized successional niches. Across spatial extents, annuals showed successful invasions for roughly the first ten years of succession, followed by population declines over longer temporal extents (Fig. 4a). For C3 grasses, growth was generally negative for the first 5 years and then positive for longer temporal extents, though growth was always positive at spatial extents above about 25 m^2, and declined somewhat at temporal extents above 20 years (Fig. 4b). Trends for C4 grasses were similar, except that positive growth only began around successional year 20 or spatial extents above 70 m^2, and then increased monotonically for larger temporal extents (Fig. 4c).

Correspondence between simulated and empirical fingerprints was almost always higher for *Levins-OF* than for *neutral-OF*, particularly for temporal extents above 10 years, or spatial extents above 50 patches (Fig. 4d-f). Unlike *neutral-OF*, *Levins-OF* successfully predicted declines in r_0 with field age for annuals, and increases in r_0 with field age at smaller spatial extents for C3 and C4 grasses (Fig. 4g-i). Interestingly, this was true even though both models predicted similar dynamics over the first decade of succession, and consistently over-predicted abundances (see Fig. S7).

Discussion

Across all models, no combination of empirically tractable measurements was guaranteed to correctly identify stable coexistence. While not surprising, this result serves as a reminder that attempts to characterize coexistence are necessarily heuristic from the perspective of an empiricist – i.e. given limited ability to make measurements and imperfect understanding of system structure (Turelli 1986; Levin 1992; Murdoch 1994; Lawton 1999; Donohue et al. 2016). Thus, definitive statements about coexistence in empirical contexts are probably neither testable nor warranted.

More encouragingly, our results show that multi-metric, multi-scale empirical tests can still produce useful results. Specifically, multivariate stability fingerprints contribute three main types of insight, described in detail below. First, based on empirically tractable measurements, fingerprints summarize major components of system dynamics which often relate closely to long-term coexistence. Second, in cases were predictions about long-term dynamics are unclear or confounded, fingerprints help identify potential drivers of uncertainty. Third, fingerprints aid in identifying potential mechanisms underlying observed dynamics. Jointly, these findings support conjectures from a broad range of studies, which suggest that the focus of stability analysis should be shifted towards more holistic, multivariate assessments (Levin 1992; Grimm and Wissel 1997; Ives and Carpenter 2007; Donohue et al. 2016; Zelnik et al. 2018).

Applying fingerprinting

Although the simulation methods we employ are complex, note that generating the stability fingerprints themselves is relatively simple: r_0 is effectively a log response ratio comparing population size at time t vs. initial observed population size, and r_e compares the log

ratio between two observed time-series (e.g. replicates with and without an experimental perturbation). Moreover, our results do not rely on a theoretically "optimal" perturbation. For example, our analysis of empirical old-field data is in many ways a "worst case scenario" for stability analyses, as all species were perturbed to low abundance simultaneously, and no control plots were available.

Despite their simplicity, fingerprints generally succeeded in characterizing complex dynamics. For example, matching a priori expectations for our models, r_0 demonstrates that re-invasion in *RPS* is transient (Fig. 3i), and r_e shows that in *neutral*, species abundances can recover from small perturbations, but only temporarily (Fig. 3e) (Hubbell 2001; Grilli et al. 2017). Likewise, for empirical old-field dynamics, r_0 shows that growth rates for C3 and C4 grasses are only positive when measured at extents above a few dozen plots, or temporal extents of greater than 5-10 years (Fig. 4a-c). This result accords with theoretical expectations for the *Levins* model, which suggests that long-term persistence of late-successional species requires large spatial extents (Tilman 1994).

Note that stability fingerprints are strongly influenced by perturbation type. We demonstrate two types of responses, which effectively represent different ends of a continuum ranging from small and instantaneous (r_e) to large and distributed across time (r_0) (Ives and Carpenter 2007). However, many other perturbation types also contain useful information. For example, although the regularly occurring mortality events in *disturbance* generally confounded our tests, species responses to these events help demonstrate the long-term viability of coexistence (see Fig. S8). Combining information from across many species, or many different perturbation types including experimental manipulations and natural events, could therefore greatly increase information availability, especially in systems where observational scales are

limited (Pimm 1984; Carpenter et al. 2001; Donohue et al. 2013, 2016; Arnoldi et al. 2018). This approach has recently been termed "probing", and suggested as a general method for classifying complex dynamics (Zelnik et al. 2018).

Addressing problems

It is tempting to imagine that deviations between our predictions from individual indices and the actual long-term persistence of species are idiosyncratic to the metrics and models we consider. However, as noted, our results accord with a broad array of studies, which suggest that many of the mischaracterizations that we demonstrate are indicative of fundamental obstacles and unavoidable trade-offs (Turelli 1986; Levin 1992; Murdoch 1994; Chesson 2000*b*; Carpenter et al. 2001; Hubbell 2001; Levine et al. 2017; Saavedra et al. 2017; Barabás et al. 2018). Nevertheless, most of these challenges can be at least partially overcome by assessing multiple metrics and scales.

First, consider alternate stable states, which can confound tests if perturbations are sufficiently strong to overcome locally stable coexistence (Chesson 2000*a*; Levine et al. 2017; Saavedra et al. 2017; Barabás et al. 2018). In *PSF*, for example, invasion tests shift simulations into a state where both species can displace the other, but the two cannot coexist (Fig. 2f; see *Appendix A.2*; n.b. although $r_0>0$, most ecologists would probably classify this result as unstable). In contrast, r_e indicates that species can coexist in the long-term at most scales (Fig. 3c). Jointly, these results (correctly) suggest that coexistence in *PSF* is locally stable, but that coexistence is not robust to large disturbances.

Conversely, just as disturbances that are too large risk obscuring local stability, perturbations that are too small can be impossible to distinguish from background noise (May

1973; Turelli 1980). For example, r_e consistently predicts instability at small spatial extents (s<0.05) regardless of underlying dynamics (Fig. 3a-e). This is because these extents harbor small populations, which allows demographic stochasticity to overwhelm effects of perturbations (see Figs. S9-S10). Similar effects confound r_e at all spatial extents for *disturbance*, because species are frequently driven to low abundance (see *Appendix B.2* and Fig. S11). Results from invasion tests are less strongly influenced by stochasticity, as effects of larger perturbations are easier to detect (Fig. 3f-j). Thus, especially when populations are small, cases where both r_e and r_0 are positive likely indicate that long-term coexistence is possible, but that dynamics are strongly influenced by stochastic fluctuations.

A related trade-off involves temporal extent. Effects of perturbations or invasions are typically diluted over time, making responses difficult to detect (Sheil and May 1996). A partial solution is to scale estimates of r_e and r_0 by t, as we do in Fig. 4, as these transformed metrics measure total recovery rather than recovery per unit time. If long-term estimates of r_et remain negative, or r_0t remain positive, then this indicates long-term coexistence. Tests conducted at shorter temporal extents also risk being confounded by transient or fluctuating dynamics, as is the case with r_0 for *RPS* (Fig. 2h,3i), or short temporal extents (t<D) in *disturbance* (Fig. 3b,g). In general, we therefore advocate caution when interpreting results from systems that may be subject to transience or fluctuations if it is suspected that surveys are too short to accurately capture the full range of potential dynamics.

One last obstacle pertains to processes acting outside of the scope of tests. For example, in *PSF* and *neutral*, r_0 incorrectly indicates stability at small scales because dispersal from outside the focal area overwhelms local dynamics (Fig. 3i,j) (Pimm 1984; Hubbell 2001). Similarly, r_e incorrectly indicates stability for *neutral* even at large spatial extents because total

community abundance is stable (see Fig. S12), which leads to compensatory increases in abundance that are shared across species (see Fig. S13). While these effects could potentially be mitigated by preventing immigration into the focal patch, or by replacing removed individuals with those of another species, such interventions would be difficult to implement in practice, and could inadvertently destabilize the system, e.g. by altering metacommunity processes (Tilman et al. 1994; Leibold and Chase 2018) (see example in Fig. S14). In these cases, there is no general solution. Either statements about stability must be limited to the observed range of spatial and temporal extents, or specific assumptions must be made about how the system behaves outside of these extents (see *Identifying mechanisms*, below).

Identifying mechanisms

In addition to providing qualitative information about system dynamics, fingerprints also appear to be useful for identifying potential underlying mechanisms that influence system dynamics. For example, our old-field analyses suggest that *Levins-OF* better explains observed dynamics than does *neutral-OF*, in accordance with results from long-term studies at Cedar Creek (Gleeson and Tilman 1990; Tilman 1990, 1994; Clark et al. 2019). Likewise, despite superficial similarities among some fingerprints – e.g. *Levins* vs. *PSF* (Fig. 3a,c) – we were typically able to distinguish among simulated models with high certainty, especially when considered across multiple extents (see *Appendix D* and Fig. S15). Critically, once an appropriate model has been identified, its stability properties can be used as a proxy of those for the empirical system, which overcomes many of the problems discussed here (Ellner et al. 2019).

An important caveat is that all information that is contained in stability fingerprints is also available in timeseries data of species abundances. Thus, similar results could probably be

achieved by fitting dynamic models to multi-scale information about species abundances. However, doing so requires complex methods that are rarely applied in practice (Detto and Muller-Landau 2016; Clark et al. 2018). Fingerprints may therefore be useful as a low-dimensional summary of dynamic information, which is comparatively easy to interpret and compare across models. This feature could be particularly important in empirical systems where community composition is influenced by large-scale processes such as metacommunity dynamics, biogeography, or evolutionary history (Lawton 1999; Chesson 2000*a*; Davis and Shaw 2001; Ricklefs 2008; Leibold and Chase 2018). Ideally, information already collected as part of other studies may provide sufficient spatial and temporal extents to identify underlying mechanisms. However, for some systems or mechanisms, available data may not be sufficient to identify unique aspects of fingerprints, and surveying the necessary scales may not be feasible. Under these circumstances, fingerprinting may still help reduce the number of potential mechanisms that could explain observed patterns.

Future directions

Here, we present an empirically tractable approach for characterizing long-term coexistence in ecological communities. By exploring some of the challenges that confront stability analysis, we hope that we have convinced readers that rigidly interpreted, binary metrics of coexistence are probably not appropriate in most empirical contexts. More importantly, by demonstrating how fingerprinting can be applied to heuristically characterize system dynamics and stability, we hope to encourage future empirical studies to employ a wider range of experimental perturbation treatments, and to conduct measurements across a broader array of spatial and temporal extents.

The next challenge will be to find combinations of metrics, scales, and perturbation types that can uniquely identify a wide range of dynamic behavior across many systems (Levin 1992; Ives and Carpenter 2007). For some systems, this may require new measurements, or methods for extrapolating dynamic behavior across unobserved spatial and temporal scales (Leibold and Chase 2018). However, in many cases, it seems likely that the data, methods, and theory necessary for conducting these tests are already in place. We are therefore optimistic that stability fingerprinting could help greatly expand our understanding of coexistence in a rich variety of theoretical and empirical systems.

Acknowledgements

We are grateful to P. Amarasekare and Y. Feng for helpful discussions regarding coexistence in diverse systems, and to the PhyDiv, sDiv, Chase, and Hillebrand lab groups for feedback on earlier versions of this manuscript. We thank editors D. Bolnick and M. Vellend, reviewer G. Barabás, and several anonymous reviewers for detailed comments that helped us focus our manuscript more specifically on empirical applications. A.T.C. was supported by a Catalyst Postdoctoral fellowship through sDiv. Support for developing the *PSF* model was provided to W.S.H. by the Center for Invasive Plant Management and National Science Foundation (DEB-0614168, DEB-1001807). Computing resources were provided by iDiv and UFZ. Lastly, we are grateful to D. Tilman for use of his old-field time-series data. Collection of these data was made possible by more than three decades of work by the staff, researchers, and interns at Cedar Creek, and was supported by the US National Science Foundation Long-Term Ecological Research Program (DEB-8114302, DEB-8811884, DEB-9411972, DEB-0080382, DEB-

0620652, and DEB-1234162), the Cedar Creek Ecosystem Science Reserve, and the University of Minnesota.

Data accessibility statement: All data and source code needed to reproduce the analyses and figures presented here are available in the electronic supplement, in the archive "stability_scaling.zip". Data used in this study can be accessed on the LTER network data portal: https://doi.org/10.6073/pasta/aa029df8f7a6091ea879ceb5c6673963 (Knops, 2018).

Appendices:

Table of Contents:

Appendix A. Model parameterization

A.1: Model derivations

At large spatial extents, effects of demographic stochasticity can be ignored, and dynamics for *Levins* follow the master equation form described in Eq. (1) in the main text (equation reproduced here for convenience):

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = c_i p_i\left(1 - \sum_{j \le i} p_j\right) - m_i p_i - p_i \sum_{j<i} c_j p_j \qquad (1)$$

where p_i is the fraction of sites occupied by species i, c_i and m_i are per-capita colonization and mortality rates, respectively, and species indices are ordered by competitive ability, such that $j<i$

implies that species *j* is competitively superior to species *i*. Note that this is identical to the Levins model, *sensu* Tilman (1994), and is the master equation for our stochastic implementation of the model.

In the model *disturbance*, disturbance events occur at regular time intervals, killing off fixed fraction d_i of the individuals of each species. Between disturbance events, dynamics follow the same form as *Levins*. Thus, dynamics at large extents can be expressed as:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = c_i p_i\left(1 - \sum_{j \le i} p_j\right) - m_i p_i - p_i \sum_{j<i} c_j p_j \qquad \text{if } t \bmod D \neq 0 \qquad \text{(A1a)}$$

$$p_{i|D} = (1 - d_i) p_{i|D'} \qquad \text{if } t \bmod D = 0 \qquad \text{(A1b)}$$

where $p_{i|D}$ is abundance of species *i* immediately after the disturbance event, $p_{i|D'}$ is the abundance immediately before, and the modulo operator "*x* mod *y*" takes the remainder of *x*/*y* (thus, *t* mod *D* = 0 only if *t* is a multiple of *D*). Note that rate d_i is consequently a rate in discrete-time, whereas m_i is a rate in continuous-time, which is why the log transformation $\log(1 - d_i)/D$ is required in our calculation of mean mortality rate for the master equation.

For *RPS*, we simulated a system of four species in a competitive loop, such that species *i* is competitively superior to species *j*, *j* is competitively superior to species *k*, *k* is competitively superior to species *l*, and *l* is competitively superior to species *i*. All other species pairs (i.e. *i* vs *k*; *j* vs. *l*; and *k* vs. *i*) are competitively equivalent, meaning that either can displace the other. Demographic rates are equal across all species, yielding a system of equations of the form:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = cp_i(1 - p_i - p_l) - mp_i - cp_i(p_k + p_l) \qquad \text{(A2a)}$$
$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = cp_j\left(1 - p_j - p_i\right) - mp_j - cp_j(p_i + p_l) \qquad \text{(A2b)}$$
$$\frac{\mathrm{d}p_k}{\mathrm{d}t} = cp_k\left(1 - p_k - p_j\right) - mp_k - cp_k(p_i + p_j) \qquad \text{(A2c)}$$
$$\frac{\mathrm{d}p_l}{\mathrm{d}t} = cp_l(1 - p_l - p_k) - mp_l - cp_l(p_j + p_k) \qquad \text{(A2d)}$$

For *neutral*, dynamics are identical to Eq. (1), except that species share the same demographic rates, and do not have direct competitive interactions. Thus, given a community of *n* species, dynamics can be written as:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = cp_i\left(1 - \sum_{j=1}^{n} p_j\right) - mp_i \qquad \text{(A3)}$$

Note that dynamics match those of Hubbel's neutral model (Hubbell 2001), except that we include time-explicit colonization and mortality, rather than following the zero-net-sum assumption (i.e. that empty sites are immediately filled). Thus, fewer than 100% of sites are occupied at any given moment in time.

Lastly, for *PSF*, we used a model presented by Suding *et al*. (2013). Spatial structure and demography are similar to that for *Levins*, except that there are no direct competitive interactions among species (i.e. species cannot displace one another). Rather, species influence local conditions in sites that they occupy in ways that either hurt themselves and favor their competitors (negative feedbacks), or favor themselves and hurt their competitors (positive feedbacks). Given two species *i* and *j* ("exotic" and "native", respectively, in the original model), survival probability in any given time-step is calculated as:

$$s_i = \frac{(1 - sign(f_i)S)}{2} - m_i \qquad \text{(A4a)}$$

$$s_j = \frac{(1 + sign(f_j)S)}{2} - m_j \qquad \text{(A4b)}$$

where f_i indicates the strength and direction of feedbacks, *S* is the local site condition, $sign(f_i)$ indicates whether f_i is positive or negative, and *m* is the background rate of stochastic mortality. Local site state *S* varies from –1 to 1.

In our implementation of the model, soil state declines by 1 percentage point every time step that a site is occupied by species *i*, whereas every time step that it is occupied by species *j*, *S* increases by 1 percentage point, with respective maxima and minima determined by –abs(f_i) and

+abs(f_j). In unoccupied sites, S approaches 0 by 1 percentage point per time step. Note that with negative feedbacks for both species, each species increases its own mortality in sites that it occupies, and decreases mortality for its competitor. Lastly, in our implementation of the model, probability of establishment in empty sites is determined following the formula:

$$p_i = (1 - p_0)\frac{c_i s_i}{c_i s_i + c_j s_j} \quad \text{(A5a)}$$

$$p_j = (1 - p_0)(1 - p_i) \quad \text{(A5b)}$$

where c_i is the seed production of species i relative to j, and p_0 is the probability that neither species successfully recruits into the patch. Note that this equation differs slightly from that in the original publication, for which $p_0=0$.

A.2: Analytical stability properties

Recall that unless noted otherwise, all discussions here and in the main text relate to population stability (i.e. the stability of the least stable species in the community, meaning the species with the highest r_e or the smallest r_0). Nevertheless, results for community stability (i.e. stability of total community biomass) are similar to those for population stability (see Fig. S12 for results for community stability).

As shown in the main text, we can solve for the analytical expectation of species equilibrium abundances by setting Eq. (1) to zero, and solving for p_i, which yields Eq. (2) from the main text (equation reproduced here for convenience):

$$p_i^* = 1 - \frac{m_i}{c_i} - \sum_{j<i} p_j^* \left(1 + \frac{c_j}{c_i}\right) \text{ if } p_i^* > 0; = 0 \text{ otherwise} \quad (2)$$

This estimate is accurate at spatial extents that are large enough to avoid strong effects from demographic stochasticity (e.g. thousands of cells). At these scales, *Levins* is globally stable around these equilibrium values (i.e. they are approached from any non-trivial starting point)

(Tilman 1994). As discussed in the main text, species achieve positive equilibrium biomass given a sufficiently strong trade-off between competitive and colonization ability. Note that because the equilibrium abundance of each species is globally stable, community abundance and composition are necessarily also stable.

For *disturbance*, stability properties are identical to those in *Levins* provided that all species are predicted to have positive abundances following Eq. (2) based on their time-averaged demographic rates (again, assuming that the grid is large enough to avoid large effects of demographic stochasticity). By "time-averaged", we mean the average rates when considered across time periods with and without disturbances. These rates can be calculated as $m_i - \log(1 - d_i)/D$, where m_i is the mortality rate in the absence of disturbance, d_i is the fraction of individuals destroyed during disturbances, and D is the time between disturbance events (n.b. c_i parameters are not influenced by disturbance). Given a trade-off between species competitive abilities and their tolerance to disturbance events (i.e. d_i), stable coexistence is therefore possible even among combinations of species that would competitively exclude one another in the absence of disturbance (Fig. S1). Average abundances of species through time in *disturbance* can be calculated exactly from the standard *Levins* model using Eq. (2), by setting the mortality rates to equal the time-averaged rate. Nevertheless, *disturbance* displays substantial transient dynamics between disturbance events which differ from the average abundances, and significantly influence perceived system stability (as discussed in the main text).

For *PSF*, stability properties have been extensively analyzed in Suding *et al.* (2013). Locally stable coexistence occurs given sufficiently strong negative feedbacks for both species. Under these circumstances, species fall into fixed oscillatory cycles. In our model, these cycles are compensatory, such that total community abundance remains at a fixed level (Fig. 1e-f in the

main text). Note, however, that the model is subject to alternate stable states. Given positive feedbacks between species and their environment (i.e. such that positive effects build up over time as species occupy a site), these alternate stable states occur from almost any starting point (Fig. S2). Given negative feedbacks, as we explore in the main text (i.e. where negative effects build up over time as species occupy a site), alternate stable states are possible, but only if species have spent a prolonged time growing in monoculture before they are invaded (Fig. S3). These alternate stable states occur because negative feedbacks become so intense that the resident species is no longer able to resist competitive exclusion by a reintroduced competitor. Although this result does not accord directly with the classical plant-soil-feedback model of Bever *et al.* (1997) (which does not predict alternate stable states in cases with negative feedbacks), this is not due to a major conceptual difference between our models, but rather because we consider stronger negative feedbacks (n.b. an equivalent result in their model would arise from feedbacks that are sufficiently strong to drive the phase cycle in their Fig. 2 into one of the zero boundaries).

For *neutral*, recall that because all individuals share the same demographic rates, dynamics for total community abundance can be treated as a single species in the *Levins* model, with equilibrium abundance (equation reproduced here for convenience):

$$\sum_{i=1}^{n} p_i^* = 1 - \frac{m}{c} \quad (3)$$

Thus, though population abundances and composition are purely subject to ecological drift and should therefore not be indicated as stable by any of our metrics (Hubbell 2001), total community abundance has the same stability properties as it does in *Levins*. Note that because of this drift, long-term coexistence is not possible.

Lastly, for *RPS*, dynamics follow a four-species competitive loop, *sensu* Fig. 1b in Grilli *et al.* (2017). For communities including all four species, all species are inhibited by a competitor, leading to a neutral, oscillatory system (Allesina and Levine 2011; Grilli et al. 2017). Because the system is neutrally stable, even small perturbations can lead to changes in the period and amplitude of oscillations, and long-term coexistence is not possible due to the effects of demographic stochasticity. However, because oscillations are compensatory, total community abundance remains relatively constant regardless of perturbation or invasion events (e.g. Fig. 1g-h in the main text). Thus, community abundance is likely to be indicated as stable following any of our metrics. Conversely, our perturbation and invasion metrics should indicate instability for population abundance and composition.

For communities in *RPS* composed of a single species, or any pair of species that have equal competitive abilities (i.e. *i* vs *k*; *j* vs. *l*; and *k* vs. *i*), species dynamics in *RPS* reduce to the *neutral* model, since individuals are equivalent in terms of demographic rates. Any combinations of species that include unconstrained competitive interactions (e.g. *i* and *j* alone; or *i*, *j*, and *l*) is inherently unstable, as it results in asymmetrical competitive interactions among otherwise identical species, which leads to rapid competitive displacement by the unconstrained species. For example, given competition between species *i* and *j*, equilibrium abundances predicted from Eq. (2) are:

$$p_i^* = 1 - \frac{m}{c} \tag{A6a}$$

$$p_j^* = 1 - \frac{m}{c} - p_i^*\left(1 + \frac{c}{c}\right) = 1 - \frac{m}{c} - 2\left(1 - \frac{m}{c}\right) = -(1 - \frac{m}{c}) \tag{A6b}$$

which implies competitive exclusion of species *j* (since $m < c$). Note that removal of a single species can therefore initiate a cascade of events, which ultimately prevent its successful re-introduction. For example, removing species *i* leaves growth of species *j* unchecked, thereby

driving species *k* extinct. Without species *k*, species *l* is able to grow unchecked, ultimately excluding *i* if it is re-introduced to the community (see Fig. 1h in the main text).

A.3: Parameter values

For all simulations, including all figures in the main text and supplement, we used the following parameter values:

1. *Levins:*
    a. $c_1 = 0.45$; $c_2 = 1.05$
    b. $m_1 = m_2 = 0.3$
2. *disturbance:*
    a. $c_1 = 0.435$; $c_2 = 0.600$
    b. $m_1 = m_2 = 0.3$
    c. disturbance frequency = 1 per 50 time-steps
    d. d_1 = 95% mortality; d_2 = 0% mortality
3. *PSF*:
    a. $c_1 = 1/c_2 = 1.5$
    b. $f_1 = f_2 = –80\%$
    c. $m_1 = m_2 = 5\%$
    d. initial soil state = 0
    e. $p_0 = 50\%$
4. *RPS:*
    a. $c = 0.48$ for all species
    b. $m = 0.224$ for all species
5. *neutral:*
    a. $c = 1.5$ for all species
    b. $m = 0.7$ for all species

For each set of extents, we simulated 20,000 iterations of each model. We simulated 300 time-steps before conducting manipulations, and 200 post-manipulation time-steps, including an additional 200 time-step burn-in period for invasion tests to account for transient dynamics.

A.4: Simulation procedure

All simulations were carried out on a 100-by-100 cell rectangular grid. In the simulations, each of the 10,000 resulting cells is either empty, or holds a single individual of a single species. We assume global dispersal in all models (i.e. all available cells are equally likely to be reached by propagules, regardless of distance). Results in the main text are based on the median of 20,000 simulated iterations of each model, at each of seven spatial extents (0.5%, 1%, 5%, 10%, 50%, 75%, and 100% of the maximum extent).

To calculate stability metrics, we first ran each model for a 300 time-step burn in period to avoid transient dynamics. We then paused the simulation, and branched it into three classes of runs: *null*, *perturbation*, and *invasion*. These branches all retained the same spatial and species layout, random number seeds, and projected future events (for the Gillespie simulations – see *Appendix B*), meaning that they would produce identical time series if left un-disturbed. For the *null* test, only a single branch was simulated, no additional changes were made, and dynamics were simulated for an additional 200 time-steps. For the *perturbation* tests, one branch was simulated per species. For each branch, the abundance of a single species was reduced by 20% at the beginning of the simulation, and the simulation was then run for an additional 200 time-steps to observe recovery. For the *invasion* tests, one branch was simulated per species. For each branch, the abundance of a single species was reduced to zero, and the remaining community was allowed to adjust to the change for 200 time-steps. Then, the removed species was re-introduced at an abundance equal to 5% of the total unoccupied sites.

A.5: Justification of perturbation metric

Perturbation analyses are usually conducted by calculating eigenvalues of the Jacobian matrix that describes interactions among system components (typically individual species), evaluated at equilibrium (May 1973). However, because eigenvalues represent a linear approximation of a high-dimensional system, it is impossible to characterize their response to noise without specific assumptions about the functional form of interactions among system components (e.g. mechanism of competition) and the statistical distribution of the noise (e.g. Gaussian, log-normal) (Chesson 1990; Anderson et al. 2010). Consequently, even minor observation error can lead to enormous uncertainty in eigenvalue estimates (Dormann 2008; Clark and Neuhauser 2018).

We therefore use a metric of response to perturbation measured at the level of individual species, rather than at the level of eigenvectors – i.e. r_e from Eq. (4) in the main text. Though our metric r_e can potentially be confounded by transient dynamics over short time periods (Fig. C1 in Arnoldi et al. 2018), its sign is guaranteed to match that of the dominant Eigenvalue over sufficiently long temporal extents, as both metrics correspond to asymptotic stability of the least stable species. Moreover, note that r_e is formally an approximation of the most positive Lyapunov exponent of our system, which is a very general metric commonly used to test for stability in complex dynamical systems.

Appendix B. Implementation of Gillespie's method

B.1: Primer on method

As noted in the main text, we used an implementation of Gillespie's method to simulate model dynamics. A major advantage of this method is that it perfectly matches analytical

expectations (e.g. Eq. 1) for large grids at the maximum spatial extent, but produces discrete, spatial results at smaller scales. By "discrete", we mean that birth and death events for individuals are explicitly considered, which allows us to account for temporal structure caused by demographic stochasticity (Durrett and Levin 1994). By "spatial", we mean that individuals occupied specific locations in a grid, which introduces spatial structure into our model. Note, however, that because we consider only interactions at the scale of an individual grid cell, and include global dispersal for all simulations, the specific spatial arrangement of individuals is not relevant for our study.

To implement Gillespie's method, we used the procedure and source code described in Lehman *et al.* (2012). This implementation simulates dynamics event-by-event (i.e. rather than time-step-by-time-step) by forecasting the expected time of future events (e.g. mortality or colonization) using an exponential waiting time distribution:

$$t_{event} = \log(-x + 1)/(-r) \quad \text{(B1)}$$

where t_{event} is number of time-steps in the future that the event will occur, x is a random uniform number drawn from over the range (0, 1), and r is a rate constant (e.g. m or c from the equations above). Note that t_{event} need not be an integer.

The method described in Lehman *et al.* (2012) is particularly fast and efficient because it bins future events into pre-sorted locations in computer memory, which allows rapid recall of events without the need to sort or search through large lists. Though describing the specifics of the algorithm is beyond the scope of this paper, we strongly recommend this procedure (available either in the cited paper, or in the "*.c*" files in our source code in *Appendix E*) for any readers who wish to implement their own versions of Gillespie's method, as it typically runs

more than three orders of magnitude faster than other versions of the algorithm that we have worked with.

One downside to the method is that it does not lend itself to cases with feedbacks between species and their environments. This is because the state of the environment depends on the precise history of occupancy in the patch, which is itself influenced by the changing state of the environment. This makes it impossible to project the timing of dispersal and mortality events into the future, and instead these must be updated every timestep as is done in standard time-step based simulation methods. This is why we used the previously published model for *PSF* rather than adapting it to match Gillespie's method, as we did with the other models.

B.2: Simulating master equations

To examine the behavior of our models in the absence of demographic stochasticity, we simulated the master equations in Eqs. 1 and A1-A3. These simulations were similar to those in a model by Muller-Landau (2010), in that we included two distinct patch types in each simulation: fraction s of the habitat was subject to observations and manipulations, similar to the observed spatial extent in our spatiotemporally explicit simulations, and fraction $(1 - s)$ of the habitat was not manipulated, but still contributed to and was influenced by dynamics in s due to dispersal.

As an example, for the *Levins* model, we simulated the following two patch system:

$$dp_{i,(1-s)}/dt = c_i(p_{i,s} + p_{i,(1-s)})(1 - s - p_{i,(1-s)}) - m_i\, p_{i,(1-s)} \quad \text{(B2a)}$$

$$dp_{i,s}/dt = c_i(p_{i,s} + p_{i,(1-s)})(s - p_{i,s}) - m_i\, p_{i,s} \quad \text{(B2b)}$$

$$dp_{j,(1-s)}/dt = c_j(p_{j,s} + p_{j,(1-s)})(1 - s - p_{i,(1-s)} - p_{j,(1-s)}) - c_i(p_{i,s} + p_{i,(1-s)})\, p_{j,(1-s)} - m_j\, p_{j,(1-s)} \quad \text{(B2c)}$$

$$dp_{j,s}/dt = c_j(p_{j,s} + p_{j,(1-s)})(s - p_{i,s} - p_{j,s}) - c_i(p_{i,s} + p_{i,(1-s)})\, p_{j,s} - m_j\, p_{j,s} \quad \text{(B2d)}$$

where $p_{i,s}$ is the abundance of species i in manipulated patches, and $p_{i,(1-s)}$ is abundance in unmanipulated sites. Note that abundance is always written in terms of fraction of the total patches occupied – thus, $p_{i,s} \leq s$, $p_{i,(1-s)} \leq (1 - s)$, and $p_{i,s} + p_{i,(1-s)} \leq 1$. Functional forms for the remaining models were similar, always with two differential equations per species (see source code in *Appendix E* for specific equations relating to each model). We did not simulate a version of the *PSF* without demographic stochasticity because we had no corresponding master equation for the model.

In general, results for the simulations of the master equations were similar to those for the spatiotemporally explicit models that we discuss in the main text (Fig. S11). For r_0, fingerprints from the models with and without demographic stochasticity were almost identical. In contrast, we found fewer instances of positive r_e, especially at small spatial extents. As discussed in the main text, this indicates that the instability detected by perturbation tests at small spatial extents is driven by demographic stochasticity rather than deterministic model behavior (see also Figs. S9-S10).

For three models – *disturbance*, *RPS*, and *neutral* – we generally found more instances of negative r_e across all scales for the master equations than for the stochastic simulations. For *disturbance*, this occurs because populations of both species are driven to small sizes during oscillatory dynamics. Thus, demographic stochasticity leads to relatively large random shifts in the starting abundance of species before they begin the recovery phase of their dynamics (Fig. S8), which can mask the effects of small perturbations even at large spatial extents. For *RPS* and *neutral*, demographic stochasticity influences the effects of perturbation tests – especially at large temporal extents – precisely because the systems are not deterministically stable. Thus, random fluctuations in population sizes compound over time, eventually driving replicate

trajectories apart. Interestingly, in the simulations of the master equations, both models "appear" stable based on r_e. As discussed in the main text, this occurs because total community biomass is stable in these models, leading to partial recovery at the population level following small perturbations, even in the absence of stable coexistence (S14).

Appendix C. Old-field succession example

A detailed description of the old-field succession experiment at the Cedar Creek Ecosystem Science Reserve is available in Clark (2017), in chapter 1, and in Clark *et al.* (2019). Briefly, Cedar Creek is a reserve run by the University of Minnesota and US Long Term Ecological Research Program, and located in Minnesota, USA (45.4°N, 93.2°W). The site is near the boundary between deciduous forest, boreal forest, and prairie biomes, though prior to European colonization, vegetation was primarily prairie, oak savanna, deciduous forests, and wetlands (Cushing 1963). Mean annual precipitation is a bit below 800 mm per year, with most occurring between April and August. Temperatures are highly variable, with summer highs averaging 27°C, and winter lows averaging -14°C. Soils are very sandy, and plant communities are strongly nitrogen limited (Tilman 1987).

Since 1983, successional dynamics of herbaceous plants have been followed in >20 old-fields at Cedar Creek. Of these, 23 are still surveyed roughly every 5 years for species-level percent cover. Each field includes 100 permanent plots, each 0.5-by-1 m. Plots are arranged into four parallel transects, each containing 25 plots. Transects are spaced 25 m apart, with 1 meter between plots within the same transect. Plots in four fields have experienced heavy afforestation, and since about 2008, half of the plots in most fields have been experimentally burned every 2-3

years. We excluded forested and burned plots, yielding a total of 1100 plots in 21 fields (roughly 50 plots per field).

To calculate r_0, we measured change in abundance as a function of successional age (i.e. years since abandonment), relative to initial abundance in year 1. Based on mean observations across fields, starting abundances were set to 0.6% for C4 grasses, 2.3% for C3 grasses, and 11.8% for annuals (n.b. annuals tend to be highly abundant in the seed bank even after many decades of agricultural use (Kitajima and Tilman 1996)). For all plots, we added 0.01% cover (i.e. 0.5 cm^2) to observed abundances, to represent the detection limit and prevent infinite growth rate estimates.

To parameterize the *Levins-OF* model, we used estimates of mean colonization and mortality rates for the three species groups in unburned plots, as calculated in Clark (2017; Ch. 1, Fig. 2a). These resulted in $c_1 = 0.1$ for C4 grasses, $c_2 = 0.3$ for C3 grasses, and $c_3 = 0.4$ for annuals, with $m = 0.02$ for all species groups. For *neutral-OF*, we used the average of these parameters, *i.e.* $c = 0.27$ and $m = 0.02$. We then simulated this three species model at 15 discrete spatial extents (0.01%, 0.02%, 0.04%, 0.06%, 0.09%, 0.12%, 0.20%, 0.30%, 0.49%, 0.72%, 1.00%, 1.96%, 3.06%, 4.00%, and 8.12% of the total 100-by-100-unit grid – numbers were chosen to ensure an integer number of grid cells).

Appendix D. Model identification

To test whether fingerprints derived from r_e and r_0 could be used to identify the underlying model responsible for generating observed dynamics, we constructed density functions summarizing observed inter-simulation variation in r_e and r_0 for each model and extent using the density function in R (2017, version 3.4.2). These were effectively multivariate

probability distributions, that allowed us to calculate the likelihood of a particular observed stability fingerprint given a hypothesized underlying model.

We then used these density functions to determine the likelihood of each simulated stability fingerprint under all five of the candidate models at each extent, $L_{m,n}(s,t)$, where m is the model that generated the fingerprint, and n is the model used to define the density function. Third, to determine how including information from across multiple extents altered likelihoods, we calculated cumulative likelihoods for subsets of nested extents:

$$CL_{m,n}(s,t) = \prod_{j=1}^{s} \prod_{k=1}^{t} L_{m,n}(j,k) \quad \text{(D1)}$$

Finally, we calculated the likelihood that model m matched its own density function relative to the density function of all other models, as:

$$RCL_{m|m}(s,t) = CL_{m,m}(s,t) / \sum_n CL_{m,n}(s,t) \quad \text{(D2)}$$

Eq. (D2) thus describes the relative likelihood of correctly identifying the underlying model responsible for generating each fingerprint.

To quantify the relative explanatory power the *Levins-OF* and *neutral-OF* models relative to the observed old-field stability fingerprints, we calculated $RCL_{m|OF}$, where *OF* is the density function derived from observed old-field dynamics. Because of the limited amount of empirical information, especially at large extents, density functions were calculated based on the sign of fingerprints (i.e. $r_0>0$ or $r_0<0$) observed across bootstrapped replicates of the empirical data, and simulated iterations of the two models rather than raw values.

Interestingly, attempts to identify underlying simulated models based on their stability fingerprints was highly successful (Fig. S15). In general, $RCL_{m|m}$ (i.e. likelihood of correctly matching a simulated fingerprint to the model that generated it) increased with spatial and

temporal extents, and r_0 yielded higher rates of correct identification than r_e. For large spatial and temporal extents, identification success was always greater than 99%.

Importantly, across all tests, we found almost monotonic increases in identification success with increasing spatial and temporal extents (Fig. S15). This property is particularly valuable, as it suggests that increasing the scale of sampling also increases identification success. In contrast, larger extents did not guarantee that measured stability better coincided with long-term persistence of system components (Pimm 1984; Levin 1992; Leibold and Chase 2018). These results support the notion that it may be easier to identify underlying mechanisms that contribute to system dynamics than it is to identify whether or not a system is stable, even given rigorously defined tests and metrics (Ives and Carpenter 2007).

Appendix E. Source code for replicating study

The full source code for replicating the analyses in this manuscript, including all data needed for fitting and testing models, is available in the file `stability_scaling.tar.gz` in the supplement, and at https://github.com/adamtclark/coexistence_scale.

The file "`HPC_iterate_spatial_scale_array.R`" is an automated script that loads and runs all functions needed to replicate simulations for the five models described in the main text. Note that for 20,000 iterations, the script requires roughly 2500 processor hours, and is therefore written to be implemented in parallel on a high-performance computing cluster. The script "`plot_iterate_out_array.R`" reproduces all summary figures related to these simulations.

The file "`HPC_calcualte_empirical_staility_e014.R`" is an automated script that loads and runs all functions needed to replicate the empirical examples and

simulations related to the old-fields. Again, note that for 20,000 iterations, the script requires roughly 400 processor hours, and is therefore written to be implemented in parallel on a high-performance computing cluster. The script "`plot_stability_e014.R`" reproduces all summary figures related to these simulations.

The script "`plot_matchcor_out_array.R`" runs all analyses and reproduces all figures related to identifying simulated models, "`plot_matchcor_out_array_emp.R`" runs all analyses and reproduces all figures related to identifying empirical dynamics, "`plot_example_timeseries.R`" reproduces figures related to examples of time-series and stability analyses, and "`plot_disturbance_example.R`" shows the example of persistence for the *disturbance* model shown in Fig. S1. The script and "`plot_Levins_nodemstoch.R`" runs simulations of the master equations for the *Levins*, *disturbance*, *RPS*, and *neutral* models (i.e. dynamics without demographic stochasticity), and plots the resulting stability fingerprints.

The script "`plot_stability_e014.R`" plots the mean observed and simulated dynamics for the old fields. The script "`plot_demstoch_vs_scale.R`" runs analyses and plots figures demonstrating the strength of demographic stochasticity as a function of spatial extent. The scripts "`plot_psf_example_negtivetransient.R`" and "`plot_psf_example_positive.R`" run analyses and plot figures related to examples of alternate stable states in the *PSF* model. The directory "`util`" and the script "`run_metapopulation_wrapper.R`" contain annotated helper functions for the other scripts.

Lastly, several "`.c`" scrips are compiled and run automatically by the R scripts described above, and include the source code for implementing Gillespie's method with our five models.

See “run_metapopulation_wrapper.R”, and comments within the individual files, for more details.

Literature Cited

Adler, P. B., A. Fajardo, A. R. Kleinhesselink, and N. J. B. Kraft. 2013. Trait-based tests of coexistence mechanisms. Ecology Letters 16:1294–1306.

Allesina, S., and J. M. Levine. 2011. A competitive network theory of species diversity. Proceedings of the National Academy of Sciences of the USA 108:5638–5642.

Anderson, H. M., V. Hutson, and R. Law. 1992. On the Conditions for Permanence of Species in Ecological Communities. American Naturalist 139:663–668.

Armstrong, R. A., and R. McGehee. 1980. Competitive Exclusion. American Naturalist 115:151.

Arnoldi, J.-F., A. Bideault, M. Loreau, and B. Haegeman. 2018. How ecosystems recover from pulse perturbations: A theory of short- to long-term responses. Journal of Theoretical Biology 436:79–92.

Arnoldi, J.-F., M. Loreau, and B. Haegeman. 2016. Resilience, reactivity and variability: A mathematical comparison of ecological stability measures. Journal of Theoretical Biology 389:47–59.

Barabás, G., R. D'Andrea, and S. M. Stump. 2018. Chesson's coexistence theory. Ecological Monographs 88:277–303.

Barabás, G., and A. Ostling. 2013. Community robustness in discrete-time periodic environments. Ecological Complexity 15:122–130.

Black, A. J., and A. J. McKane. 2012. Stochastic formulation of ecological models and their applications. Trends in Ecology & Evolution 27:337–345.

Carpenter, S., B. Walker, J. M. Anderies, and N. Abel. 2001. From Metaphor to Measurement: Resilience of What to What? Ecosystems 4:765–781.

Chesson, P. 2000*a*. Mechanisms of Maintenance of Species Diversity. Annual Review of Ecology and Systematics 31:343–366.

———. 2000*b*. General Theory of Competitive Coexistence in Spatially-Varying Environments. Theoretical Population Biology 58:211–237.

———. 2003. Quantifying and testing coexistence mechanisms arising from recruitment fluctuations. Theoretical Population Biology 64:345–357.

———. 2017. AEDT: A new concept for ecological dynamics in the ever-changing world. PLoS Biology 15:e2002634.

———. 2018. Updates on mechanisms of maintenance of species diversity. Journal of Ecology 106:1773–1794.

Chesson, P., and N. Huntly. 1997. The Roles of Harsh and Fluctuating Conditions in the Dynamics of Ecological Communities. American Naturalist 150:519–553.

Chesson, P., and J. J. Kuang. 2008. The interaction between predation and competition. Nature 456:235–238.

Chesson, P. L., and S. Ellner. 1989. Invasibility and stochastic boundedness in monotonic competition models. Journal of Mathematical Biology 27:117–138.

Clark, A. T. 2017. Constraints and tradeoffs: Toward a predictive, mechanism-based understanding of ecological communities. University of Minnesota, St. Paul, MN.

Clark, A. T., M. Detto, H. C. Muller-Landau, S. A. Schnitzer, S. J. Wright, R. Condit, and S. P. Hubbell. 2018. Functional traits of tropical trees and lianas explain spatial structure across multiple scales. Journal of Ecology 106:795–806.

Clark, A. T., J. M. H. Knops, and D. Tilman. 2019. Contingent factors explain average divergence in functional composition over 88 years of old field succession. Journal of Ecology 107: 545-558.

Davis, M. B., and R. G. Shaw. 2001. Range shifts and adaptive responses to Quaternary climate change. Science 292:673–679.

Detto, M., and H. C. Muller-Landau. 2016. Rates of formation and dissipation of clumping reveal lagged responses in tropical tree populations. Ecology 97:1170–1181.

Donohue, I., H. Hillebrand, J. M. Montoya, O. L. Petchey, S. L. Pimm, M. S. Fowler, K. Healy, et al. 2016. Navigating the complexity of ecological stability. Ecology Letters 19:1172–1185.

Donohue, I., O. L. Petchey, J. M. Montoya, A. L. Jackson, L. McNally, M. Viana, K. Healy, et al. 2013. On the dimensionality of ecological stability. Ecology Letters 16:421–429.

Durrett, R., and S. Levin. 1994. The importance of being discrete (and spatial). Theoretical Population Biology 46:363–394.

Ellner, S. P., R. E. Snyder, P. B. Adler, and G. Hooker. 2019. An expanded modern coexistence theory for empirical applications. Ecology Letters 22: 3-18.

Gleeson, S. K., and D. Tilman. 1990. Allocation and the transient dynamics of succession on poor soils. Ecology 71:1144–1155.

Grilli, J., G. Barabás, M. J. Michalska-Smith, and S. Allesina. 2017. Higher-order interactions stabilize dynamics in competitive network models. Nature 548:210–213.

Grimm, V., and C. Wissel. 1997. Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. Oecologia 109:323–334.

Hillebrand, H., S. Langenheder, K. Lebret, E. Lindström, Ö. Östman, and M. Striebel. 2018. Decomposing multiple dimensions of stability in global change experiments. Ecology Letters 21:21–30.

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

Inouye, R. S., N. J. Huntly, D. Tilman, J. R. Tester, M. Stillwell, and K. C. Zinnel. 1987. Old-field succession on a Minnesota sand plain. Ecology 68:12–26.

Ives, A. R., and S. R. Carpenter. 2007. Stability and diversity of ecosystems. Science 317:58–62.

Lawton, J. H. 1999. Are there general laws in ecology? Oikos 84:177.

Lehman, C., A. Keen, and R. Barnes. 2012. Trading Space for Time: Constant-Speed Algorithms for Managing Future Events in Scientific Simulations. Page 1 *in* Proceedings of the International Conference on Scientific Computing (CSC). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Leibold, M. A., and J. M. Chase. 2018. Metacommunity Ecology, Volume 59. Princeton University Press, Princeton NJ.

Levin, S. A. 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73:1943.

Levine, J. M., J. Bascompte, P. B. Adler, and S. Allesina. 2017. Beyond pairwise mechanisms of species coexistence in complex communities. Nature 546:56–64.

Levins, R. 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America 237–240.

MacArthur, R. H. 1970. Species packing and competitive equilibrium for many species. Theoretical Population Biology 1:1–11.

May, R. M. 1973. Stability and complexity in model ecosystems. Monographs in Population Biology 6. Princeton University Press, Princeton NJ.

———. 1974. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. Science (New York, N.Y.) 186.

Meyer, K., A. Hoyer-Leitzel, S. Iams, I. Klasky, V. Lee, S. Ligtenberg, E. Bussmann, et al. 2018. Quantifying resilience to recurrent ecosystem disturbances using flow–kick dynamics. Nature Sustainability 1:671–678.

Murdoch, W. W. 1994. Population regulation in theory and practice. Ecology 75:271.

Pimm, S. L. 1984. The complexity and stability of ecosystems. Nature 307:321–326.

Ricklefs, R. E. 2008. Disintegration of the ecological community. American Naturalist 172:741–750.

Saavedra, S., R. P. Rohr, J. Bascompte, O. Godoy, N. J. B. Kraft, and J. M. Levine. 2017. A structural approach for understanding multispecies coexistence. Ecological Monographs 87:470–486.

Sheil, D., and R. M. May. 1996. Mortality and recruitment rate evaluations in heterogeneous tropical forests. Journal of Ecology 84:91.

Siepielski, A. M., and M. A. McPeek. 2010. On the evidence for species coexistence: a critique of the coexistence program. Ecology 91:3153–3164.

Stommel, H. 1963. Varieties of Oceanographic Experience: The ocean can be investigated as a hydrodynamical phenomenon as well as explored geographically. Science 139:572–576.

Suding, K. N., W. Stanley Harpole, T. Fukami, A. Kulmatiski, A. S. MacDougall, C. Stein, and W. H. van der Putten. 2013. Consequences of plant-soil feedbacks in invasion. Journal of Ecology 101:298–308.

Tilman, D. 1990. Constraints and tradeoffs – toward a predictive theory of competition and succession. Oikos 58:3–15.

———. 1994. Competition and biodiversity in spatially structured habitats. Ecology 75:2–16.

Tilman, D. 2004. Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. Proceedings of the National Academy of Sciences of the USA 101:10854–10861.

Tilman, D., R. May, C. Lehman, and M. Nowak. 1994. Habitat destruction and the extinction debt. Nature 371:65–66.

Turelli, M. 1980. Niche overlap and invasion of competitors in random environments II. The effects of demographic stochasticity. Pages 119–129 *in* W. Jäger, H. Rost, and P. Tautu, eds. Biological Growth and Spread (Vol. 38). Springer Berlin Heidelberg, Berlin, Heidelberg.

———. 1981. Niche overlap and invasion of competitors in random environments I. Models without demographic stochasticity. Theoretical Population Biology 20:1–56.

———. 1986. Stochastic community theory: A partially guided tour. Pages 321–339 *in* T. G. Hallam and S. A. Levin, eds. Mathematical Ecology (Vol. 17). Springer Berlin Heidelberg, Berlin, Heidelberg.

Zelnik, Y. R., J.-F. Arnoldi, and M. Loreau. 2018. The impact of spatial and temporal dimensions of disturbances on ecosystem stability. Frontiers in Ecology and Evolution 6:224.

References Cited Only in the Online Enhancements

Bever, J. D., K. M. Westover, and J. Antonovics. 1997. Incorporating the soil community into plant population dynamics: The utility of the feedback approach. Journal of Ecology 85:561.

Borer, E. T., W. S. Harpole, P. B. Adler, E. M. Lind, J. L. Orrock, E. W. Seabloom, and M. D. Smith. 2014. Finding generality in ecology: a model for globally distributed experiments. Methods in Ecology and Evolution 5:65–73.

Chesson, P. 1990. MacArthur's consumer-resource model. Theoretical Population Biology 37:26–38.

Clark, A. T., and C. Neuhauser. 2018. Harnessing uncertainty to approximate mechanistic models of interspecific interactions. Theoretical Population Biology 123:35–44.

Cushing, E. J. 1963. Late-Wisconsin pollen stratigraphy in east-central Minnesota. University of Minnesota, Minneapolis, MN, USA.

Dormann, C. F. 2008. On community matrix theory in experimental plant ecology. Web Ecology 8:108–115.

Kitajima, K., and D. Tilman. 1996. Seed banks and seedling establishment on an experimental productivity gradient. Oikos 76:381.

Knops, J. 2018. Plant species percent cover data: Successional dynamics on a resampled chronosequence. Environmental Data Initiative. https://doi.org/10.6073/pasta/aa029df8f7a6091ea879ceb5c6673963.

Muller-Landau, H. C. 2010. The tolerance-fecundity trade-off and the maintenance of diversity in seed size. Proceedings of the National Academy of Sciences of the USA 107:4242–4247.

R. Development Core Team,. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing.

Tilman, D. 1987. Secondary succession and the pattern of plant dominance along experimental nitrogen gradients. Ecological Monographs 57:189–214.

Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie, and E. Siemann. 1997. The influence of functional diversity and composition on ecosystem processes. Science 277:1300–1302.

Weisser, W. W., C. Roscher, S. T. Meyer, A. Ebeling, G. Luo, E. Allan, H. Beßler, et al. 2017. Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions. Basic and Applied Ecology 23:1–73.

Figure legends

Figure 1: Because of a combination of physical limitations, financial barriers, and historical legacy, the spatial and temporal scales at which one can reasonably observe and manipulate ecological systems do not always correspond to those that are useful for studying coexistence. Consider a time-series comprised of three observations, in each of three different types of communities. **(A)** Samples of microbial communities might include millions of individuals, and hundreds of generations can pass between surveys. **(B)** Herbaceous plant communities can contain hundreds of individuals per square meter, and measurements of species-level biomass are sufficiently destructive and time-consuming that only a small fraction of the total landscape can be surveyed. **(C)** For tree communities, survey plots might only contain a few individuals, but dynamics can be very slow, playing out over decades or centuries.

Figure 2: Example dynamics of each model at the maximum spatial extent (i.e. across all simulated sites). Each row shows results for a single model. Colored lines show species abundances. Vertical dashed lines and arrows show perturbation events for perturbation tests **(a,c,e,g,i)**, or removals and invasion events for invasion test **(b,d,f,h,j)**, always for the red species. Colored dashed lines in **(a,c,e,g,i)** show trajectory in the absence of perturbation. Horizontal lines in **(a,b)** show equilibria from Eq. (2), and in **(c,d)** show mean abundances predicted from temporally averaged mortality rates (see Appendix *A.2* for details). Vertical dotted lines in **(c,d)** show disturbance events.

Figure 3: Stability fingerprints summarizing median results across spatial and temporal extents, based on 20,000 iterations of each model. Columns show different metrics, rows show different models. Colors show median value for maximum r_e **(a-e)** or minimum r_0 **(f-j)** observed across species. Cooler colors indicate greater measured stability (recall that negative r_e or

positive r_0 indicate stable coexistence). Bright red regions indicate extinction of at least one species in more than one third of simulations. Dark grey shading shows regions where between-simulation variability is sufficiently large that >50 observations would be required to detect a significant difference from zero. Shading is semi-transparent, so that the sign of the index is still visible. For more information on the distribution of results, see Fig. S4.

Figure 4: Stability fingerprints for successional dynamics in old-fields at Cedar Creek generated from approximated invasion statistic r_0. Rows show results for different species groups, columns show fingerprints for observed dynamics **(a-c)**, simulated results from *Levins-OF* **(d-f)**, and simulated results from *neutral-OF* **(g-i)**. Colors show median stability metrics, either from 20,000 bootstrapped samples drawn from old-field surveys **(a-c)**, or 20,000 simulations **(d-i)**. Contour lines show relative cumulative likelihood of observed pattern given the *Levins-OF* **(d-f)** vs. *neutral-OF* **(g-i)** models (defined as $RCL_{m|OF}$ in the supplement – see *Appendix D* and Fig. S16 for more details). Note that unlike Eq. (5) and Fig. (3f-j), r_0 is multiplied by temporal extent t to make patterns at larger temporal extents clearer (Sheil and May 1996).

Figure 1

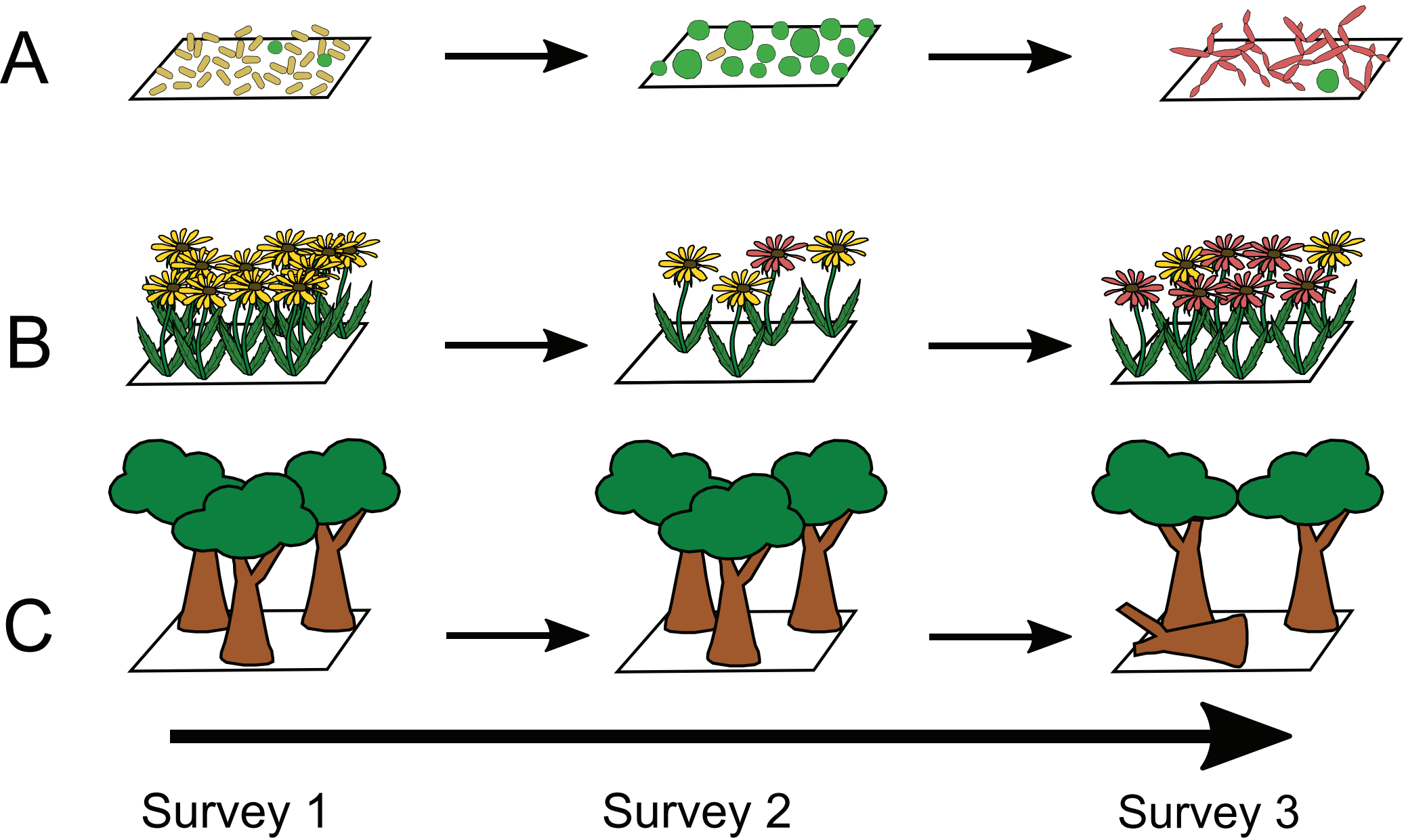


Figure 2

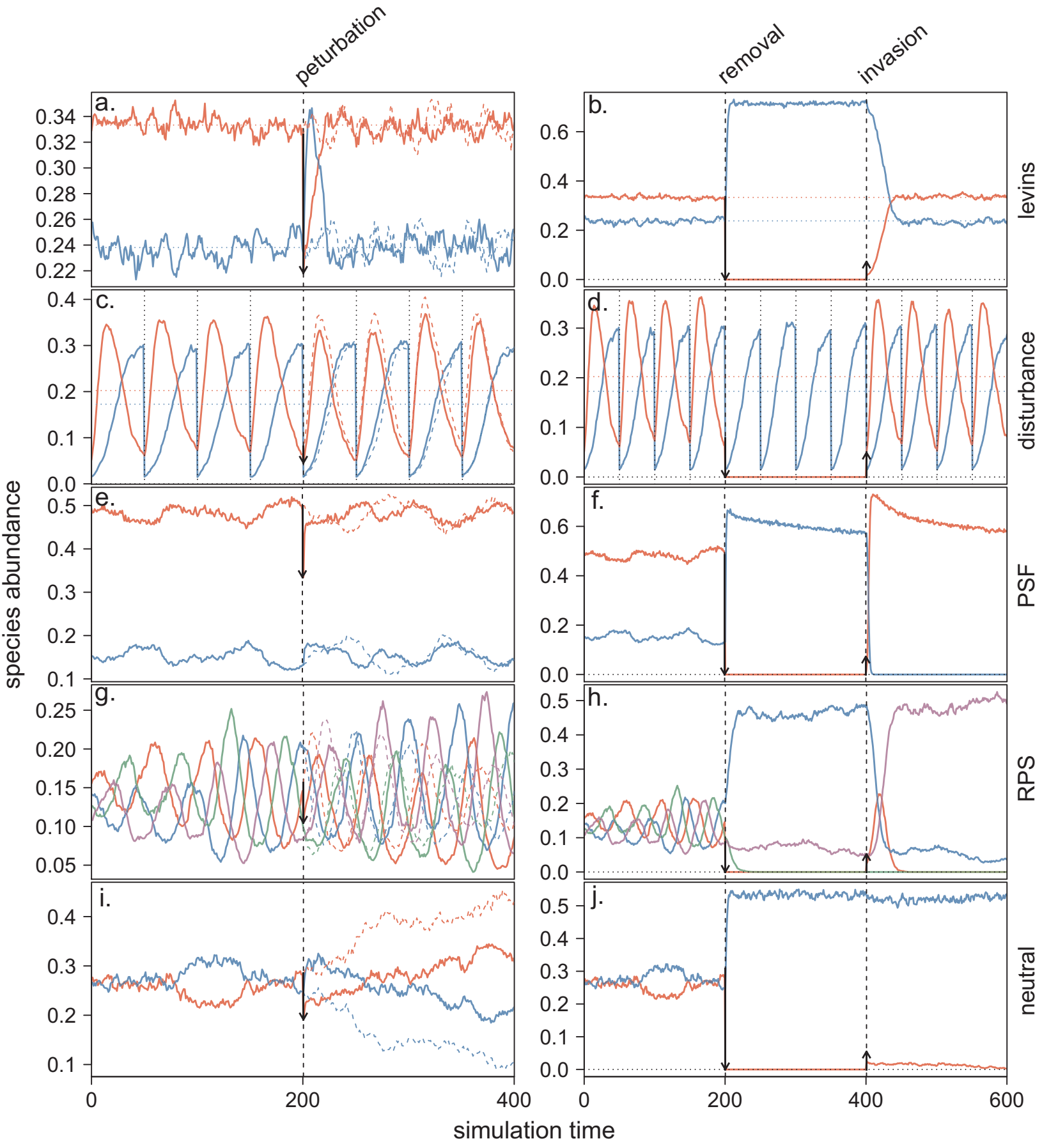


Figure 3

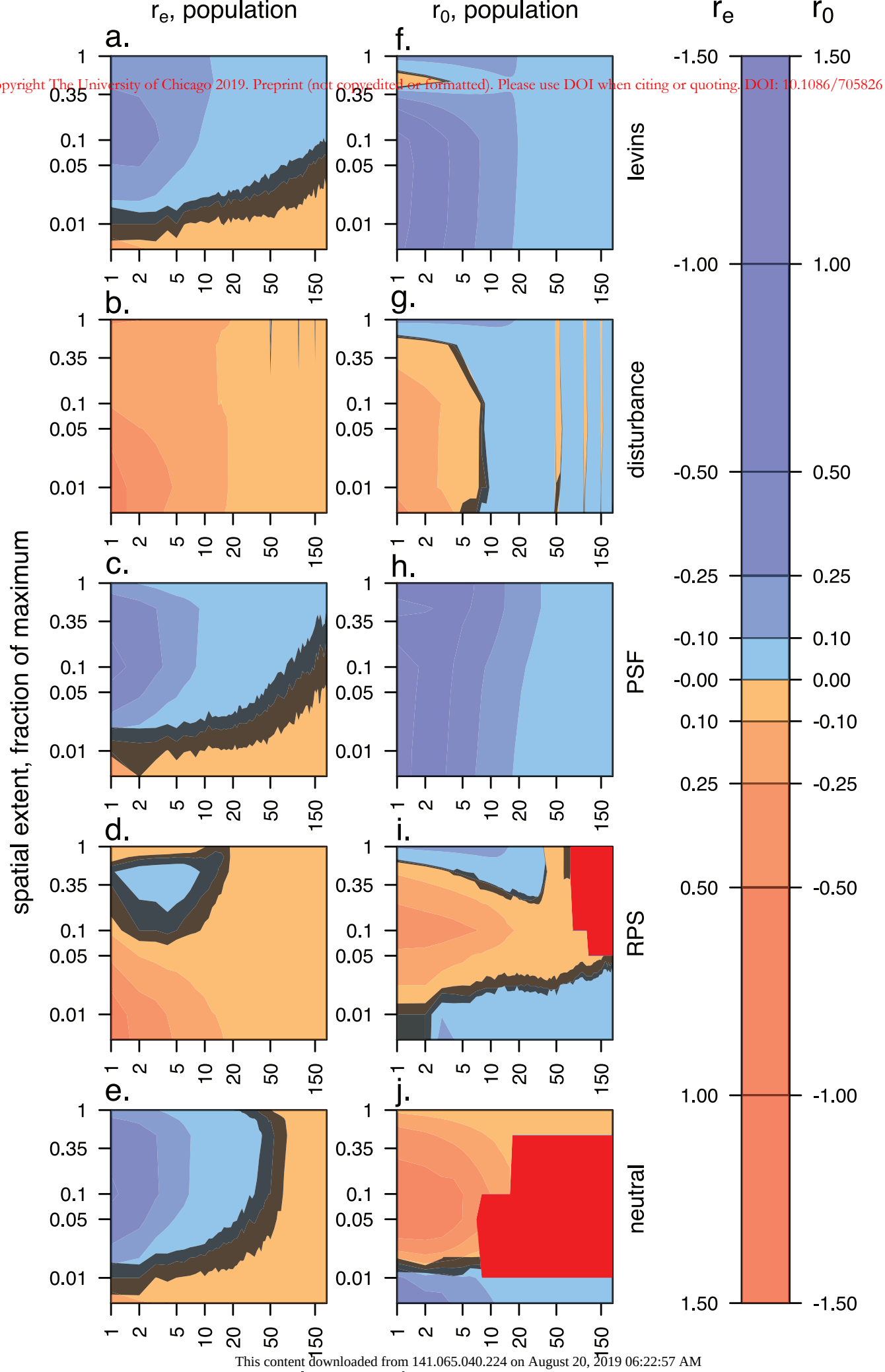
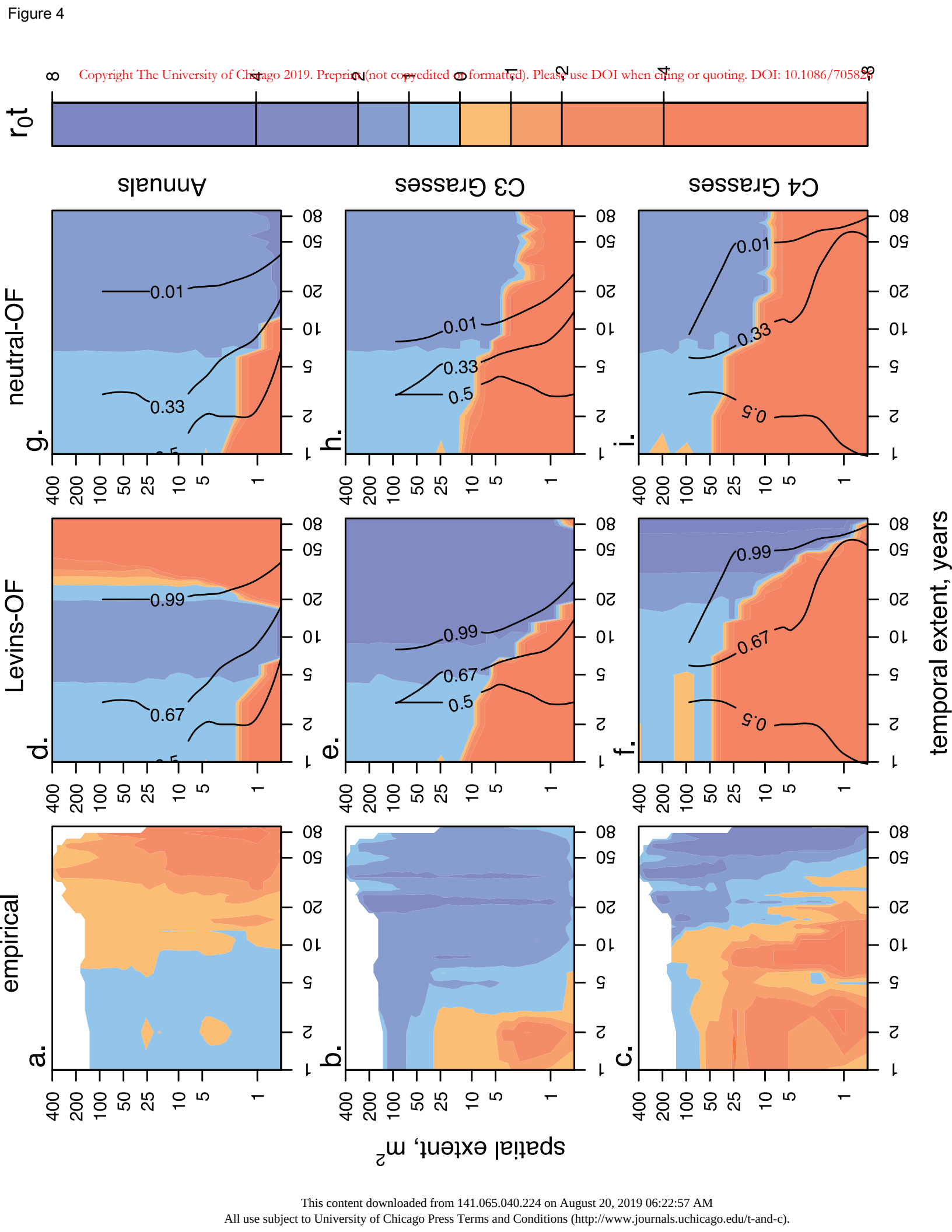


Figure 4



Supplement for:
Scale both confounds and informs characterization of species coexistence in empirical systems

Authors:
Adam Clark*
Helmut Hillebrand
W. Stanley Harpole

*Corresponding author: *adam.tclark@gmail.com*

Contents:

I. Model parameterization

I.1: Model derivations

At large spatial extents, effects of demographic stochasticity can be ignored, and dynamics for *Levins* follow the master equation form described in Eq. (1) in the main text (equation reproduced here for convenience):

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = c_i p_i\left(1 - \sum_{j \le i} p_j\right) - m_i p_i - p_i \sum_{j<i} c_j p_j \qquad (1)$$

where p_i is the fraction of sites occupied by species i, c_i and m_i are per-capita colonization and mortality rates, respectively, and species indices are ordered by competitive ability, such that $j<i$ implies that species j is competitively superior to species i. Note that this is identical to the Levins model, *sensu* Tilman (1994), and is the master equation for our stochastic implementation of the model.

In the model *disturbance*, disturbance events occur at regular time intervals, killing off fixed fraction d_i of the individuals of each species. Between disturbance events, dynamics follow the same form as *Levins*. Thus, dynamics at large extents can be expressed as:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = c_i p_i\left(1 - \sum_{j \le i} p_j\right) - m_i p_i - p_i \sum_{j<i} c_j p_j \qquad \text{if } t \bmod D \neq 0 \qquad \text{(S1a)}$$

$$p_{i|D} = (1 - d_i) p_{i|D'} \qquad \text{if } t \bmod D = 0 \qquad \text{(S1b)}$$

where $p_{i|D}$ is abundance of species i immediately after the disturbance event, $p_{i|D'}$ is the abundance immediately before, and the modulo operator "x mod y" takes the remainder of x/y (thus, t mod $D = 0$ only if t is a multiple of D). Note that rate d_i is consequently a rate in discrete-time, whereas m_i is a rate in continuous-time, which is why the log transformation $\log(1 - d_i)/D$ is required in our calculation of mean mortality rate for the master equation.

For *RPS*, we simulated a system of four species in a competitive loop, such that species i is competitively superior to species j, j is competitively superior to species k, k is competitively superior to species l, and l is competitively superior to species i. All other species pairs (i.e. i vs k; j vs. l; and k vs. i) are competitively equivalent, meaning that either

can displace the other. Demographic rates are equal across all species, yielding a system of equations of the form:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = cp_i(1 - p_i - p_l) - mp_i - cp_i(p_k + p_l) \quad \text{(S2a)}$$
$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = cp_j(1 - p_j - p_i) - mp_j - cp_j(p_i + p_l) \quad \text{(S2b)}$$
$$\frac{\mathrm{d}p_k}{\mathrm{d}t} = cp_k(1 - p_k - p_j) - mp_k - cp_k(p_i + p_j) \quad \text{(S2c)}$$
$$\frac{\mathrm{d}p_l}{\mathrm{d}t} = cp_l(1 - p_l - p_k) - mp_l - cp_l(p_j + p_k) \quad \text{(S2d)}$$

For *neutral*, dynamics are identical to Eq. (1), except that species share the same demographic rates, and do not have direct competitive interactions. Thus, given a community of *n* species, dynamics can be written as:

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = cp_i\left(1 - \sum_{j=1}^{n} p_j\right) - mp_i \quad \text{(S3)}$$

Note that dynamics match those of Hubbel's neutral model (Hubbell 2001), except that we include time-explicit colonization and mortality, rather than following the zero-net-sum assumption (i.e. that empty sites are immediately filled). Thus, fewer than 100% of sites are occupied at any given moment in time.

Lastly, for *PSF*, we used a model presented by Suding *et al*. (2013). Spatial structure and demography are similar to that for *Levins*, except that there are no direct competitive interactions among species (i.e. species cannot displace one another). Rather, species influence local conditions in sites that they occupy in ways that either hurt themselves and favor their competitors (negative feedbacks), or favor themselves and hurt their competitors (positive feedbacks). Given two species *i* and *j* ("exotic" and "native", respectively, in the original model), survival probability in any given time-step is calculated as:

$$s_i = \frac{(1 - sign(f_i)S)}{2} - m_i \quad \text{(S4a)}$$
$$s_j = \frac{(1 + sign(f_j)S)}{2} - m_j \quad \text{(S4b)}$$

where f_i indicates the strength and direction of feedbacks, S is the local site condition, $sign(f_i)$ indicates whether f_i is positive or negative, and m is the background rate of stochastic mortality. Local site state S varies from –1 to 1.

In our implementation of the model, soil state declines by 1 percentage point every time step that a site is occupied by species *i*, whereas every time step that it is occupied by species *j*, *S* increases by 1 percentage point, with respective maxima and minima determined by –abs(f_i) and +abs(f_j). In unoccupied sites, *S* approaches 0 by 1 percentage point per time step. Note that with negative feedbacks for both species, each species increases its own mortality in sites that it occupies, and decreases mortality for its competitor. Lastly, in our implementation of the model, probability of establishment in empty sites is determined following the formula:

$$p_i = (1 - p_0)\frac{c_i s_i}{c_i s_i + c_j s_j} \quad \text{(S5a)}$$

$$p_j = (1 - p_0)(1 - p_i) \quad \text{(S5b)}$$

where c_i is the seed production of species *i* relative to *j*, and p_0 is the probability that neither species successfully recruits into the patch. Note that this equation differs slightly from that in the original publication, for which $p_0=0$.

I.2: Analytical stability properties

Recall that unless noted otherwise, all discussions here and in the main text relate to population stability (i.e. the stability of the least stable species in the community, meaning the species with the highest r_e or the smallest r_0). Nevertheless, results for community stability (i.e. stability of total community biomass) are similar to those for population stability (see Fig. S12 for results for community stability).

As shown in the main text, we can solve for the analytical expectation of species equilibrium abundances by setting Eq. (1) to zero, and solving for p_i, which yields Eq. (2) from the main text (equation reproduced here for convenience):

$$p_i^* = 1 - \frac{m_i}{c_i} - \sum_{j<i} p_j^* \left(1 + \frac{c_j}{c_i}\right) \text{ if } p_i^* > 0; = 0 \text{ otherwise} \quad (2)$$

This estimate is accurate at spatial extents that are large enough to avoid strong effects from demographic stochasticity (e.g. thousands of cells). At these scales, *Levins* is globally stable

around these equilibrium values (i.e. they are approached from any non-trivial starting point) (Tilman 1994). As discussed in the main text, species achieve positive equilibrium biomass given a sufficiently strong trade-off between competitive and colonization ability. Note that because the equilibrium abundance of each species is globally stable, community abundance and composition are necessarily also stable.

For *disturbance*, stability properties are identical to those in *Levins* provided that all species are predicted to have positive abundances following Eq. (2) based on their time-averaged demographic rates (again, assuming that the grid is large enough to avoid large effects of demographic stochasticity). By "time-averaged", we mean the average rates when considered across time periods with and without disturbances. These rates can be calculated as $m_i - \log(1 - d_i)/D$, where m_i is the mortality rate in the absence of disturbance, d_i is the fraction of individuals destroyed during disturbances, and D is the time between disturbance events (n.b. c_i parameters are not influenced by disturbance). Given a trade-off between species competitive abilities and their tolerance to disturbance events (i.e. d_i), stable coexistence is therefore possible even among combinations of species that would competitively exclude one another in the absence of disturbance (Fig. S1). Average abundances of species through time in *disturbance* can be calculated exactly from the standard *Levins* model using Eq. (2), by setting the mortality rates to equal the time-averaged rate. Nevertheless, *disturbance* displays substantial transient dynamics between disturbance events which differ from the average abundances, and significantly influence perceived system stability (as discussed in the main text).

For *PSF*, stability properties have been extensively analyzed in Suding *et al.* (2013). Locally stable coexistence occurs given sufficiently strong negative feedbacks for both species. Under these circumstances, species fall into fixed oscillatory cycles. In our model, these cycles are compensatory, such that total community abundance remains at a fixed level (Fig. 1e-f in the main text). Note, however, that the model is subject to alternate stable states.

Given positive feedbacks between species and their environment (i.e. such that positive effects build up over time as species occupy a site), these alternate stable states occur from almost any starting point (Fig. S2). Given negative feedbacks, as we explore in the main text (i.e. where negative effects build up over time as species occupy a site), alternate stable states are possible, but only if species have spent a prolonged time growing in monoculture before they are invaded (Fig. S3). These alternate stable states occur because negative feedbacks become so intense that the resident species is no longer able to resist competitive exclusion by a reintroduced competitor. Although this result does not accord directly with the classical plant-soil-feedback model of Bever *et al.* (1997) (which does not predict alternate stable states in cases with negative feedbacks), this is not due to a major conceptual difference between our models, but rather because we consider stronger negative feedbacks (n.b. an equivalent result in their model would arise from feedbacks that are sufficiently strong to drive the phase cycle in their Fig. 2 into one of the zero boundaries).

For *neutral*, recall that because all individuals share the same demographic rates, dynamics for total community abundance can be treated as a single species in the *Levins* model, with equilibrium abundance (equation reproduced here for convenience):

$$\sum_{i=1}^{n} p_i^* = 1 - \frac{m}{c} \quad (3)$$

Thus, though population abundances and composition are purely subject to ecological drift and should therefore not be indicated as stable by any of our metrics (Hubbell 2001), total community abundance has the same stability properties as it does in *Levins*. Note that because of this drift, long-term coexistence is not possible.

Lastly, for *RPS*, dynamics follow a four-species competitive loop, *sensu* Fig. 1b in Grilli *et al.* (2017). For communities including all four species, all species are inhibited by a competitor, leading to a neutral, oscillatory system (Allesina and Levine 2011; Grilli et al. 2017). Because the system is neutrally stable, even small perturbations can lead to changes in the period and amplitude of oscillations, and long-term coexistence is not possible due to the

effects of demographic stochasticity. However, because oscillations are compensatory, total community abundance remains relatively constant regardless of perturbation or invasion events (e.g. Fig. 1g-h in the main text). Thus, community abundance is likely to be indicated as stable following any of our metrics. Conversely, our perturbation and invasion metrics should indicate instability for population abundance and composition.

For communities in *RPS* composed of a single species, or any pair of species that have equal competitive abilities (i.e. *i* vs *k*; *j* vs. *l*; and *k* vs. *i*), species dynamics in *RPS* reduce to the *neutral* model, since individuals are equivalent in terms of demographic rates. Any combinations of species that include unconstrained competitive interactions (e.g. *i* and *j* alone; or *i*, *j*, and *l*) is inherently unstable, as it results in asymmetrical competitive interactions among otherwise identical species, which leads to rapid competitive displacement by the unconstrained species. For example, given competition between species *i* and *j*, equilibrium abundances predicted from Eq. (2) are:

$$p_i^* = 1 - \frac{m}{c} \tag{S6a}$$

$$p_j^* = 1 - \frac{m}{c} - p_i^*\left(1 + \frac{c}{c}\right) = 1 - \frac{m}{c} - 2\left(1 - \frac{m}{c}\right) = -(1 - \frac{m}{c}) \tag{S6b}$$

which implies competitive exclusion of species *j* (since $m < c$). Note that removal of a single species can therefore initiate a cascade of events, which ultimately prevent its successful re-introduction. For example, removing species *i* leaves growth of species *j* unchecked, thereby driving species *k* extinct. Without species *k*, species *l* is able to grow unchecked, ultimately excluding *i* if it is re-introduced to the community (see Fig. 1h in the main text).

I.3: Parameter values

For all simulations, including all figures in the main text and supplement, we used the following parameter values:

1. *Levins:*
    a. $c_1 = 0.45$; $c_2 = 1.05$
    b. $m_1 = m_2 = 0.3$

2. *disturbance:*
    a. $c_1 = 0.435$; $c_2 = 0.600$
    b. $m_1 = m_2 = 0.3$
    c. disturbance frequency = 1 per 50 time-steps
    d. d_1 = 95% mortality; d_2 = 0% mortality
3. *PSF*:
    a. $c_1 = 1/c_2 = 1.5$
    b. $f_1 = f_2 = -80\%$
    c. $m_1 = m_2 = 5\%$
    d. initial soil state = 0
    e. $p_0 = 50\%$
4. *RPS:*
    a. $c = 0.48$ for all species
    b. $m = 0.224$ for all species
5. *neutral:*
    a. $c = 1.5$ for all species
    b. $m = 0.7$ for all species

For each set of extents, we simulated 20,000 iterations of each model. We simulated 300 time-steps before conducting manipulations, and 200 post-manipulation time-steps, including an additional 200 time-step burn-in period for invasion tests to account for transient dynamics.

I.4: Simulation procedure

All simulations were carried out on a 100-by-100 cell rectangular grid. In the simulations, each of the 10,000 resulting cells is either empty, or holds a single individual of a single species. We assume global dispersal in all models (i.e. all available cells are equally likely to be reached by propagules, regardless of distance). Results in the main text are based on the median of 20,000 simulated iterations of each model, at each of seven spatial extents (0.5%, 1%, 5%, 10%, 50%, 75%, and 100% of the maximum extent).

To calculate stability metrics, we first ran each model for a 300 time-step burn in period to avoid transient dynamics. We then paused the simulation, and branched it into three classes of runs: *null*, *perturbation*, and *invasion*. These branches all retained the same spatial and species layout, random number seeds, and projected future events (for the Gillespie

simulations – see *Appendix II* in the supplement), meaning that they would produce identical time series if left un-disturbed. For the *null* test, only a single branch was simulated, no additional changes were made, and dynamics were simulated for an additional 200 time-steps. For the *perturbation* tests, one branch was simulated per species. For each branch, the abundance of a single species was reduced by 20% at the beginning of the simulation, and the simulation was then run for an additional 200 time-steps to observe recovery. For the *invasion* tests, one branch was simulated per species. For each branch, the abundance of a single species was reduced to zero, and the remaining community was allowed to adjust to the change for 200 time-steps. Then, the removed species was re-introduced at an abundance equal to 5% of the total unoccupied sites.

I.5: Justification of perturbation metric

Perturbation analyses are usually conducted by calculating eigenvalues of the Jacobian matrix that describes interactions among system components (typically individual species), evaluated at equilibrium (May 1973). However, because eigenvalues represent a linear approximation of a high-dimensional system, it is impossible to characterize their response to noise without specific assumptions about the functional form of interactions among system components (e.g. mechanism of competition) and the statistical distribution of the noise (e.g. Gaussian, log-normal) (Chesson 1990; Anderson et al. 2010). Consequently, even minor observation error can lead to enormous uncertainty in eigenvalue estimates (Dormann 2008; Clark and Neuhauser 2018).

We therefore use a metric of response to perturbation measured at the level of individual species, rather than at the level of eigenvectors – i.e. r_e from Eq. (4) in the main text. Though our metric r_e can potentially be confounded by transient dynamics over short time periods (Fig. C1 in Arnoldi et al. 2018), its sign is guaranteed to match that of the dominant Eigenvalue over sufficiently long temporal extents, as both metrics correspond to

asymptotic stability of the least stable species. Moreover, note that r_e is formally an approximation of the most positive Lyapunov exponent of our system, which is a very general metric commonly used to test for stability in complex dynamical systems.

II. Implementation of Gillespie's method

II.1: Primer on method

As noted in the main text, we used an implementation of Gillespie's method to simulate model dynamics. A major advantage of this method is that it perfectly matches analytical expectations (e.g. Eq. 1) for large grids at the maximum spatial extent, but produces discrete, spatial results at smaller scales. By "discrete", we mean that birth and death events for individuals are explicitly considered, which allows us to account for temporal structure caused by demographic stochasticity (Durrett and Levin 1994). By "spatial", we mean that individuals occupied specific locations in a grid, which introduces spatial structure into our model. Note, however, that because we consider only interactions at the scale of an individual grid cell, and include global dispersal for all simulations, the specific spatial arrangement of individuals is not relevant for our study.

To implement Gillespie's method, we used the procedure and source code described in Lehman *et al.* (2012). This implementation simulates dynamics event-by-event (i.e. rather than time-step-by-time-step) by forecasting the expected time of future events (e.g. mortality or colonization) using an exponential waiting time distribution:

$$t_{event} = \log(-x + 1)/(-r) \quad \text{(S7)}$$

where t_{event} is number of time-steps in the future that the event will occur, x is a random uniform number drawn from over the range (0, 1), and r is a rate constant (e.g. m or c from the equations above). Note that t_{event} need not be an integer.

The method described in Lehman *et al.* (2012) is particularly fast and efficient because it bins future events into pre-sorted locations in computer memory, which allows rapid recall of events without the need to sort or search through large lists. Though describing the specifics of the algorithm is beyond the scope of this paper, we strongly recommend this procedure (available either in the cited paper, or in the "*.c*" files in our source code in *Appendix V* in the supplement) for any readers who wish to implement their own versions of

Gillespie's method, as it typically runs more than three orders of magnitude faster than other versions of the algorithm that we have worked with.

One downside to the method is that it does not lend itself to cases with feedbacks between species and their environments. This is because the state of the environment depends on the precise history of occupancy in the patch, which is itself influenced by the changing state of the environment. This makes it impossible to project the timing of dispersal and mortality events into the future, and instead these must be updated every timestep as is done in standard time-step based simulation methods. This is why we used the previously published model for *PSF* rather than adapting it to match Gillespie's method, as we did with the other models.

II.2: Simulating master equations

To examine the behavior of our models in the absence of demographic stochasticity, we simulated the master equations in Eqs. 1 and S1-S3. These simulations were similar to those in a model by Muller-Landau (2010), in that we included two distinct patch types in each simulation: fraction s of the habitat was subject to observations and manipulations, similar to the observed spatial extent in our spatiotemporally explicit simulations, and fraction $(1 - s)$ of the habitat was not manipulated, but still contributed to and was influenced by dynamics in s due to dispersal.

As an example, for the *Levins* model, we simulated the following two patch system:

$$dp_{i,(1-s)}/dt = c_i(p_{i,s} + p_{i,(1-s)})(1 - s - p_{i,(1-s)}) - m_i\, p_{i,(1-s)} \quad \text{(S8a)}$$

$$dp_{i,s}/dt = c_i(p_{i,s} + p_{i,(1-s)})(s - p_{i,s}) - m_i\, p_{i,s} \quad \text{(S8b)}$$

$$dp_{j,(1-s)}/dt = c_j(p_{j,s} + p_{j,(1-s)})(1 - s - p_{i,(1-s)} - p_{j,(1-s)}) - c_i(p_{i,s} + p_{i,(1-s)})\, p_{j,(1-s)} - m_j\, p_{j,(1-s)} \quad \text{(S8c)}$$

$$dp_{j,s}/dt = c_j(p_{j,s} + p_{j,(1-s)})(s - p_{i,s} - p_{j,s}) - c_i(p_{i,s} + p_{i,(1-s)})\, p_{j,s} - m_j\, p_{j,s} \quad \text{(S8d)}$$

where $p_{i,s}$ is the abundance of species i in manipulated patches, and $p_{i,(1-s)}$ is abundance in unmanipulated sites. Note that abundance is always written in terms of fraction of the total

patches occupied – thus, $p_{i,s} \leq s$, $p_{i,(1-s)} \leq (1 - s)$, and $p_{i,s} + p_{i,(1-s)} \leq 1$. Functional forms for the remaining models were similar, always with two differential equations per species (see source code in *Appendix V* for specific equations relating to each model). We did not simulate a version of the *PSF* without demographic stochasticity because we had no corresponding master equation for the model.

In general, results for the simulations of the master equations were similar to those for the spatiotemporally explicit models that we discuss in the main text (Fig. S11). For r_0, fingerprints from the models with and without demographic stochasticity were almost identical. In contrast, we found fewer instances of positive r_e, especially at small spatial extents. As discussed in the main text, this indicates that the instability detected by perturbation tests at small spatial extents is driven by demographic stochasticity rather than deterministic model behavior (see also Figs. S9-S10).

For three models – *disturbance*, *RPS*, and *neutral* – we generally found more instances of negative r_e across all scales for the master equations than for the stochastic simulations. For *disturbance*, this occurs because populations of both species are driven to small sizes during oscillatory dynamics. Thus, demographic stochasticity leads to relatively large random shifts in the starting abundance of species before they begin the recovery phase of their dynamics (Fig. S8), which can mask the effects of small perturbations even at large spatial extents. For *RPS* and *neutral*, demographic stochasticity influences the effects of perturbation tests – especially at large temporal extents – precisely because the systems are not deterministically stable. Thus, random fluctuations in population sizes compound over time, eventually driving replicate trajectories apart. Interestingly, in the simulations of the master equations, both models "appear" stable based on r_e. As discussed in the main text, this occurs because total community biomass is stable in these models, leading to partial recovery at the population level following small perturbations, even in the absence of stable coexistence (S14).

III. Old-field succession example

A detailed description of the old-field succession experiment at the Cedar Creek Ecosystem Science Reserve is available in Clark (2017), in chapter 1, and in Clark *et al.* (2019). Briefly, Cedar Creek is a reserve run by the University of Minnesota and US Long Term Ecological Research Program, and located in Minnesota, USA (45.4°N, 93.2°W). The site is near the boundary between deciduous forest, boreal forest, and prairie biomes, though prior to European colonization, vegetation was primarily prairie, oak savanna, deciduous forests, and wetlands (Cushing 1963). Mean annual precipitation is a bit below 800 mm per year, with most occurring between April and August. Temperatures are highly variable, with summer highs averaging 27°C, and winter lows averaging -14°C. Soils are very sandy, and plant communities are strongly nitrogen limited (Tilman 1987).

Since 1983, successional dynamics of herbaceous plants have been followed in >20 old-fields at Cedar Creek. Of these, 23 are still surveyed roughly every 5 years for species-level percent cover. Each field includes 100 permanent plots, each 0.5-by-1 m. Plots are arranged into four parallel transects, each containing 25 plots. Transects are spaced 25 m apart, with 1 meter between plots within the same transect. Plots in four fields have experienced heavy afforestation, and since about 2008, half of the plots in most fields have been experimentally burned every 2-3 years. We excluded forested and burned plots, yielding a total of 1100 plots in 21 fields (roughly 50 plots per field).

To calculate r_0, we measured change in abundance as a function of successional age (i.e. years since abandonment), relative to initial abundance in year 1. Based on mean observations across fields, starting abundances were set to 0.6% for C4 grasses, 2.3% for C3 grasses, and 11.8% for annuals (n.b. annuals tend to be highly abundant in the seed bank even after many decades of agricultural use (Kitajima and Tilman 1996)). For all plots, we added 0.01% cover (i.e. 0.5 cm^2) to observed abundances, to represent the detection limit and prevent infinite growth rate estimates.

To parameterize the *Levins-OF* model, we used estimates of mean colonization and mortality rates for the three species groups in unburned plots, as calculated in Clark (2017; Ch. 1, Fig. 2a). These resulted in $c_1 = 0.1$ for C4 grasses, $c_2 = 0.3$ for C3 grasses, and $c_3 = 0.4$ for annuals, with $m = 0.02$ for all species groups. For *neutral-OF*, we used the average of these parameters, *i.e.* $c = 0.27$ and $m = 0.02$. We then simulated this three species model at 15 discrete spatial extents (0.01%, 0.02%, 0.04%, 0.06%, 0.09%, 0.12%, 0.20%, 0.30%, 0.49%, 0.72%, 1.00%, 1.96%, 3.06%, 4.00%, and 8.12% of the total 100-by-100-unit grid – numbers were chosen to ensure an integer number of grid cells).

IV. Model identification

To test whether fingerprints derived from r_e and r_0 could be used to identify the underlying model responsible for generating observed dynamics, we constructed density functions summarizing observed inter-simulation variation in r_e and r_0 for each model and extent using the density function in R (2017, version 3.4.2). These were effectively multivariate probability distributions, that allowed us to calculate the likelihood of a particular observed stability fingerprint given a hypothesized underlying model.

We then used these density functions to determine the likelihood of each simulated stability fingerprint under all five of the candidate models at each extent, $L_{m,n}(s,t)$, where m is the model that generated the fingerprint, and n is the model used to define the density function. Third, to determine how including information from across multiple extents altered likelihoods, we calculated cumulative likelihoods for subsets of nested extents:

$$CL_{m,n}(s,t) = \prod_{j=1}^{s} \prod_{k=1}^{t} L_{m,n}(j,k) \quad \text{(S9)}$$

Finally, we calculated the likelihood that model m matched its own density function relative to the density function of all other models, as:

$$RCL_{m|m}(s,t) = CL_{m,m}(s,t) / \sum_n CL_{m,n}(s,t) \quad \text{(S10)}$$

Eq. (S10) thus describes the relative likelihood of correctly identifying the underlying model responsible for generating each fingerprint.

To quantify the relative explanatory power the *Levins-OF* and *neutral-OF* models relative to the observed old-field stability fingerprints, we calculated $RCL_{m|OF}$, where *OF* is the density function derived from observed old-field dynamics. Because of the limited amount of empirical information, especially at large extents, density functions were calculated based on the sign of fingerprints (i.e. $r_0>0$ or $r_0<0$) observed across bootstrapped replicates of the empirical data, and simulated iterations of the two models rather than raw values.

Interestingly, attempts to identify underlying simulated models based on their stability fingerprints was highly successful (Fig. S15). In general, $RCL_{m|m}$ (i.e. likelihood of correctly

matching a simulated fingerprint to the model that generated it) increased with spatial and temporal extents, and r_0 yielded higher rates of correct identification than r_e. For large spatial and temporal extents, identification success was always greater than 99%.

Importantly, across all tests, we found almost monotonic increases in identification success with increasing spatial and temporal extents (Fig. S15). This property is particularly valuable, as it suggests that increasing the scale of sampling also increases identification success. In contrast, larger extents did not guarantee that measured stability better coincided with long-term persistence of system components (Pimm 1984; Levin 1992; Leibold and Chase 2018). These results support the notion that it may be easier to identify underlying mechanisms that contribute to system dynamics than it is to identify whether or not a system is stable, even given rigorously defined tests and metrics (Ives and Carpenter 2007).

V. Source code for replicating study

The full source code for replicating the analyses in this manuscript, including all data needed for fitting and testing models, is available in the file `stability_scaling.zip` in the supplement, and at https://github.com/adamtclark/coexistence_scale.

The file "`HPC_iterate_spatial_scale_array.R`" is an automated script that loads and runs all functions needed to replicate simulations for the five models described in the main text. Note that for 20,000 iterations, the script requires roughly 2500 processor hours, and is therefore written to be implemented in parallel on a high-performance computing cluster. The script "`plot_iterate_out_array.R`" reproduces all summary figures related to these simulations.

The file "`HPC_calcualte_empirical_staility_e014.R`" is an automated script that loads and runs all functions needed to replicate the empirical examples and simulations related to the old-fields. Again, note that for 20,000 iterations, the script requires roughly 400 processor hours, and is therefore written to be implemented in parallel on a high-performance computing cluster. The script "`plot_stability_e014.R`" reproduces all summary figures related to these simulations.

The script "`plot_matchcor_out_array.R`" runs all analyses and reproduces all figures related to identifying simulated models, "`plot_matchcor_out_array_emp.R`" runs all analyses and reproduces all figures related to identifying empirical dynamics, "`plot_example_timeseries.R`" reproduces figures related to examples of time-series and stability analyses, and "`plot_disturbance_example.R`" shows the example of persistence for the *disturbance* model shown in Fig. S1. The script and "`plot_Levins_nodemstoch.R`" runs simulations of the master equations for the *Levins*, *disturbance*, *RPS*, and *neutral* models (i.e. dynamics without demographic stochasticity), and plots the resulting stability fingerprints.

The script "plot_stability_e014.R" plots the mean observed and simulated dynamics for the old fields. The script "plot_demstoch_vs_scale.R" runs analyses and plots figures demonstrating the strength of demographic stochasticity as a function of spatial extent. The scripts "plot_psf_example_negtivetransient.R" and "plot_psf_example_positive.R" run analyses and plot figures related to examples of alternate stable states in the *PSF* model. The directory "util" and the script "run_metapopulation_wrapper.R" contain annotated helper functions for the other scripts.

Lastly, several ".c" scrips are compiled and run automatically by the R scripts described above, and include the source code for implementing Gillespie's method with our five models. See "run_metapopulation_wrapper.R", and comments within the individual

References

Allesina, S., and J. M. Levine. 2011. A competitive network theory of species diversity. Proceedings of the National Academy of Sciences 108:5638–5642.

Anderson, G. W., A. Guionnet, and O. Zeitouni. 2010. An introduction to random matrices. Cambridge studies in advanced mathematics. Cambridge University Press, New York.

Arnoldi, J.-F., A. Bideault, M. Loreau, and B. Haegeman. 2018. How ecosystems recover from pulse perturbations: A theory of short- to long-term responses. Journal of Theoretical Biology 436:79–92.

Bever, J. D., K. M. Westover, and J. Antonovics. 1997. Incorporating the Soil Community into Plant Population Dynamics: The Utility of the Feedback Approach. The Journal of Ecology 85:561.

Borer, E. T., W. S. Harpole, P. B. Adler, E. M. Lind, J. L. Orrock, E. W. Seabloom, and M. D. Smith. 2014. Finding generality in ecology: a model for globally distributed experiments. (R. Freckleton, ed.)Methods in Ecology and Evolution 5:65–73.

Chesson, P. 1990. MacArthur's consumer-resource model. Theoretical Population Biology 37:26–38.

Clark, A. T. 2017. *Constraints and tradeoffs: Toward a predictive, mechanism-based understanding of ecological communities*. University of Minnesota, St. Paul, MN.

Clark, A. T., J. M. H. Knops, and D. Tilman. 2019. Contingent factors explain average divergence in functional composition over 88 years of old field succession. (R. Bardgett, ed.)Journal of Ecology.

Clark, A. T., and C. Neuhauser. 2018. Harnessing uncertainty to approximate mechanistic models of interspecific interactions. Theoretical Population Biology 123:35–44.

Cushing, E. J. 1963. *Late-Wisconsin pollen stratigraphy in east-central Minnesota*. University of Minnesota, Minneapolis, MN, USA.

Dormann, C. F. 2008. On community matrix theory in experimental plant ecology. Web Ecology 8:108–115.

Durrett, R., and S. Levin. 1994. The Importance of Being Discrete (and Spatial). Theoretical Population Biology 46:363–394.

Grilli, J., G. Barabás, M. J. Michalska-Smith, and S. Allesina. 2017. Higher-order interactions stabilize dynamics in competitive network models. Nature.

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

Ives, A. R., and S. R. Carpenter. 2007. Stability and Diversity of Ecosystems. Science 317:58–62.

Kitajima, K., and D. Tilman. 1996. Seed Banks and Seedling Establishment on an Experimental Productivity Gradient. Oikos 76:381.

Lehman, C., A. Keen, and R. Barnes. 2012. Trading Space for Time: Constant-Speed Algorithms for Managing Future Events in Scientific Simulations. Page 1 *in*Proceedings of the International Conference on Scientific Computing (CSC). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Leibold, M. A., and J. M. Chase. 2018. Metacommunity Ecology, Volume 59. Princeton University Press.

Levin, S. A. 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73:1943.

May, R. M. 1973. Stability and complexity in model ecosystems. Princeton landmarks in biology (1st Princeton landmarks in biology ed.). Princeton University Press, Princeton.

Muller-Landau, H. C. 2010. The tolerance-fecundity trade-off and the maintenance of diversity in seed size. Proceedings of the National Academy of Sciences 107:4242–4247.

Pimm, S. L. 1984. The complexity and stability of ecosystems. Nature 307:321–326.

R. Development Core Team,. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing.
Suding, K. N., W. Stanley Harpole, T. Fukami, A. Kulmatiski, A. S. MacDougall, C. Stein, and W. H. van der Putten. 2013. Consequences of plant-soil feedbacks in invasion. (M. Hutchings, ed.)Journal of Ecology 101:298–308.
Tilman, D. 1987. Secondary succession and the pattern of plant dominance along experimental nitrogen gradients. Ecological Monographs 57:189–214.
———. 1994. Competition and biodiversity in spatially structured habitats. Ecology 75:2–16.
Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie, and E. Siemann. 1997. The influence of functional diversity and composition on ecosystem processes. Science 277:1300–1302.
Weisser, W. W., C. Roscher, S. T. Meyer, A. Ebeling, G. Luo, E. Allan, H. Beßler, et al. 2017. Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions. Basic and Applied Ecology 23:1–73.

Supplementary Figures

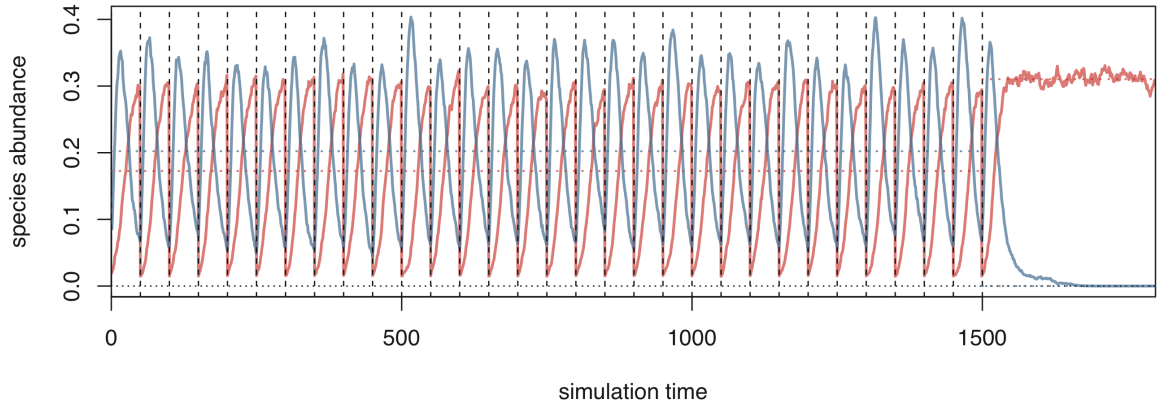


Figure S1: Example of long-term persistence of multiple species caused by periodic disturbance events in the *disturbance* model. Parameter values are identical to those used for the simulations described in the main text. Because disturbance events (vertical dashed lines) affect the superior competitor (red) more than the inferior competitor (blue), regular disturbances prevent the superior competitor from remaining at a high abundance for long enough to drive the inferior competitor extinct. This can lead to persistence by both species for arbitrarily long periods of time, depending on their relative growth rates and relative responses to disturbance. When disturbance events are discontinued after time step 1500, the superior competitor is able to rapidly drive the inferior competitor extinct. Note that species average abundances correspond to the mean expectation from Eq. (2), based on time-averaged mortality rates (horizontal dotted lines), as discussed in Appendix *I.2* in the supplement.

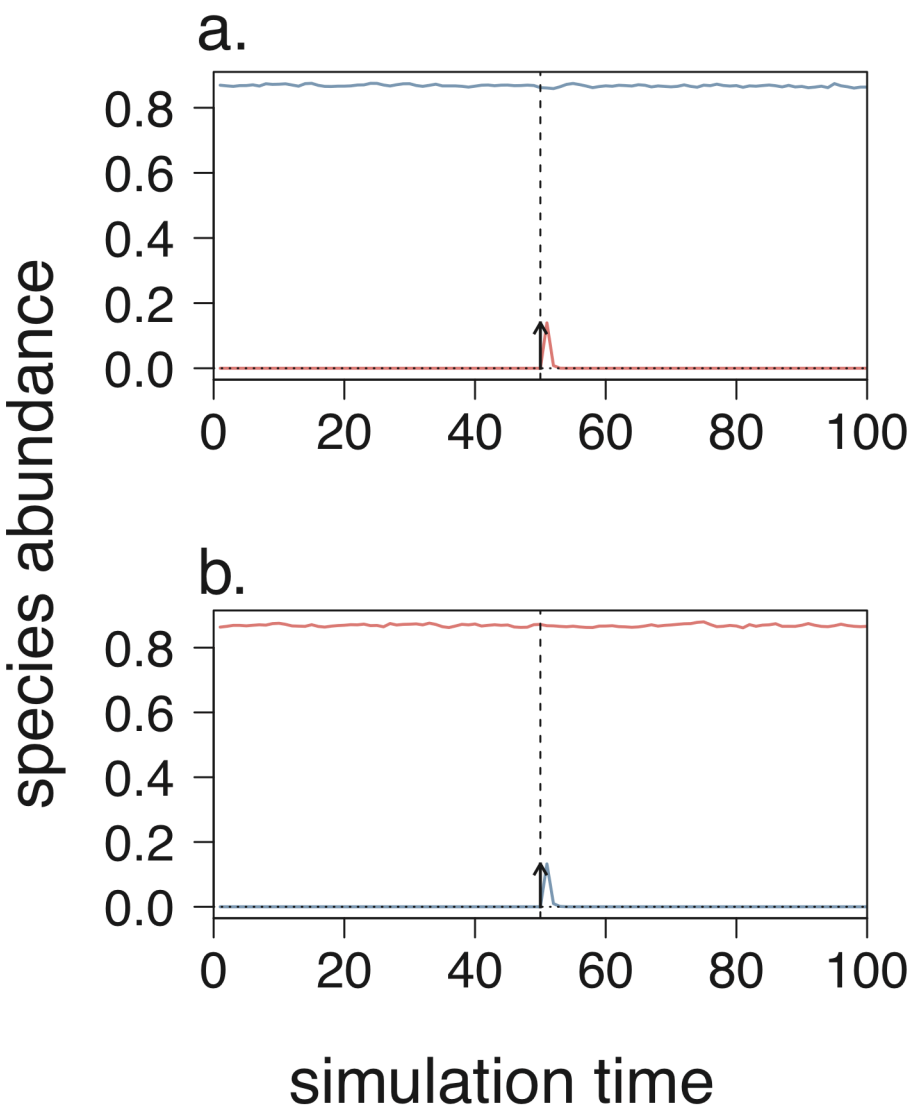


Figure S2: Example dynamics of the *RPS* model at the maximum spatial extent, for a case with positive feedbacks (in contrast to the simulations in the main text, which are for systems with negative feedbacks). Colored lines show species abundances. Vertical dashed lines and arrows show invasion events for invasion tests. In accordance with common results for such models, we find no coexistence, and strong priority effects, such that both species is able to prevent its competitor from invading.

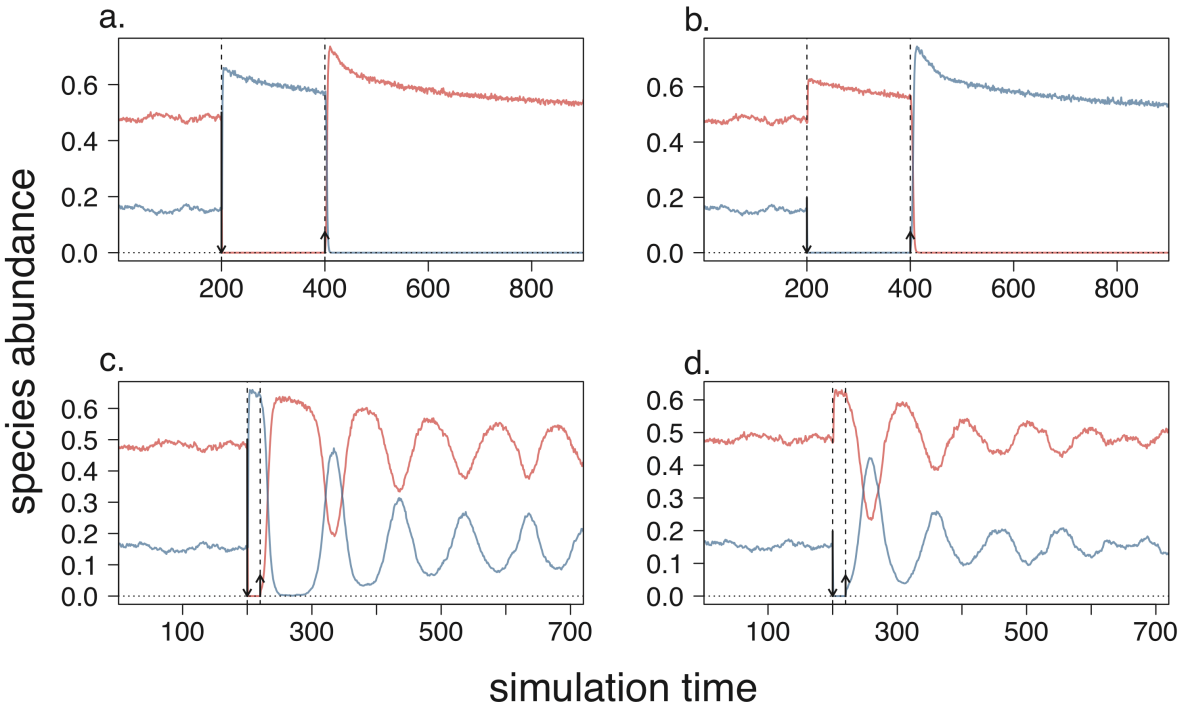


Figure S3: Example dynamics of the *RPS* model at the maximum spatial extent, for a case with negative feedbacks (i.e. the same model as shown in Fig. 2f in the main text). Colored lines show species abundances. Vertical dashed lines and arrows show invasion events for invasion tests. In cases where species are allowed to approach their monoculture equilibrium state before re-invasion of their competitor **(a,b)**, we find alternate stable states, as discussed in the main text. This is because negative feedbacks build up over time in monocultures, eventually reducing the local fitness of the resident species so profoundly that they can no longer coexist with their competitor. In contrast, in cases where only a short time interval passes between species removal and subsequent reinvasion **(c,d)**, negative feedbacks do not build up sufficiently to allow exclusion of the resident species. Thus, the two species return to their initial state of coexistence, following a period of damped oscillations.

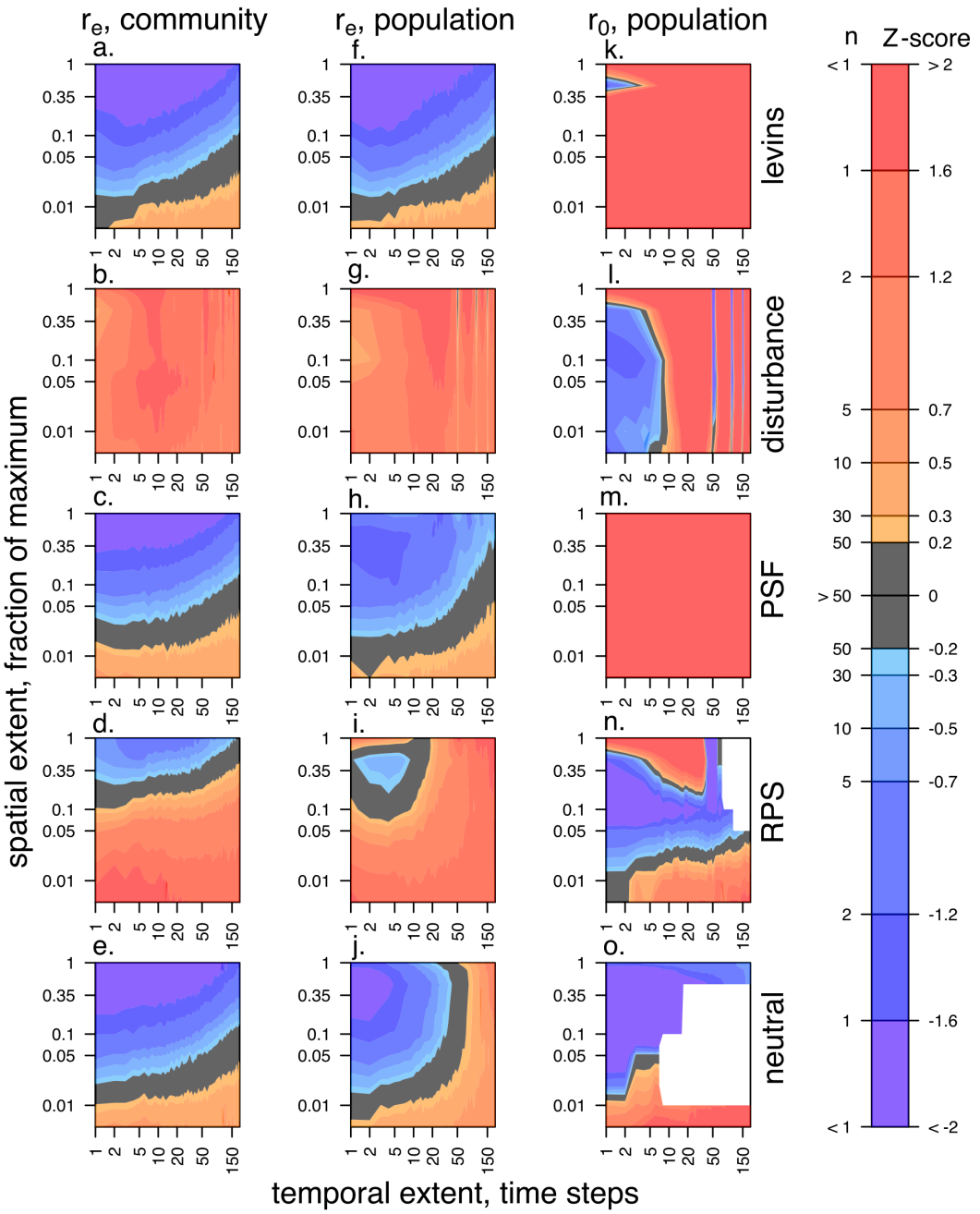


Figure S4: Distributions of perturbation and invasion metrics. Panels and axes match those described in Fig. 3 in the main text. Colors and contour lines show z-score and average sample size (n) needed to detect a significant difference from zero, based on variation among 20,000 simulations. Dark grey shows regions where z > 50 observations would be required to detect a significant difference from zero. Note that unlike in Fig. 3 in the main text, we do not invert axis colors for r_e vs r_0. In this figure, cooler colors always indicate more negative numbers and thus greater stability for r_e (panels **a-j**), whereas warmer colors always indicate more positive numbers and thus greater stability for r_0 (panels **k-o**).

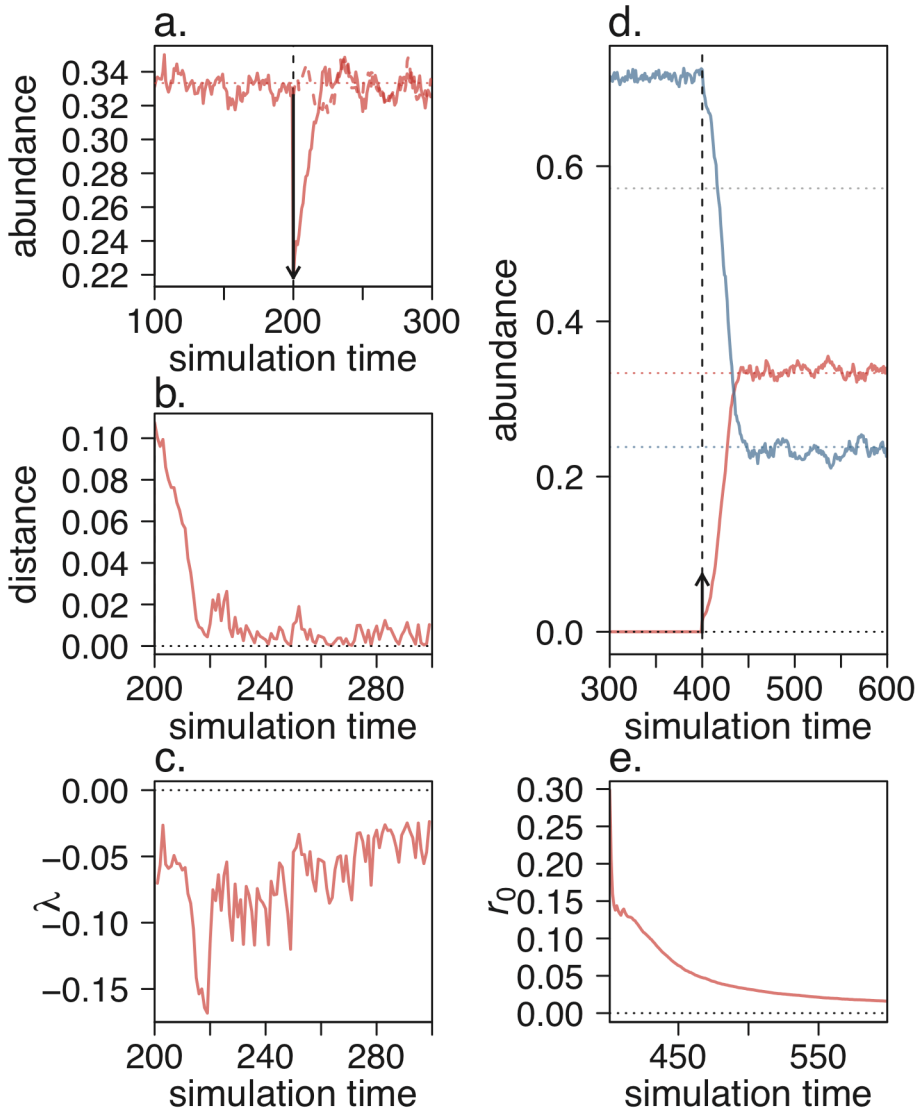


Figure S5: Example calculation of perturbation **(a-c)** and invasion **(d-e)** statistics. Time series match the *Levins* simulation from Fig. 1a-b. **(a)** Solid line shows species response to a pulse perturbation applied at time step 200. Dashed line shows trajectory in the absence of the perturbation. **(b)** Distance between the population trajectories with and without the pulse perturbation. **(c)** Time-averaged growth of distance, r_e, following Eq. (4) in the main text. Negative values indicate return to pre-perturbation state. **(d)** Invasion at time step 400 of removed (red) species into an established population (blue species). **(e)** Time-averaged growth of removed species after re-invasion, r_0, following Eq. **(5)** in the main text. Positive values indicate successful invasion.

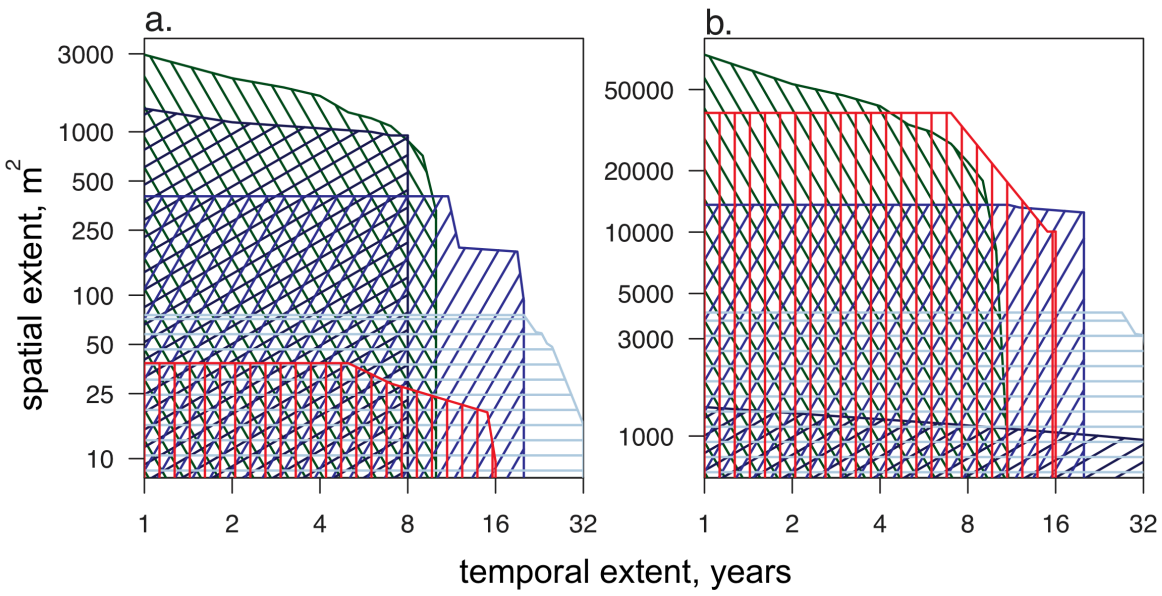


Figure S6: Examples of available spatial and temporal scales of data for surveys of grassland plants. Hatched regions show **(a)** actual sampled portions of each plots, and actual number of sampling years, or **(b)** total plot size and range of sampling years. Green shows the NutNet experiment (Borer et al. 2014), red the main experimental plots at Jena (Weisser et al. 2017), and dark blue, blue, and light blue show, respectively, the old field experiment (E014) (Clark et al. 2019), the "big" biodiversity experiment (E120) (Tilman et al. 1997), and the nitrogen addition experiments (E001/E002) at Cedar Creek (Tilman 1987).

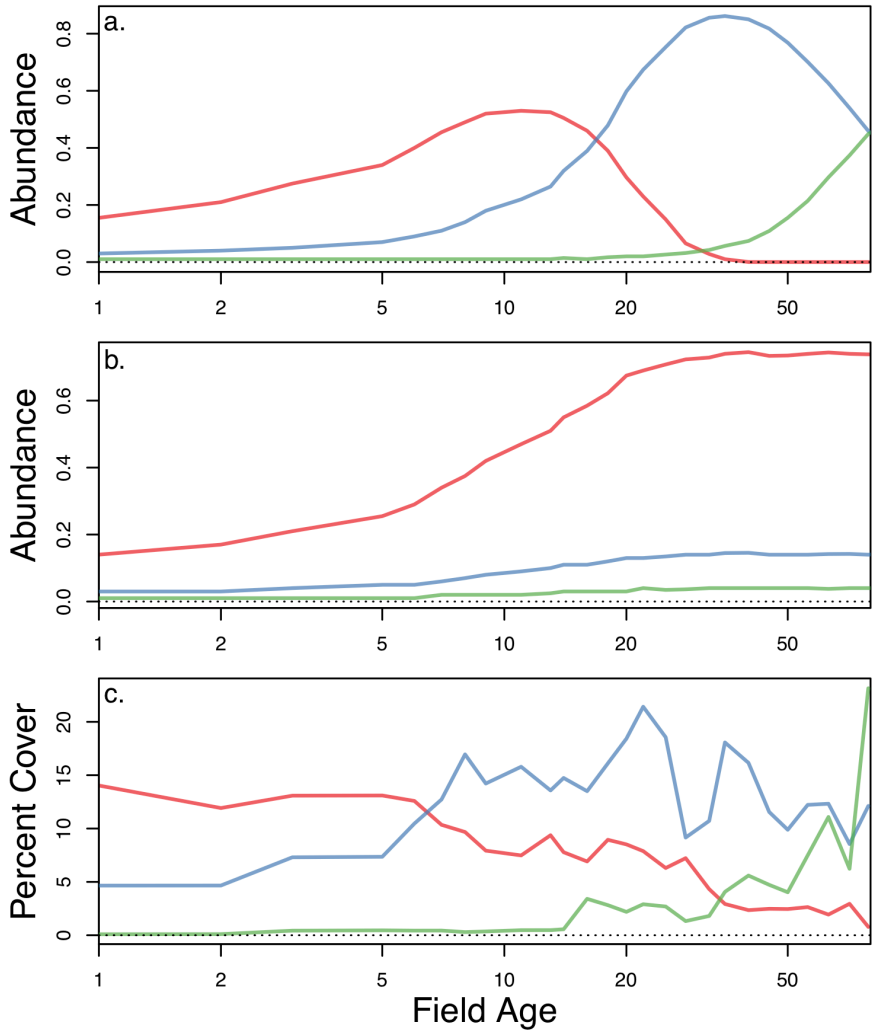


Figure S7: Temporal trends for species group dynamics. Results for simulated *Levins-OF* **(a)** and *neutral-OF* **(b)**, vs. mean observed trends in old-fields at Cedar Creek **(c)**. Model parameters are estimated from observed colonization and mortality events at Cedar Creek. Red shows annuals, blue shows C3 grasses, and green shows C4 grasses. Though simulations and empirical observations differ in terms of absolute cover, the general trends and timing of increases and decreases within each species group are quite similar between the empirical data and predictions from *Levins-OF*. Note that initial differences in species dynamics in the *neutral* case are a result of founder effects, due to differences in starting abundances.

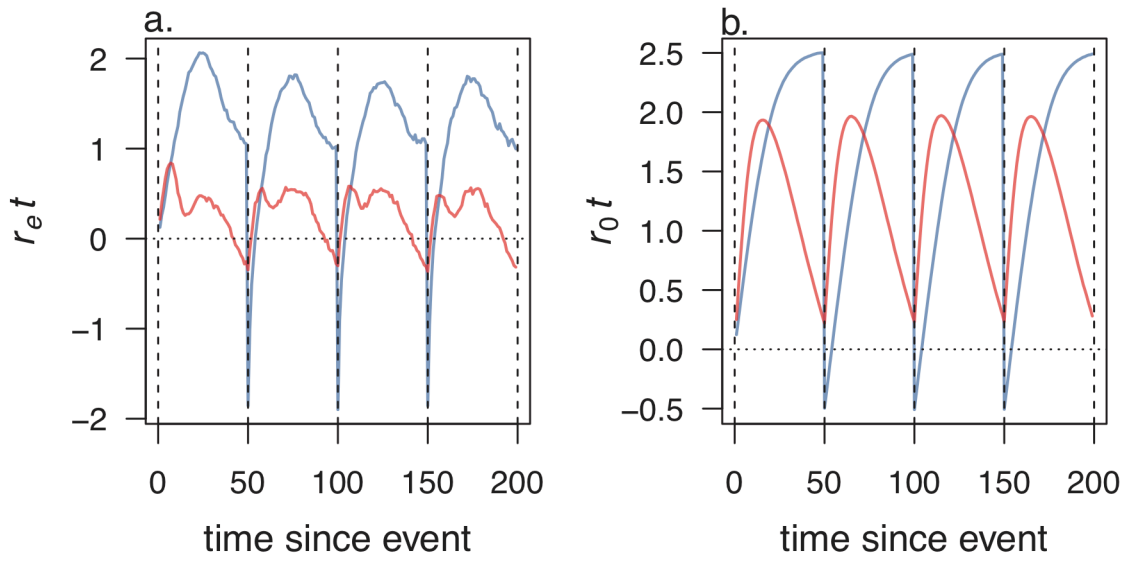


Figure S8: Population-level responses to **(a)** perturbation and **(b)** invasion tests for the *disturbance* model, observed at the maximum spatial extent. Recall that the red species is the inferior competitor and is not influenced directly by disturbance events, whereas the blue species is the superior competitor and is negatively influenced by disturbances. Solid red and blue lines show median results across simulations for each species. Vertical dashed lines show disturbance events. **(a)** For the perturbation test, the system is only indicated as stable (i.e. all $r_e < 0$) around the time of disturbances, when the inferior competitor is able to recover from competitive suppression and when the superior competitor is recovering from disturbances. **(b)** For the invasion test, both species contribute to the slowest growth rate when rare (i.e. minimum r_0). However, negative growth (i.e. $r_0 < 0$), and thus predictions of instability, occur only for the blue species, and during disturbance events, because of the sharp decline that disturbance causes in the abundance of the superior competitor. Note that for clarity, both r_e and r_0 are scaled by time (i.e. as in Fig. 4 in the main text).

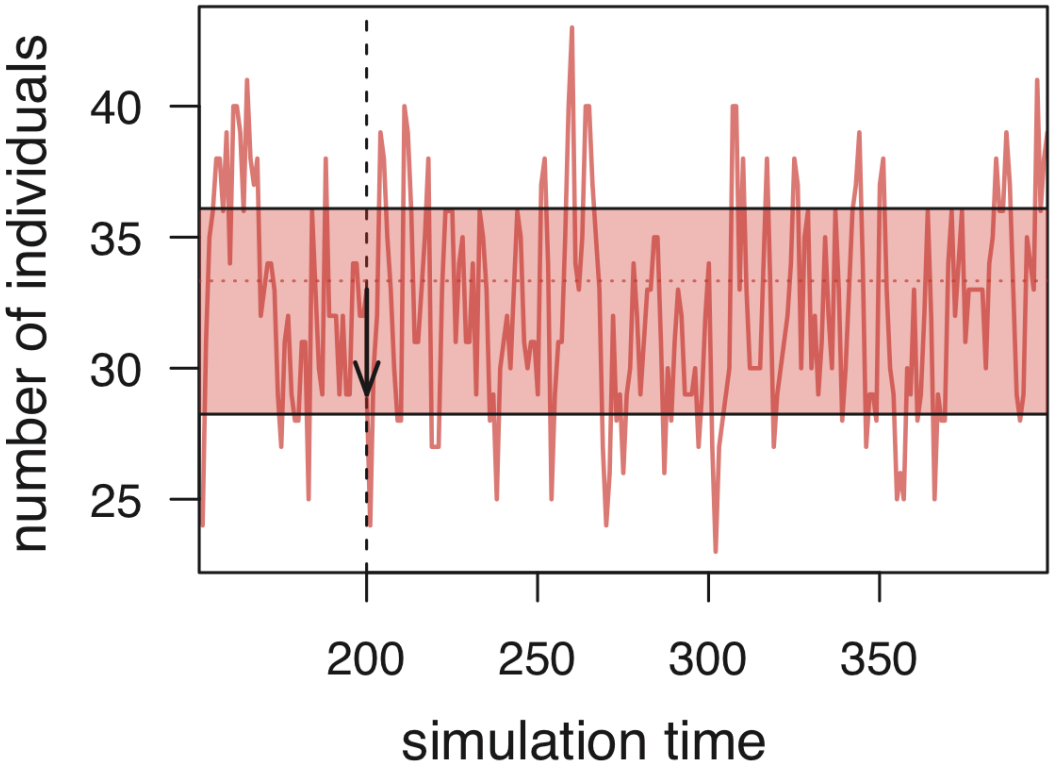


Figure S9: Example of perturbation tests at a small spatial extent. Dynamics match those from the *Levins* model in Fig. 1a, given a spatial extent of 1% of the maximum. Dashed vertical line and arrow shows a local perturbation, removing 20% of individuals (length of arrow indicates size of perturbation). Horizontal dotted line shows expected equilibrium from the analytical model. Shaded region shows mean abundance ± one standard deviation due to demographic stochasticity (i.e. stochastic birth/death). Note that demographic stochasticity leads to larger average deviations from the equilibrium than does the perturbation at this extent. Thus, the effect of the perturbation is difficult to detect.

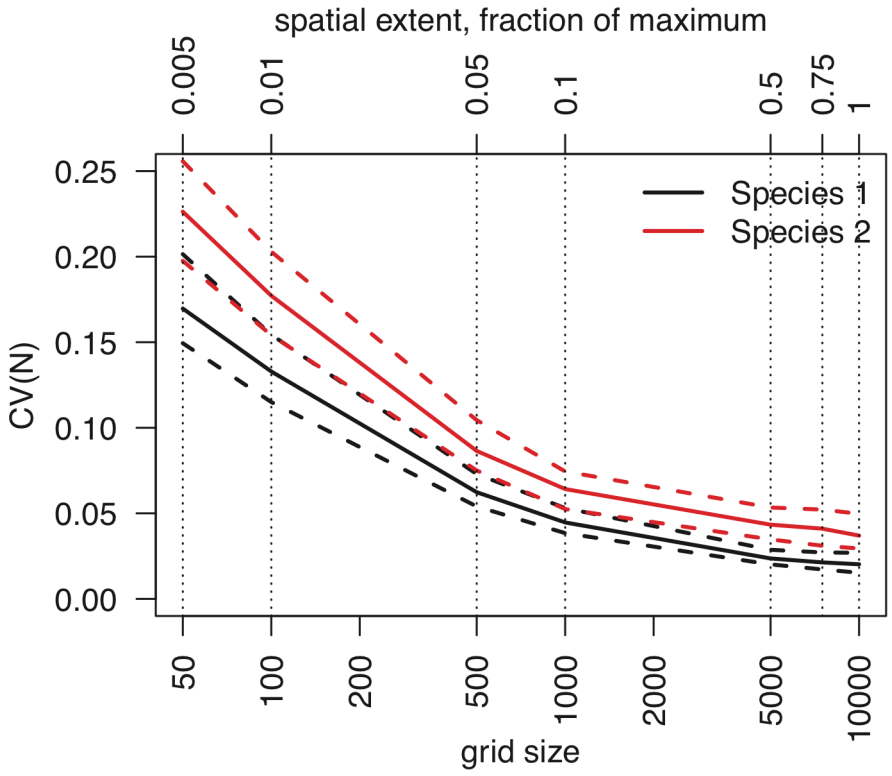


Figure S10: Effect of demographic stochasticity on dynamics of the *Levins* model as a function of spatial scale. Vertical axis shows coefficient of variation (*CV*), measured as the standard deviation of species abundance divided by the mean abundance that would be analytically expected in the absence of demographic stochasticity, following Eq. 2 in the main text. Solid and dashed lines show mean ± one standard deviation, based on 100 iterations of the model. At extents below 500 grid cells (i.e. 5% of the maximum extent), demographic stochasticity leads to relatively large variability in abundance, but declines rapidly as a function of spatial extent. At larger spatial extents, this variability is small relative to mean abundance (i.e. $CV < 0.1$), suggesting that demographic stochasticity has minor effects at these spatial scales. Note that variability is always higher for species 2 than for species 1, because it is influenced not only by its own demographic stochasticity, but also that of the superior competitor.

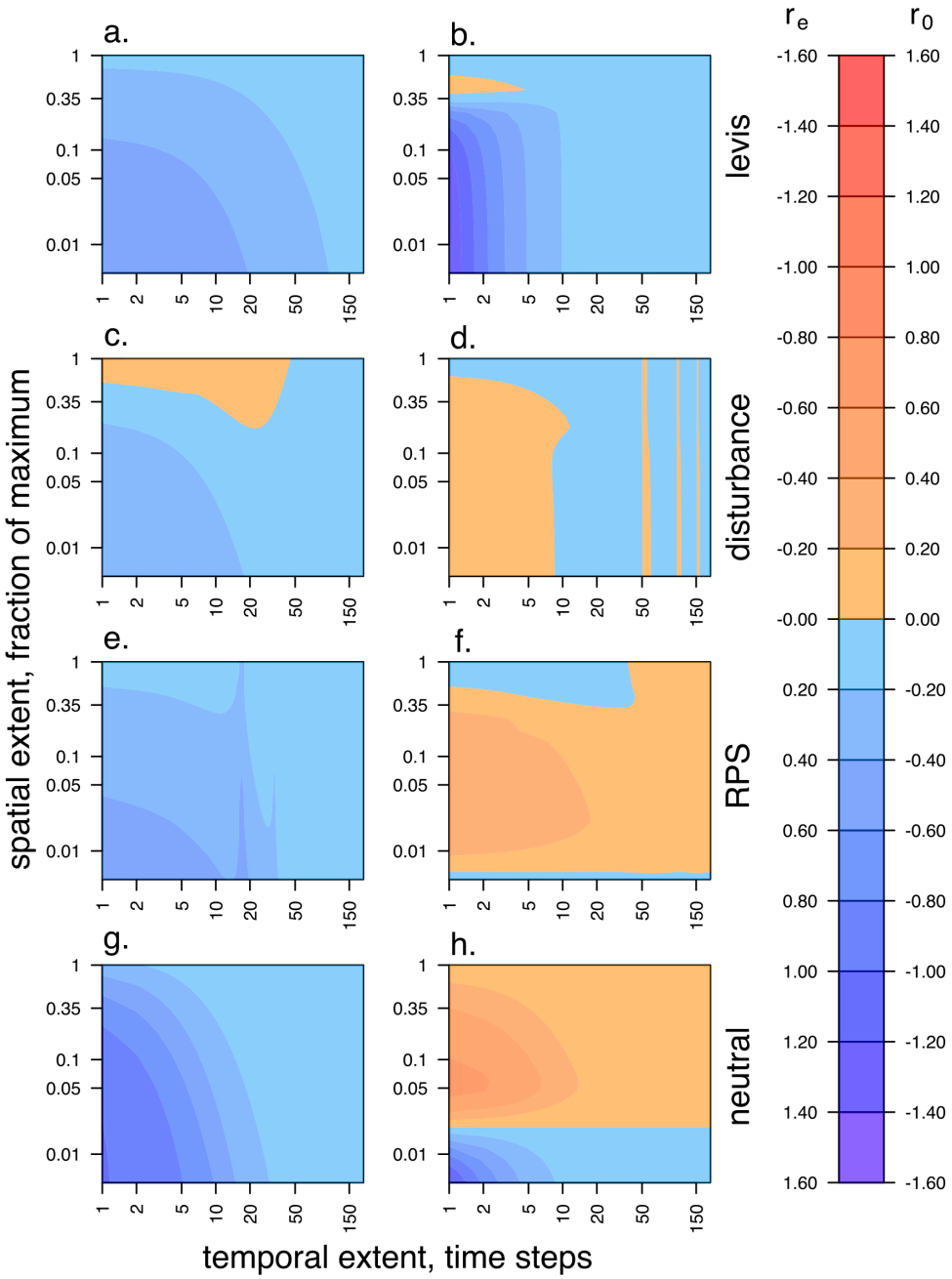


Figure S11: Stability fingerprints summarizing r_e and r_0 across spatial and temporal extents for the analytical versions the *Levins*, *disturbance*, *RPS*, and *neutral* models, following Eq. 1 and S1-S4. Colors and labels are as described in Fig. 3 in the main text. Because results are simulated directly from the underlying master equation rather than from a spatially explicit implementation based on Gillespie's method, there is no impact of demographic stochasticity on model dynamics. Note that results are almost identical to those in Fig. 3, except for spatial extents below 0.05, for perturbation tests of *neutral* at large temporal extents, and for perturbation tests of *disturbance* and *RPS* at all scales.

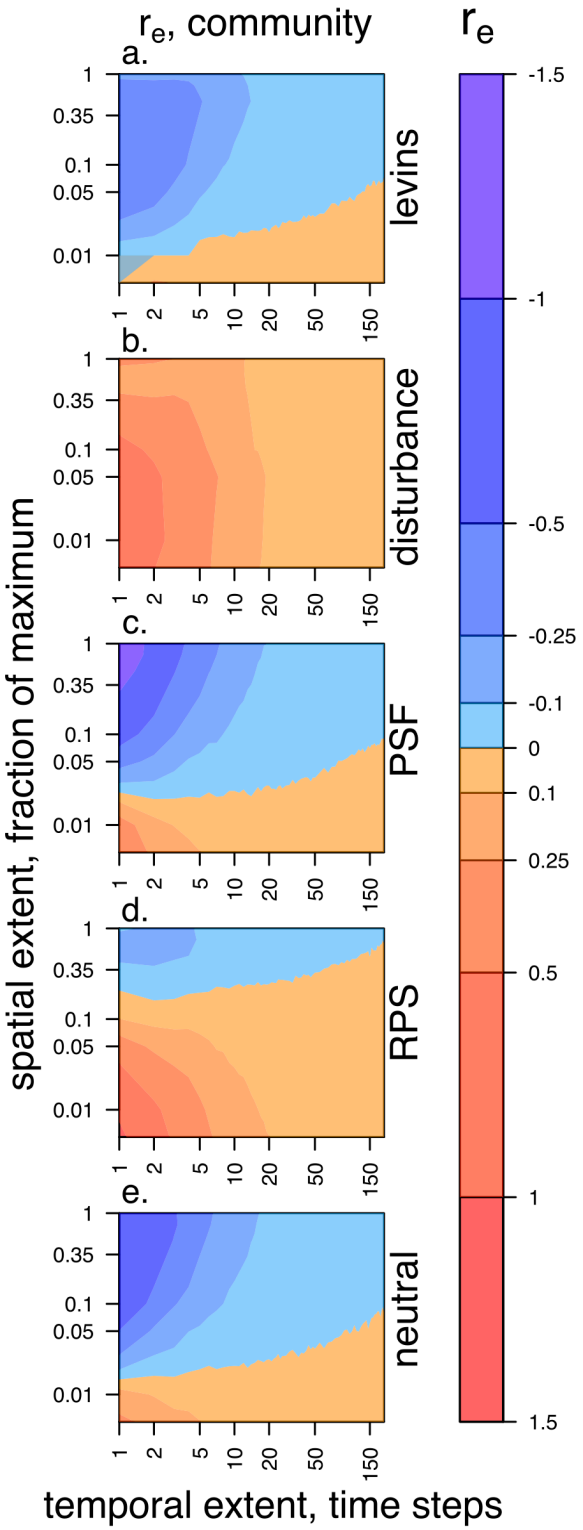


Figure S12: Simulation results for perturbation statistic across all models (columns) and spatial/temporal extents (rows). Unlike Fig. 2 in the main text, contours show median r_e from 20,000 simulations for total summed community biomass, rather than population-level biomass.

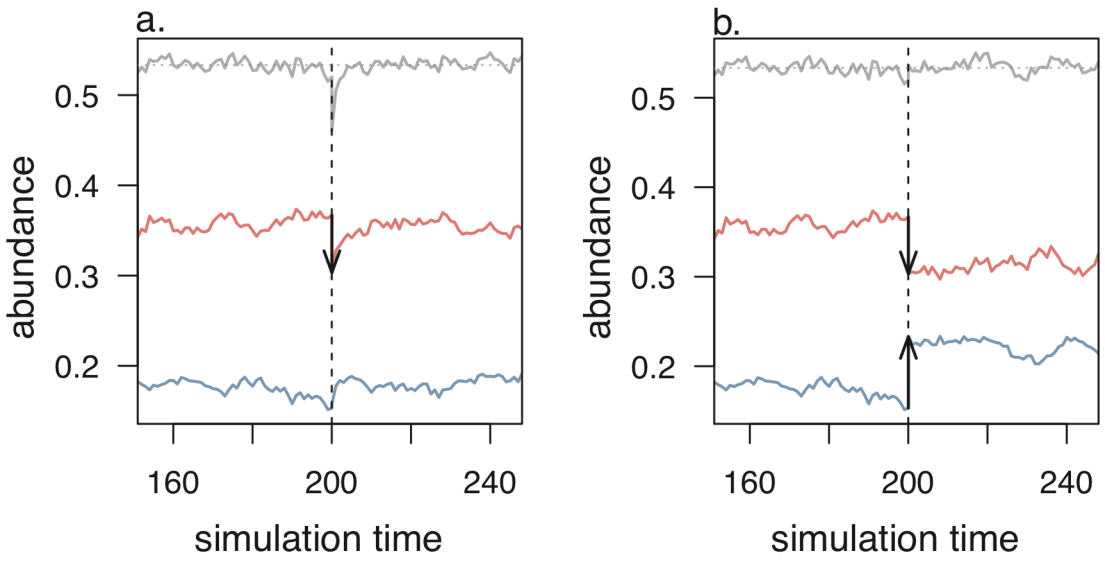


Figure S13: Example of stability at the community level leading to spurious detection of stability at the population level, from the *neutral* model. Solid black line shows summed community abundance, while red and blue lines show results for individual species. Vertical dashed lines and arrows show perturbations. **(a)** A perturbation applied to the red species with no subsequent correction in the abundance of other species leads to a decline in total community biomass. Because the neutral model is stable at the community level, this decline leads to an increase in the growth rate of all species. Thus, the red species appears to "recover" from the perturbation, even though species cannot actually coexist in the long-term. **(b)** If the population size of the blue species is increased to counteract the perturbation to the red species, there is no net change in summed community abundance. Thus, the red species no longer appears to "recover" following the perturbation.

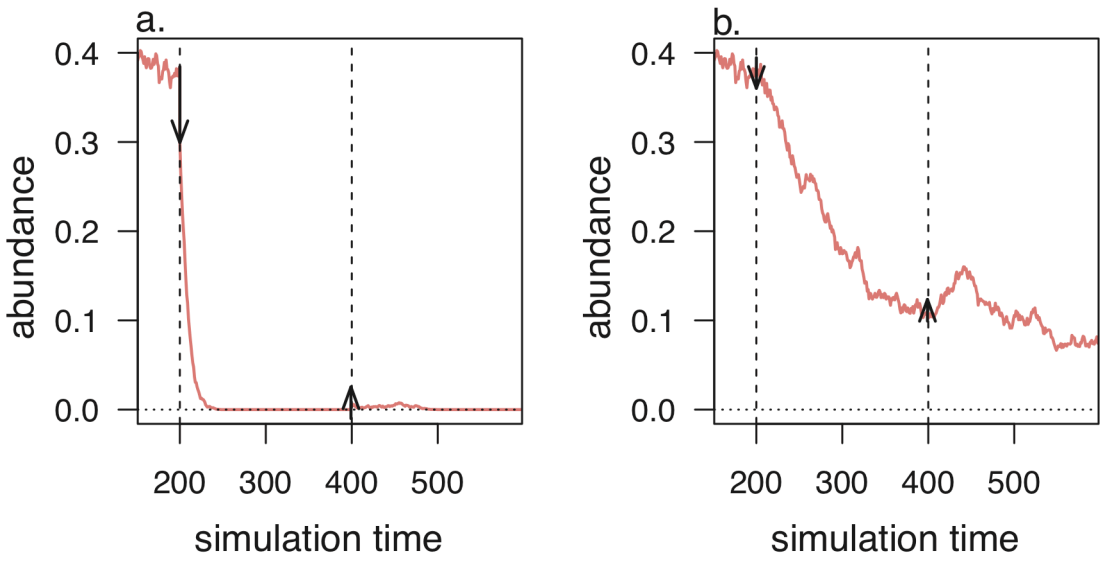


Figure S14: Example of invasion test for the *neutral* model applied at spatial extents of **(a)** 20% and **(b)** 1% of the maximum. Solid lines show abundance dynamics at the maximum extent. Vertical dashed lines and arrows show removal and re-invasion events. **(a)** At intermediate spatial extents, removal leads to deterministic extinction of the species at the maximum extent, because subsequent exclusion from the experimental region causes an effective decrease in colonization rate. Though the species is later re-introduced in the experimental plots, the resulting abundance at the maximum extent is very low. Thus, demographic stochasticity frequently leads to extinction. Note that for tests implemented at the maximum spatial extent, resulting population sizes are larger, and successful re-invasion is therefore more likely. **(b)** At small spatial extents, removal has weaker effects on abundance and dynamics at the maximum spatial extent, and the resulting deterministic declines in abundance are much slower. Thus, the species rarely goes extinct at the global extent.

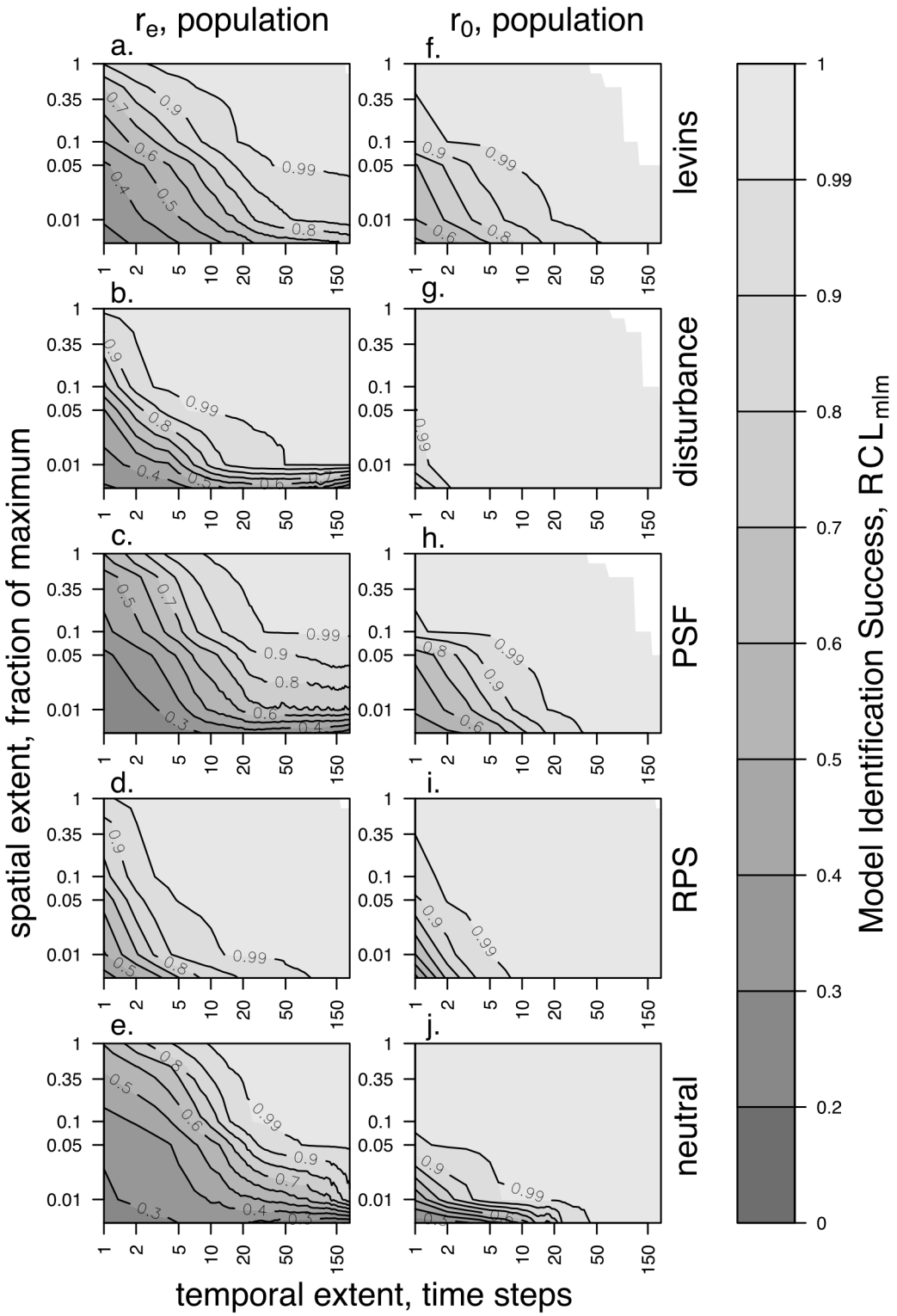


Figure S15: Relative cumulative likelihood ($RCL_{m|m}$) of successful identification of underlying models across spatial and temporal extents, based on the methods described in *Appendix IV*. Rows show results for different models. Contours show median likelihood observed across simulations, either for r_e **(a-e)** or r_0 **(f-j)**. Note $RCL_{m|m}$>0.2 implies identification is more accurate than a random guess among the five models.

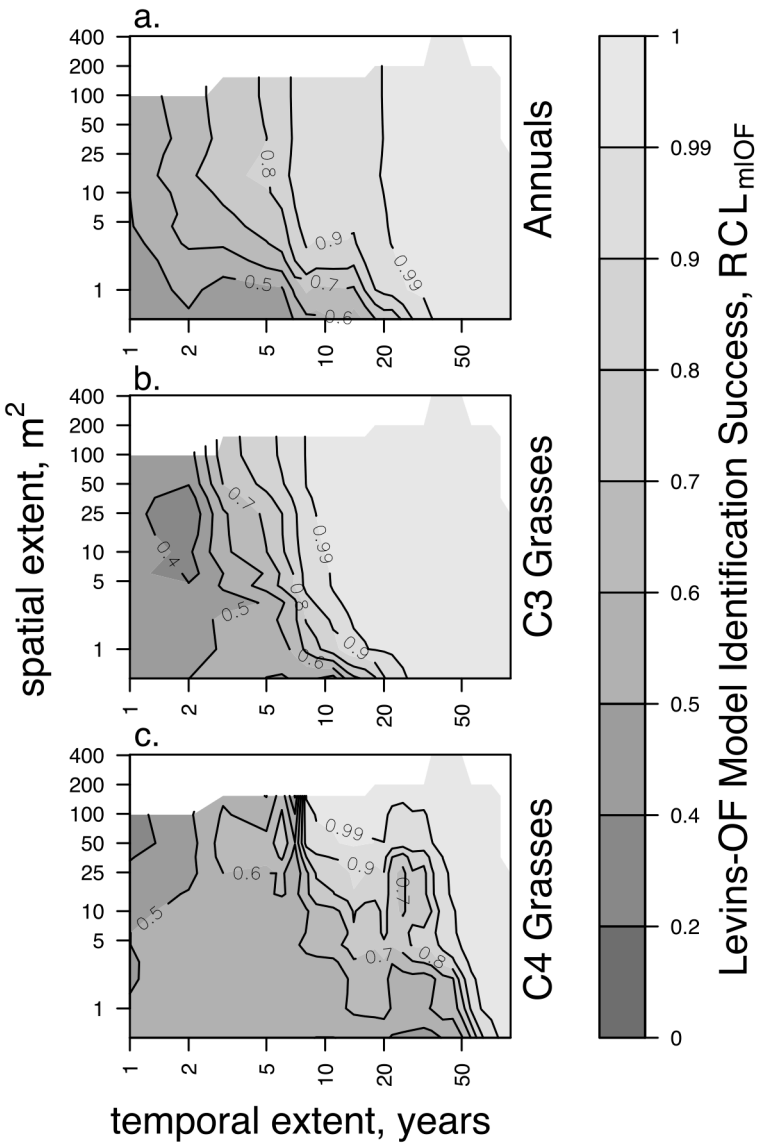


Figure S16: Cumulative likelihood of *Levins-OF* relative to *neutral-OF*, based on their abilities to explain empirically observed stability fingerprints generated from old-field dynamics at Cedar Creek ($RCL_{m|OF}$). See *Appendix IV* for details. Contours show median likelihood observed across simulations. Rows show different species groups. RCL_m>0.5 implies that empirical fingerprints are better matched by *Levins-OF* than by *neutral-OF*.