Title

Interpreting random forest analysis of ecological models to move from prediction to explanation

Journal

Scientific Reports, 13(1)

ISSN

2045-2322

Authors

Simon, Sophia M

Glaum, Paul

Valdovinos, Fernanda S

Publication Date

2023

DOI

10.1038/s41598-023-30313-8

scientific reports

Check for updates

OPEN

Interpreting random forest analysis of ecological models to move from prediction to explanation

Sophia M. Simon[1,3]✉, Paul Glaum[1,2,3]✉ & Fernanda S. Valdovinos[1,2]

As modeling tools and approaches become more advanced, ecological models are becoming more complex. Traditional sensitivity analyses can struggle to identify the nonlinearities and interactions emergent from such complexity, especially across broad swaths of parameter space. This limits understanding of the ecological mechanisms underlying model behavior. Machine learning approaches are a potential answer to this issue, given their predictive ability when applied to complex large datasets. While perceptions that machine learning is a "black box" linger, we seek to illuminate its interpretive potential in ecological modeling. To do so, we detail our process of applying random forests to complex model dynamics to produce both high predictive accuracy and elucidate the ecological mechanisms driving our predictions. Specifically, we employ an empirically rooted ontogenetically stage-structured consumer-resource simulation model. Using simulation parameters as feature inputs and simulation output as dependent variables in our random forests, we extended feature analyses into a simple graphical analysis from which we reduced model behavior to three core ecological mechanisms. These ecological mechanisms reveal the complex interactions between internal plant demography and trophic allocation driving community dynamics while preserving the predictive accuracy achieved by our random forests.

As ecologists expand the range and depth of ecological theory they necessarily integrate advanced modeling techniques and computational tools to produce increasingly complex models. This increase in model complexity has been driven, in part, by the long-standing debate on the relationship between diversity and stability in ecosystems[1]. Differing levels of diversity (in terms of number of species and interactions) directly affects the number of interacting variables and, therefore, model complexity[2,3]. Indeed, numerous studies continue to raise unique sources of complexity and separate mechanisms tying diversity to increased stability. These mechanisms include weak species interactions[4], adaptive foraging[5]; allometric scaling of interaction strength[6]; and omnivorous[7], mutualistic[8], or high-order interactions[9].

These diverse sources of ecological complexity can produce intricate patterns of model behavior governed by layered nonlinearities and interacting effects. Traditional local sensitivity analyses struggle to detect the full range of these patterns and the complex interacting relationships between model parameters and model behavior. This is because local sensitivity analyses necessarily addresses sensitivity relative to changes in a restricted parameter space. Such limitations consequently hinder ecological modelers' ability to identify the ecological mechanisms behind model output. To ensure that advances in model complexity accompany advances in utility for the field of ecology, researchers must (1) utilize tools to detect the complex relationships between model input and output, (2) develop processes to identify the unifying ecological mechanisms driving those relationships, and (3) do so across broad distributions of model parameter sets.

Interpretable statistical models, such as generalized additive models (GAMs), can offer utility in this regard[10]. However, GAMs do not necessarily "find" interactions. They instead test the validity of explicitly modeled hypothetical interactions. This frequently requires exhaustive model selection to identify and verify critical interactions driving model behavior. This can be particularly challenging with a large numbers of smoothing parameters, with highly interactive input parameters, when statistical power is limited, or when smoothing in higher dimensions (however, see Discussion for more on the utility of GAMs).

Other options for producing analyses of complex model data lie in machine learning. Modern machine learning algorithms are flexible, powerful tools used to study many complex systems and are increasingly applied

[1]Department of Environmental Science and Policy, University of California-Davis, Davis, CA 95616, USA. [2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA. [3]These authors contributed equally: Sophia M. Simon and Paul Glaum. ✉email: sosimon@ucdavis.edu; prglaum@umich.edu

to ecological datasets[11]. Recently, ecologists have applied machine learning algorithms to predicting species interactions from empirical trait data[12], improving estimation of viral host ranges from incomplete datasets[13], and examining food web responses to variable functional diversity across trophic levels[14]. Random forests are a particularly intuitive option with limited parameter tuning, readily applicable to various types of data with built in limits on overfitting, and capable of regression and classification tasks. Finally, unlike GAMs, random forests can also identify higher order features or interactions without requiring specific hypotheses a priori, a useful characteristic for data exploration before intuition for model behavior is developed.

While random forests and other machine learning approaches are powerful, they have been called "black boxes" because of their complex algorithmic nature and a perceived lack of interpretability behind their high predictive power[15]. Recently, methods for the interpretation of machine learning results have been reviewed in the hopes of transforming this "black box" into a "translucent box"[11]. However, developing comprehensively interpretable results via random forest remains difficult. This is partially because machine learning algorithms identify key connections between a variety of inputs and outcome, but do not necessarily develop unifying principles behind the systems they study. When applied to ecological models this can potentially limit ecological generalizability as model complexities lead to incommensurable (i.e., lacking common measurement standards) results across different formulations and parameterizations[14].

Here, we describe a process to expand random forest interpretability when applied to ecological models and demonstrate their utility in producing comprehensive mechanistic descriptions of model behavior. To do so, we use a key source of ecological complexity as an example, demographic heterogeneity. Demographic heterogeneity is frequently modeled as stage-structured organismal ontogeny[16]. Explicitly modeled ontogenetic stages increase the number of unique ecological actors within a single species. This consequently increases the total number of species interactions in an ecosystem because each ontogenetic stage tends to have its unique interactions with other species or stages[17]. Stage-structured models have made clear the importance of demographic heterogeneity in influencing population dynamics, particularly in plants[18,19]. At the community level, the few food web studies addressing the role of organismal ontogeny on community stability have found mixed results. Rudolf and Lafferty[20] found that diet shifts across organismal development (ontogenetic niche shifts) destabilize food webs. However, in formulating a community model without ontogenetic niche shifts, de Roos[21] found that explicit ontogeny stabilizes food webs so long as ontogenetic development involves substantial asymmetries between juveniles and adults. These qualitatively different results indicate that the specific demographic and ecological formulation of ontogeny influences its effect on broader dynamics. Again, here we see the potential for incommensurable results.

Studying how dynamics change with specific ecological/ontogenetic structures is best addressed initially through simpler, more tractable models that facilitate a deeper sensitivity analysis across model formulations, providing a useful basis and reference as model complexity is scaled up to represent entire communities. Modeling plant–herbivore interactions is well suited for this purpose given that plants frequently have clear ontogenetic stages or size distinctions that affect their intra- and interspecific interactions, and plants' autotrophic nature allows us to build small-scale tractable models that can form the basis of larger food webs. Furthermore, as a consequence of the long history of plant ontogeny in ecology, there are empirical resources to aid in vetting basic model formulation[22].

Therefore, given the suitability of plant ontogeny to the study ecological complexity, we focus on a plant–herbivore model with empirically informed ontogenetic stage structure and parameterization. Using this simulation model's parameters as random forest input features (see "Methods"), we then implement a machine learning based analysis of model behavior with the goal of exhibiting its utility in identifying and categorizing context dependent dynamics. Then, given the high degree of context-dependent results observed from our simulation model (and others[14]), we also demonstrate how random forests can facilitate the development of a comprehensive model analysis unifying the explanation of seemingly incommensurable results under ecological mechanisms that maintain high predictive accuracy even when applied in comparatively simple linear models.

Methods

Model development and justification. We implemented the plant ontogeny framework from Glaum & Vandermeer ([19]; see Eq. (1), Fig. 1, and Appendices S1.1 and S1.2). Plant ontogeny is divided into three stages: a seed bank (S_1), non-reproductive seedlings (S_2), and fecund adults (F). While there are clearly a variety of potential ontogenetic structures, a three-stage ontogenetic structure provides initial tractability in analysis and is well-represented across plant species (see Appendix S1.1). Public data repositories[22] hold nearly 1200 instances of plant species demographics represented by a three-stage structure representing 107 unique species. These three-stage plant species span a wide phylogenetic distribution, representing both eudicots and monocots across 3 phyla, 3 classes, 28 orders, 45 families, and 87 genera. Botanically, these species represent shrubs, succulents, trees, epiphytes, annuals, and herbaceous plants. Finally, three-stage plant species are also widespread geographically, representing eleven terrestrial ecoregions across all continents except Antarctica, making three-stage structures worthy of concentrated theoretical evaluation. For further details, please see SI File 1 and Appendix S1.

The stage-structured plant in our model is consumed by a single herbivore species (H) (Eq. 1; Fig. 1A). Real world herbivory can certainly involve multiple species interactions. However, specialization of herbivore species on a single plant taxa (especially at the family level) is common (especially amongst insect herbivores) and geographically widespread[23,24]. The modeled herbivore is limited to eating the vegetative stages (S_2 and F) as the role of "seed predator" is rarely filled by a species which also consumes vegetative tissue, again especially in insects due to the requirements of different mouth parts. Among herbivore species which consume vegetation (specialist or otherwise), past work has cataloged examples of herbivores exhibiting a range from clearly distinct

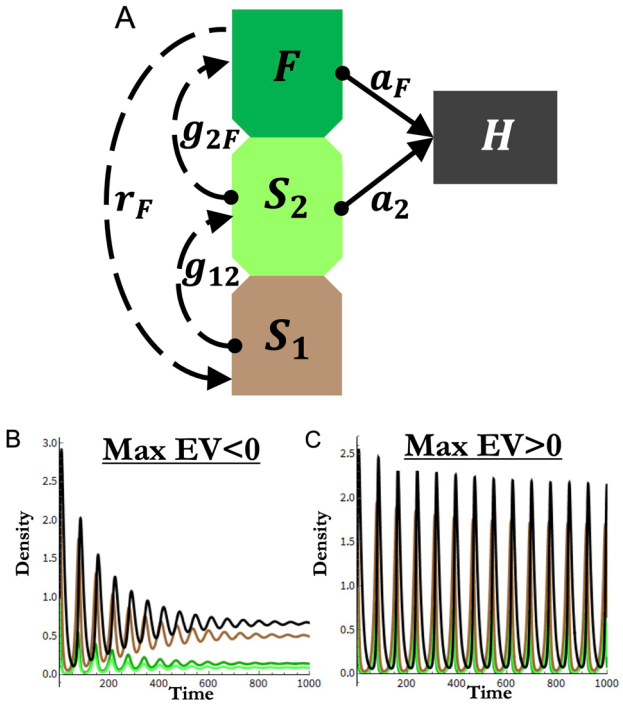

Figure 1. Simulation model overview. (**A**) Model diagram and major parameters mediating model flow. Dashed lines indicate plant population dynamics, solid lines indicate trophic interactions. S_1 indicates seed density, S_2 indicates seedling density, F indicates fecund adult density, H indicates density of herbivorous insects. Arrows indicate direction of density gain and circles indicate direction of density loss. Parameters detailed in "Methods" and Table 1. (**B**) Example of stable population trajectories indicated by negative max eigenvalue (Max EV < 0). (**C**) Example of persistent population oscillations indicated by positive max eigenvalue (Max EV > 0). Line colors for (**B**) and (**C**) correspond to (**A**).

to non-existent ontogenetic preferences in their plant resources[25–29]. Still other examples have found variable preferences, potentially resulting from local plant chemistry, leaf palatability, and microclimate[28,30]. Reflecting this empirical range, we vary the focus of herbivory to either specialize on a particular ontogenetic stage (adult or seedling) or both stages to varying degrees. The model is shown in Eq. (1) with each component (germination, consumption, etc.) labeled as different sub-functions. Sub-functions and model parameters are detailed in Table 1.

$$
\begin{aligned}
\frac{dF}{dt} &= \overbrace{\frac{g_{2F}S_2}{1+\alpha_{g2}(F+\varepsilon S_2)}}^{\substack{\gamma_{2F}\\ maturation}} - \overbrace{\frac{a_F F H}{1+a_F h_F F + a_2 h_2 S_2}}^{\substack{\theta_F\\ consumption}} - \overbrace{d_F F}^{\substack{background\\ mortality}} \\
\frac{dS_2}{dt} &= \overbrace{\frac{g_{12}S_1}{1+\alpha_{g1}(F+\varepsilon(S_1+S_2))}}^{\substack{\gamma_{12}\\ maturation}} - \overbrace{\frac{g_{2F}S_2}{1+\alpha_{g2}(F+\varepsilon S_2)}}^{\substack{\gamma_{2F}\\ maturation}} - \overbrace{\frac{a_2 S_2 H}{1+a_F h_F F + a_2 h_2 S_2}}^{\substack{\theta_2\\ consumption}} - \overbrace{d_S S_2}^{\substack{background\\ mortality}} \\
\frac{dS_1}{dt} &= \overbrace{\mathrm{Max}(F(r_F-\alpha_F F),0)}^{\substack{\delta\\ seedproduction}} - \overbrace{\frac{g_{12}S_1}{1+\alpha_{g1}(F+\varepsilon(S_1+S_2))}}^{\substack{\gamma_{12}\\ maturation}} - \overbrace{S_1(d_S+\alpha_{FS}F)}^{\substack{background\\ mortality}} \\
\frac{dH}{dt} &= c_{FH}\overbrace{\frac{a_F F H}{1+a_F h_F F + a_2 h_2 S_2}}^{\substack{\theta_F\\ consumption}} + c_{2H}\overbrace{\frac{a_2 S_2 H}{1+a_F h_F F + a_2 h_2 S_2}}^{\substack{\theta_2\\ consumption}} - \overbrace{d_H H}^{\substack{background\\ mortality}}
\end{aligned}
\tag{1}
$$

Simulation design. Simulations and numerical analyses were done in Mathematica 11. We focus analysis on three specific demographic parameters and two trophic parameters. The demographic parameters are germination rates of seeds into seedlings (g_{12}), seedling maturation rates into fecund adults (g_{2F}), and seed production

Parameter	Definition
r_F	Intrinsic reproduction (seed) rate of fecund plant individuals (F). $0.2 \leq r_F \leq 4.0$
g_{12}	Germination rate of seeds (S_1) into seedlings (S_2).$0.12 \leq g_{12} \leq 0.88$
g_{2F}	Maturation rate of seedlings (S_2) into fecund adults (F).$0.12 \leq g_{2F} \leq 0.88$
a_F	Attack rate of the herbivore (H) on the fecund adult plant population (F).$0.0 \leq a_F \leq 2.0$
a_2	Attack rate of the herbivore (H) on the seedling population (S_2).$0.0 \leq a_2 \leq 2.0$
c_{FH}, c_{2H}	Conversion rate of eaten adult plants (c_{FH}) or seedlings (c_{2H}) into herbivores.$c_{FH} = c_{2H} = 0.6$
$\alpha_F, \alpha_{g1}, \alpha_{g2}, \alpha_{FS}$	Strength of density dependence affecting seed production, seed germination, seedling maturation, and seed survival respectively. $\alpha_{FS} = 0.001$; $\alpha_{g1}, \alpha_{g2}, \alpha_F \in \{0.06, 1.0\}$
ϵ	Parameter mitigating density dependent attenuation of seeds on germination.$\epsilon = 0.2$
d_F, d_S, d_H	Death rates for the flowering plant, seeds/seedlings, & herbivore. $d_F = 0.1, d_S = 0.1, d_H = 0.2$.
h_F, h_2	Handling times for herbivory on reproductive adults (h_F) and seedling predation (h_2) $h_F \in \{0.5, 1.0\}$ & $h_2 \in \{0.5, 1.0\}$
Sub-function	**Definition**
δ	Density of seeds produced
γ_{12}	Density of seeds (S_1) maturing to seedling stage (S_2)
γ_{2F}	Density of seedlings (S_2) maturing to fecund stage (F)
θ_F	Density of adults (F) lost to consumption
θ_2	Density of seedlings (S_2) lost to consumption
Factors (linear predictors)	**Definition**
γ_{12}^*	Density of seeds (S_1) maturing to seedlings (S_2) at model equilibrium
γ_{2F}^*	Density of seedlings (S_2) maturing to fecund stage (F) at model equilibrium
$L : D$ratio	$(\theta_F^* + \theta_2^*)/(S_1^* + S_2^* + F^*)$

Table 1. Model parameters, sub-functions, and ecological factors. Parameter descriptions include values used in numerical analysis. Text formatted parameters indicate varied parameters during parameter sweeps. Bold highlighted parameters indicate parameters whose value range was varied in factorial parameter sweeps (see Appendix S1.3) for analysis via random forest. Underline highlighted parameters were varied as part of sensitivity analysis. Note, for each simulation $\alpha_{g1} = \alpha_{g2} = \alpha_F$. All parameter rates are per capita. Note, the 5 bold highlighted parameters were used as input features in our random forests. Also note, F^*, S_1^*, and S_2^* indicate equilibrium values of our time dependent variables. Similarly, θ_F^* and θ_2^* indicate density of adults (F) and seedlings (S_2) lost to consumption respectively at model equilibrium.

rate by fecund adults (r_F). The trophic parameters are the attack rates of herbivores (H) on seedlings (a_2) and fecund adults (a_F). These five rates represent both demographic and trophic flow, a functional basis for studying how plant ontogeny interacts with trophic dynamics (Fig. 1A).

We informed the ranges of demographic parameter values simulated in our parameter sweeps using empirical data of three-stage plant population dynamics[22]. Reproduction from the third stage (F) into the first stage (S_1) displayed a large range of values but was highly left skewed, with ~ 87% of values ≤ 10. Our exploratory analysis was done with $r_F \leq 10$ and detailed analysis held at $r_F \leq 4$ (Appendix S1.2; Fig. S1a). Stage transitions, on the other hand, are much more evenly distributed between 0 and 1 (Fig. S1b,c), so analysis considered $0.12 < g_{12}$, $g_{2F} \leq 0.88$. Despite the skewness in the data, analysis reveals no correlation between any demographic rates, indicating no appreciable covariation between parameters (Appendix S1.3; Fig. S2). In demographic terms, no covariation between these rates means there was no clear evidence of tradeoffs such that, for example, a higher seed production rate necessitated lower germination rates, or direct correlative relationships between the stage transition rates. The range of herbivore attack rates was chosen heuristically (see Table 1). The lack of any apparent restrictions or relationships with parameter values spurred a broad factorial parameter sweep across all five parameters via cluster computing. We duplicated the five parameter factorial sweep across a range of handling times and degrees of density dependence to test for ubiquity in qualitative results, producing nearly 5.5 million simulations.

Simulation analysis. *Random forests.* Simulation output measured various factors (Appendix S1.3), but focused on local stability indicators (maximum eigenvalues) signifying stable trajectories (Fig. 1B) or persistent oscillations (Fig. 1C). Stability indicators are a convenient and fundamental description of the ecological dynamics resultant from each parameter combination. There are multiple machine learning approaches that can provide useful inference to high dimensional nonlinear analysis. We used random forests because they are: (1) relatively easy to tune given the small number of hyper-parameters, (2) frequently used for feature/predictor selection, (3) flexibly applied across numerous biological fields, (4) and can readily be applied to both categorical and quantitative data[31,32]. Using the randomForest package in R[33], our five simulation model parameters (Fig. 1A) functioned as features/predictors with local-stability indicators serving as our predicted variables (Appendix S2). To avoid terminology confusion, we refer to simulation model parameters in general as "parameters" and refer to them specifically as "features" when used as random forest inputs (i.e., independent variables). We then use the term "predictor" only to specifically refer to independent predictors used in our liner models.

For classification tasks we used a simple indicator, locally stable or unstable. For regression tasks we used the model equilibria's eigenvalues. We trained random forests using hold out cross validation methods. As a default, random forest parameter "mtry" was set at floor(sqrt(p)) for categorization tasks (stable vs unstable) and floor(p/3) for regression tasks (max eigenvalue) where p = # of features (see Appendix S2.1). Instances where a different p produced better results are noted below. The random forest parameter "ntrees" (No. of trees) was varied from 300 to 600 with little to no effect on performance. Note these random forest parameters specifically refer to random forest formulation and are not related to the simulation model parameters. We measured random forest performance on validation/test data using area under the receiver operating characteristic curve (AUC; pROC package) for classification tasks and RMSE for regressions. We measured Variable Importance of individual features in forming predictions with Mean Accuracy Decrease and Mean Increase in MSE for classification and regression respectively (see Appendix S2; Fig. 1). An example of our use of random forests in our analysis is available in SI File 2.

Our five simulation model parameters (3 demographic and 2 trophic rates; see Table 1) served as random forest input features. We determined the degree to which they were independent or interdependent on one another in their effect on random forest predictions with Friedman's H-statistic[34]. The H-statistic can measure either (1) the overall interactivity of a single input feature with all other input features (Fig. 2A) or (2) the interactivity of specific pairs of input features (Appendix S3; Fig. S5). Higher H-statistic values indicate higher interactivity. We examined the specific details of these interacting features' effects on trophic dynamics with partial dependence (PD) plots and Individual Conditional Expectation (ICE) curves using the iml package in R[35]. These results served as the basis for our graphical analysis.

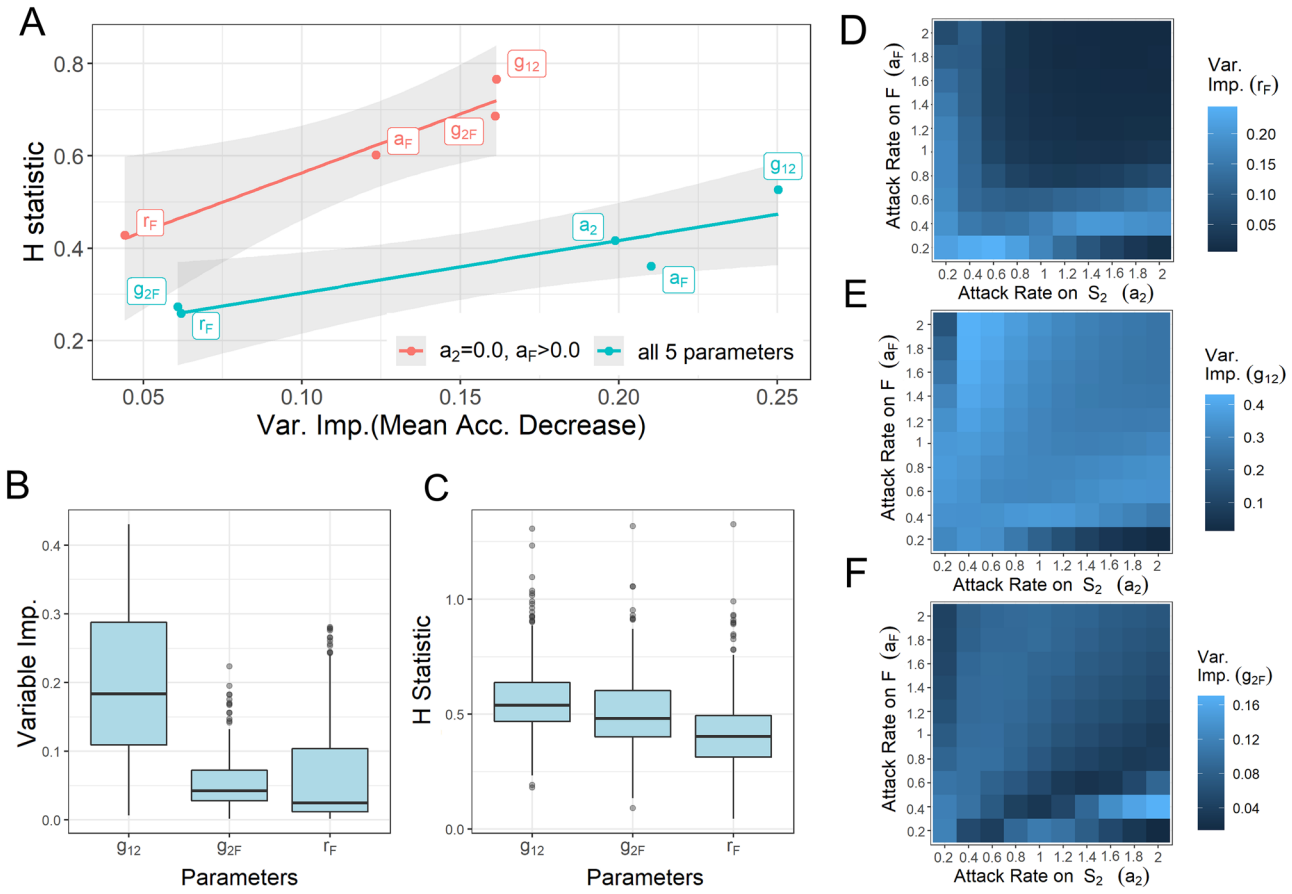

Figure 2. Random forest output overview. (**A**) Relationship between Variable Importance and interactivity (H-statistic) of parameters in Random Forest output. Blue line: RF model across all five parameters, AUC:0.998 ($\alpha_{g1} = \alpha_{g2} = \alpha_F = 0.1$; $h_2 = h_F = 0.5$). Red line: RF model with single stage herbivory ($a_2 = 0$); AUC = 0.984 ($\alpha_{g1} = \alpha_{g2} = \alpha_F = 0.1$; h_2=h_F=1). Shaded regions represent standard error. (**B**,**C**) Box and whisker plots detailing range of Variable Importance (**B**) and H-statistic (**C**) for each demographic rate in random forests run with set attack rates where a_2 and a_F vary between 0.2 and 2.0 ($\alpha_{g1} = \alpha_{g2} = \alpha_F = 0.1$; $h_2 = h_F = 0.5$). Boxes represent the interquartile range with the horizontal line showing the median, the lower box showing the 25 percentile, and the upper box showing the 75 percentile. Upper and lower lines extending from the boxes show the most extreme values within 1.5 times the 75th and 25th percentile respectively. Outliers are shown as single dots. (**D**–**F**) Heatmaps showing changing importance for each demographic rate across different allocations of consumption across plant stages. Here we found consistently but only slightly better performance with mtry = 2.

Linear regression. Random forest analysis aided in the identification of several model components, quantifiable pre-simulation, which drove simulation behavior. We refer to these components as components as "factors" (see Table 1 and "Results" for derivation). We investigated the factors' ability to both explain and predict model behavior through linear regression models. To avoid terminology confusion, we refer to these factors as predictor variables in our linear regression. For classification predictions (stable/unstable) with our ecological factors as predictor variables, we used generalized linear models and checked for accuracy using hold out cross validation and the AUC metric. For predictions of the continuous max. eigenvalue (also used to indicate stability) with our ecological factors as predictor variables, we used partial least squares regression (pls package). Partial least squares were used due to collinearity between our ecological factors. Partial least square regressions were evaluated with adjusted R^2 and checked for predictive accuracy using hold out cross validation and RMSE. Partial least square regression coefficients were then used to show how ecological drivers of model behavior change across trophic allocation (Fig. 4).

Results

Random forest: feature importance and interactivity. Our random forests produced highly accurate predictions of local stability when trained on model output from the full dataset (e.g., AUC = 0.998 across all 5 parameters, see Fig. 2A) and all tested subsets. Running random forests on the full results set with all five parameters as predictors indicated both demographic and trophic rates were important to understanding resultant model stability. Moreover, results reveal that whether in multi-stage (red line; Fig. 2A) or single stage herbivory (e.g., $a_2 = 0, a_F \geq 0$; blue line Fig. 2A), parameters' contribution to predictive power is related to their interactivity with other parameters (blue line; Fig. 2A). Note, a similar analysis with $a_2 > 0$ and $a_F = 0$ is not possible because this type of herbivory is always stable.

This interactivity was apparent in our attempts to understand how our specific parameters affected the behavior of our model in Eq. (1) via studying their effects as features in driving random forest predictions. Initial investigations into individual feature effects revealed that the effect of any single feature (parameter) on trophic dynamics could change substantially based on the values of our other features (parameters). Specifically, the average marginal effects (e.g., PD plots; Fig. S3) on simulation dynamics belied a high degree of variability in feature effects throughout the simulation data (e.g., ICE plots; Fig. S3).

Breaking down results into further subsets of set specific attack rates with varying demographic rates revealed that this variability in feature effects was largely based on the changes in feature importance and effect over different allocations of herbivory on ontogenetic stages. This breakdown affected the relationship between importance and interactivity (Fig. 2A) such that it was inconsistent but still visible in aggregate across our simulation parameters (Fig. 2B,C). Figure 2D–F depict how different allocations and intensity of herbivory across plant ontogeny change the influence of each demographic parameter in driving model stability.

Given how the influence of plant demographic rates over model behavior changed across trophic allocation (Fig. 2D–F), we first focused in depth analysis on variable demographic rates across static allocations of herbivore attack rates. By limiting the number of varying features, we use multivariate analysis to develop a fuller understanding of dynamics in subsections of the data which functioned as a scaffolding for further investigation. Specifically, we took a hierarchical approach, first developing an understanding of single-stage herbivory as a basis to study single-stage dominant herbivory (Fig. 3), which then leads us to a better overall understanding of our system's dynamics across all trophic rates.

Single stage consumption. In the case of the seedling-only herbivore (S_2; via $a_2 > 0$ and $a_F = 0$), all simulations produced stable trophic dynamics. This occurs because density loss in the seedling stage means more juveniles never reach maturity and reproduce themselves[19]. This essentially reduces the effective reproduction rate, limits the reproductive plant density, and decreases resources available to the herbivore (similar to lowering intrinsic reproduction in the classic Lotka–Volterra model). In fact, seedling herbivory only induced oscillations at higher handling times, a common effect of high handling time (results not shown).

On the other hand, concentrating consumption on the fecund stage (F) can induce both stable and oscillating trajectories (Fig. S4). Consumption of F does not induce the same regulation of reproductive potential that stabilizes under seedling-only consumption, and so is vulnerable to boom/bust populations cycles. We chose the two most consistently important (Fig. 2B) and interactive (Fig. 2C and Fig. S5) parameters, g_{12} and g_{2F}, in order to search for dominant effects on model behavior and their interactions. These parameters functioned as focal axes for our two-dimensional PD plots[36]. These PD plots depict the estimates of marginal effect of each parameter on random forest predictions, which in this case is categorical stability (Fig. 3A). We can see that stability estimates are increased by lowering either or both per-capita germination and/or maturation rates (g_{12} and g_{2F}). Demographically, reduced maturation rates shift the ratio of plant population density across its ontogeny, creating a larger juvenile population shielded from consumer pressure. Trophically, this restricts resources for the herbivore, thereby limiting losses in plant density due to herbivory (θ_F) relative to the overall plant density.

This mechanism is so influential in determining trophic dynamics, its effect on stability is statistically detectable pre-simulation via equilibrium values. Losses in plant density due to herbivory are labeled under brackets in Eq. (1) as θ_F and θ_2, which we can represent as θ_F^* and θ_2^* at equilibria. Relative to overall plant density we can define a ratio for plants of consumptive losses to total density (L:D ratio) such that:

$$\mathrm{L:Dratio} = (\theta_F^* + \theta_2^*)/(S_1^* + S_2^* + F^*). \tag{2}$$

When applied as a predictor variable on the same adult-herbivory subsection presented in Fig. 3A via a simple linear regression, we can see that L:D ratio alone explains ~ 45% of the variance of the maximum eigenvalue in simple linear models (F-statistic: 4578 on 1 and 5598 DF, p-value: < 2.2e−16) and produces an AUC score

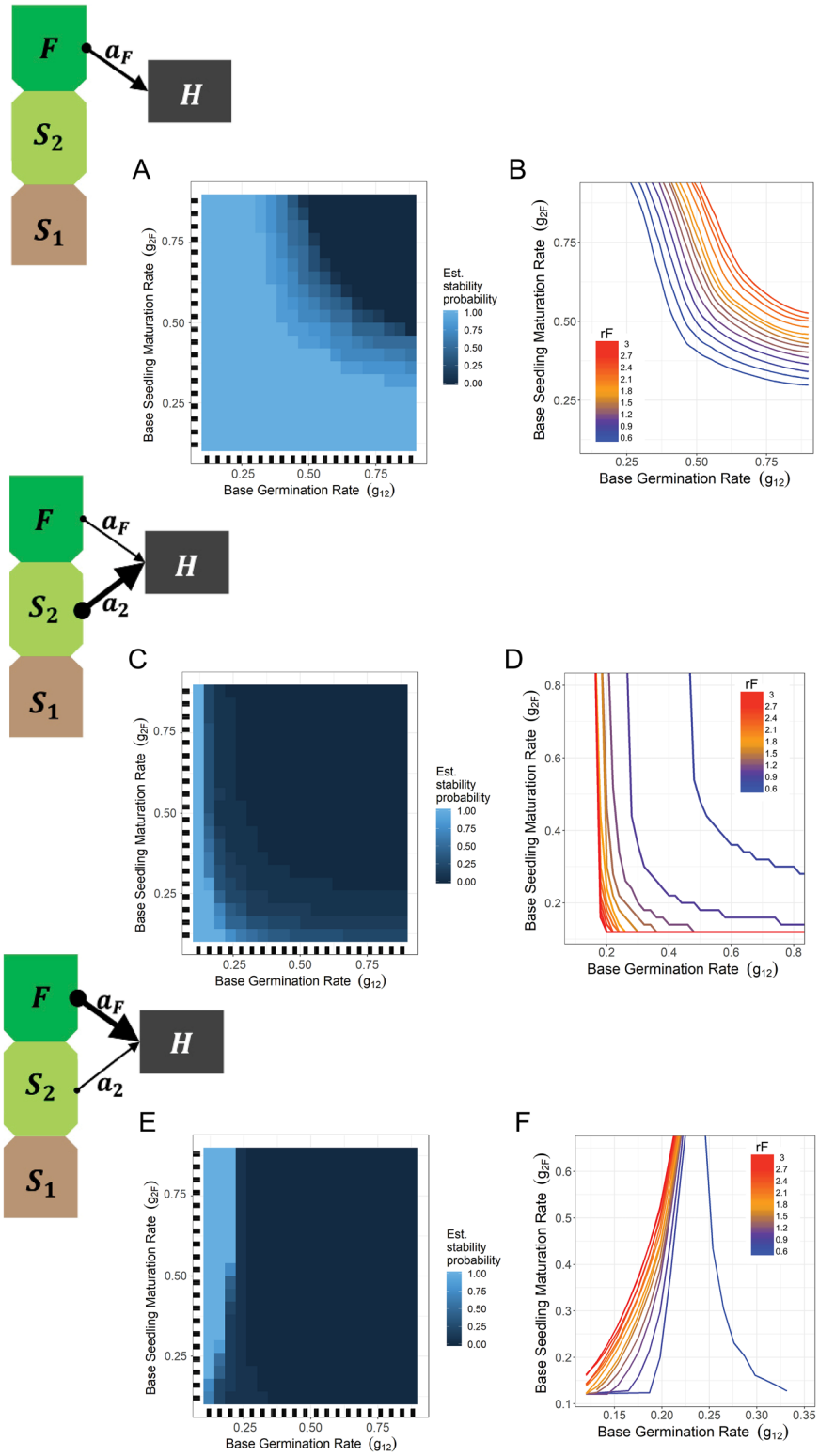

Figure 3. Interactive feature effects on model behavior. Across different herbivory allocations, partial dependence (PD) plots (**A**,**C**,**E**) show interactive effects between maturation rates on categorical simulation stability. Threshold plots (**B**,**D**,**F**) extend this analysis to include gradations of seed production rates. (**A**,**B**) Herbivory allocation $a_F = 1.0$ and $a_2 = 0.0$. (**A**) Partial dependence plot shows probability of stability across all values of r_F. (**B**) Threshold plot shows the location of the threshold between stable and unstable dynamics in $\{g_{12},g_{2F}\}$ parameter space as a function of seed production levels (r_F). (**C**,**D**) Herbivory allocation $a_F = 0.2$ and $a_2 = 1.0$. (**C**) Partial dependence plot shows probability of stability across all values of r_F. (**D**) Threshold plot shows the location of the threshold between stable and unstable dynamics in $\{g_{12}, g_{2F}\}$ parameter space as a function of seed production levels (r_F). (**E**,**F**) Herbivory allocation $a_F = 1.0$ and $a_2 = 0.2$. (**E**) Partial dependence plot shows probability of stability across all values of r_F. (**F**) Threshold plot shows the location of the threshold between stable and unstable dynamics in $\{g_{12}, g_{2F}\}$ parameter space as a function of seed production levels (r_F).

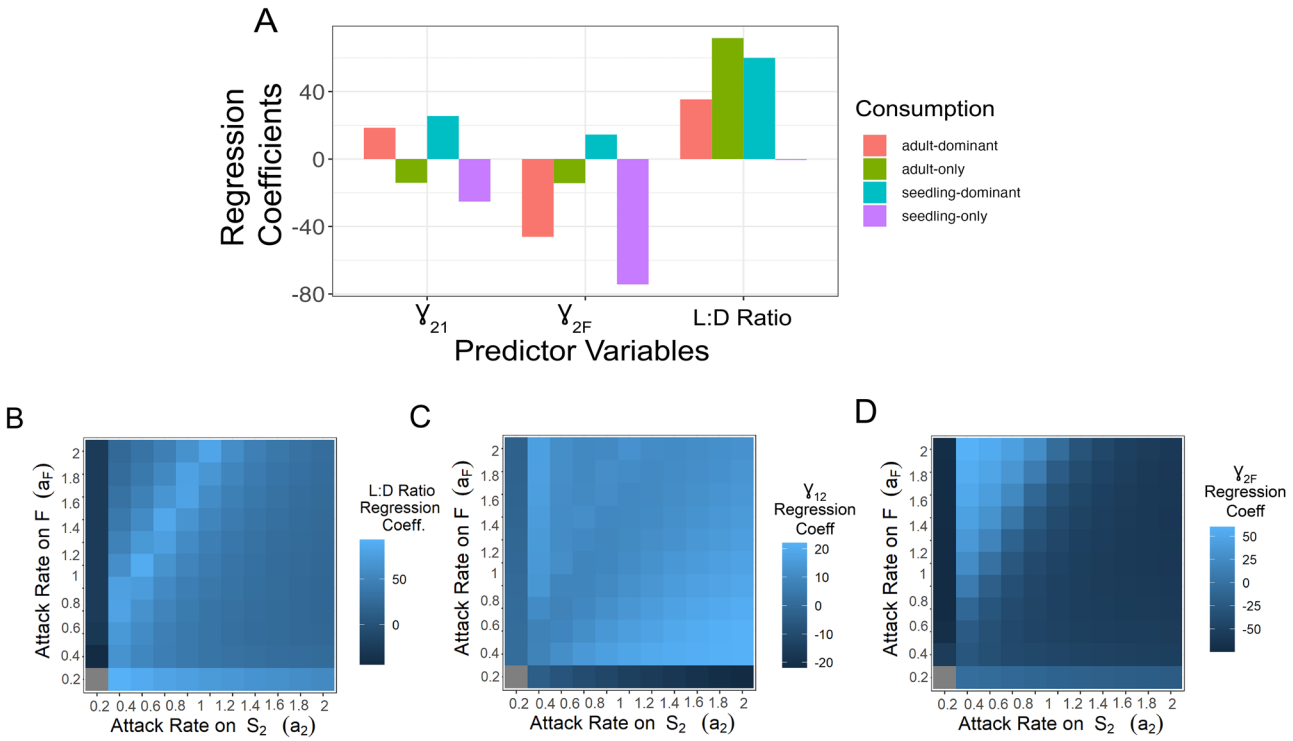

Figure 4. Ecological drivers of stability. Ecological factor effects on stability via coefficients from partial least squares regression of ecological factors versus maximum eigenvalue. Regressions are run for multiple herbivory allocations. (**A**) Bar graph showing specific changes in effect on maximum eigenvalue (regression coefficients) for each ecological factor across the four specific herbivory allocations investigated in-depth in the "Results". (**B**–**D**) Range of effects for each ecological factor on stability shown via heatmaps of regressions coefficients on maximum eigenvalue across all specific combinations of herbivory on the adult and seedling stages. Note, the gray square when both attack rates are set to 0 indicates no data given the lack of consumption.

of ~ 0.83 when predicting categorical stability. Comparatively, our random forest using simulation parameters produced an AUC of 0.98, making it clear that our L:D ratio mechanism explains some but not all the variance in stability outcomes.

Our PD plot (Fig. 3A) shows that as both g_{12} and g_{2F} increase, predictions gradually shift from stable to unstable. Based on this observation, we can make a "threshold plot" which depicts thresholds between our categorical variables, stable and unstable behavior, as a function of a third yet unexamined parameter, which in our case is seed production, r_F (Fig. 3B). Plotting the thresholds between our stability categories shows a similar dynamic between g_{12} and g_{2F} as seen in the PD plot. It also reveals that the gradual changes seen in the PD plot were in fact a function of the rate of seed production, r_F, where higher seed production supports stability at higher maturation rates. This is striking given that increased resource production is generally a destabilizing influence in the traditional Lotka-Volterra formulation. Increases in seed production are also related to increases in L:D ratio (Fig. S6), so the stabilizing effect of r_F must be coming from a different mechanism. Using a similar analysis on pre-simulation equilibrium values as was done with L:D ratio, we can integrate δ^* (see Eq. (1) and description in Table 1) into our regression given its clear connection to rF values. Doing so raises our explanatory power in our regression ($R^2 = 0.95$ in this subset of data) and our predictive power (AUC = 0.98), giving us comparable results to our random forest. However, this explains very little of the ecology given that δ^* is largely just the immediate realized effect of r_F and doesn't describe any other changes in system behavior.

Examining the effects of r_F on the rest of our system shows that r_F is also positively correlated with functions γ_{12}^* and γ_{2F}^* (see definition and description of these functions in Eq. (1) and Table 1, respectively; Fig. S6). This implies a stabilizing effect of high densities of maturing seedlings, which seems unintuitive given the stabilizing effect of lowering baseline maturation rates, g_{12} and g_{2F}. The key difference is the mechanistic difference between increasing γ_{12}^* and γ_{2F}^* via baseline maturation rates (g_{12} and g_{2F}) vs increasing overall seedling density via increases in seed production (r_F) (see Appendix S3.2 for more details). Increasing γ_{12}^* and γ_{2F}^* via baseline maturation rates (g_{12} and g_{2F}) changes the density distribution ratio in favor of F (Fig. S7a), inducing boom-bust cycles and oscillations as more plant individuals move into the adult stage and are consumed. On the other hand, increased seed production (r_F) increases overall plant density, which increases density dependent limitations on maturation, shifts the range of potential density distributions in the plant population, and saturates the younger stages, S_1 and S_2 (Fig. S7b). In adult-only herbivory, these saturated stages face no consumer pressure and therefore act as a more immediate reservoir to replace adults lost to consumption by H. The functions γ_{12}^* and γ_{2F}^* increase because there are simply more seeds and seedlings maturing. Therefore, even though increasing plant density may lead to higher overall consumption, the reservoir of density in younger stages titrates into

adult stages due to consumption, raises the trough of oscillations as r_F (and subsequently γ_{12}^* and γ_{2F}^*) increases (Fig. S8) and ultimately leads to a stabilizing consumer dynamic.

This description offers a fuller ecological explanation of how the distribution of density across plant stages mediates herbivore resource availability and drives the observed parameter context-dependence. Specifically, we observe two ways that shifting internal plant demography can promote stability. First, we observe that lowering the average per-capita maturation rates (g_{12} and g_{2F}) shields plant density in younger stages. This sequestration of plant density stabilizes by directly restricting resource availability for the herbivore and preventing overconsumption. Second, we observe that raising the intrinsic seed production rate (r_F) saturates plant density across the plant population which results in increased resource availability for the herbivore. Despite this increase in resource availability, the system is stabilized by density-dependent limitations on maturation and a robust supply of replacement plant density in younger stages which prevents overconsumption of the adult stage. The observation that increasing resource availability for the herbivore can be both stabilizing (via r_F) or destabilizing (via g_{12},g_{2F}) depending on the specific parameters underlies much of the parameter context dependence detected by our initial random forest (Fig. 2A).

Integrating the γ_{12}^* and γ_{2F}^* factors into our earlier linear regression analysis (with partial least regression due to collinearity between our variables) shows that our three factors (L:D ratio, γ_{12}^* and γ_{2F}^*) explain 91% of the variance of the maximum eigenvalue (from Fig. 3A,B) and matches our expected direction of effect (shown in Fig. 3A). Predictive power increased as well (AUC: 0.98 for categorical stability; RMSE 0.003 for regression on max eigenvalues).

Multi-stage consumption. Having established baseline dynamics, we move into multi-staged consumption by supplementing single-stage consumption with ancillary consumption on the complementary stage. Specifically, we start by supplementing a seedling-oriented herbivore with limited attack on adults ($a_2 = 1$ and $a_F = 0.2$) while revealing the importance of interactivity in parameter effects on understanding model behavior. Using this subset of simulation data, we trained a random forest and tested its predictive accuracy. As expected, random forests are sufficiently flexible to maintain high predictive accuracy despite these changes (AUC: 0.98 for categorical stability; RMSE: 0.0001 for regressions of max eigenvalue).

Simulation results reveal that the stability of the seedling-only consumer is vulnerable to destabilization from even limited multi-stage consumption (Fig. 3C,D). PD plots offer some ecological explanation in showing that oscillations can still be stabilized by restrictions in maturation rates (Fig. 3C). Extending this analysis with our threshold plots we see that the effect of r_F has effectively been reversed (Fig. 3D). Overall then, the slight addition of adult consumption institutes a relationship between g_{12} and g_{2F} that mimics an adult-only herbivore but with the caveat that higher r_F values require substantially more restricted maturation to stabilize dynamics.

We can explain this result using our earlier ecological factors (L:D ratio, γ_{12}^*, γ_{2F}^*). Compared to the adult-only herbivore, we can see that in the seedling-dominant herbivore, raising r_F has much higher proportional effect on L:D ratio compared to γ_{12}^* and γ_{2F}^* (Fig. S9). This is because L:D ratio now consists of consumption on both stages and the seedling stage can no longer act as a saturating reservoir with increased seed production. Additionally, increasing either γ_{12}^* or γ_{2F}^* via r_F (higher density) induces higher cumulative attacks (θ values; see Eq. (1) and Table 1) on both stages. This becomes clear when we regress our ecological factors on our eigenvalue data from this subset of herbivory data. We see that the effects of γ_{12}^* and γ_{2F}^* have switched from stabilizing to destabilizing (i.e., raising the max eigenvalue; Fig. 4A). Despite these changes, our linear model using our ecological factors still performs comparatively well to our random forest predictions (AUC: 0.99 for categorical stability; RMSE 0.002 ($R^2 = 0.97$) for regression on max eigenvalues).

We tested this further by supplementing an adult-oriented herbivore with a limited attack on the seedling stage ($a_2 = 0.2$, $a_F = 1$). Once again, we trained and validated a random forest on this subset of herbivory data and once again the random forest proved adept in providing accurate predictions (AUC: 0.98 for categorical stability; RMSE: 0.0002 for regressions of max eigenvalue). Similar to before, the addition of only slight amounts of multi-stage consumption qualitatively changes the relationships amongst model parameters (input features) and stability. PD plots show that lower g_{12} is again stabilizing. However, it also reveals stability at lower g_{12} values is now more dependent on *higher* g_{2F} (Fig. 3E). Extending the analysis with our threshold plots again indicates that higher seed production values (r_F) limit stable {g_{12}, g_{2F}} parameter space (Fig. 3F).

Similar to seedling-oriented herbivory, this multi-stage consumption limits the function of saturating reservoirs in the seedlings and induces more instances of oscillatory dynamics compared to single stage consumption on the fecund stage. However, using our ecological factors, we can investigate the demographic conditions which still promote stability. The stability found at high g_{2F} and low g_{12} can also be described as high γ_{2F}^* and low γ_{12}^* with exact values contingent upon the seed production rate (Fig. S10a). These demographic conditions reduce the composition of plants in the seedling stage. This limits herbivore consumption on the seedlings (Fig. S10b) and causes the interaction to function more like single stage consumption. This functional similarity to single stage consumption on the adult stage means this promotes the stabilizing effect of a reservoir in the seed bank (low g_{12}) and a high replacement rate of adults (high g_{2F}). Consequently, simulation results show higher rates of stability at these conditions (Fig. S10c) and the coefficients from our partial least squares regression correspond with our analysis (Fig. 4A). Additionally, our ecological factors once again not only aid in explaining our results but also perform well as predictors (AUC: 0.99 for categorical stability; RMSE: 0.003 ($R^2 = 0.99$) for regressions of max eigenvalue).

In fact, across all of unique herbivory allocations, our ecological factors perform well as predictors in both categorical (mean AUC: 0.99) and regression based (mean RMSE: 0.002) linear models. The accuracy of our predictions, though not equivalent to causation, affords some confidence in using our partial least squares regression coefficients (on max eigenvalues) to investigate how the effects of each ecological factor change across different

allocations of herbivory (Fig. 4B–D). With this we get a detailed view of how plant demography interacts with trophic rates to drive dynamics of our trophic interaction. These results were generally qualitatively consistent across handling times and strength of density dependence (see Table 1) with a notable exception when handling times for seedling consumption are smaller than for adult consumption (see Fig. S11).

Finally, expanding back out and considering the full simulation dataset reveals our ecological factors demonstrate predictive accuracy even across all demographic and trophic rates. Using our linear partial least squares model on maximum eigenvalues showed modest success (mean RMSE = 0.012 across all permutations of handling time and density dependence). Categorically predicting stability using our factors and the binomial regression was comparatively more successful (mean AUC = 0.91 across all permutations of handling time and density dependence).

Discussion

Increasingly realistic model frameworks in ecology present increasingly complex datasets and novel challenges in analysis. Machine learning algorithms are an obvious candidate for addressing analytical challenges. While machine learning algorithms still have limits in their interpretability, their ability to produce highly accurate predictions bolster their increasing prevalence in ecology[11]. Stopping at predicting ecological model outcome alone, however, places high credence in simulation model formulation if no mechanistic explanation behind the predicted outcome is possible. Famously, "all models are wrong"[37] but their utility lies in their ability to aid researcher's intuition of complex systems. Therefore, mechanistic explanations of model results are critical to achieving models' maximum utility in ecology. Accordingly, there is fundamental value in expanding the interpretability of machine learning (e.g., random forests) in studying simulation models which we argue connects to the core utility of modeling in science.

The variety of ecological behavior across organismal ontogeny is an important frontier in ecological modeling[16]. The context-dependent complexities resulting from even our simple inclusions of ontogeny make clear the need to develop broadly applicable methodologies which improve generalizable ecological understanding. Our random forest output produced clear descriptions of each simulation model parameters' contribution to predicting simulation behavior and Friedman's H-statistic analysis showed that these contributions were largely tied to interactions with other model parameters (both demographic and trophic; see Fig. 1, Fig. S5). Individual Conditional Expectation plots revealed the extent of this context-dependency (e.g., ICE plots Fig. S3), but our goal was to go beyond the well-known ecological axiom that parameter effects are context-dependent. We therefore subdivided our data into its component parts, focusing on the consistently important model parameters and it became clear that different parameters drove model behavior across different herbivory allocations (Fig. 2D–F). By investigating feature interactions using two-dimensional PD plots and then extending categorical analysis with our threshold plots (Fig. 3), we were able to develop intuition of how simulation parameters interact to drive model behavior. This intuition can serve as the necessary ingredients to develop the specific hypotheses required to verify our results with interpretable regression analyses. Specifically, using GAMs to test for effects from the parameter interactions revealed by the random forest analysis corroborates the results detailed in Fig. 3 (e.g., Fig. S12). Verifying our specific results via GAM is another avenue to increase our confidence in their validity (in addition to the cross validation we preformed), but the random forest approach was the preferable data exploration approach in our case given its significant advantage in finding higher order features and interactions at superior computing speed. Regardless, understanding and verifying these interactions between simulation parameters moved our analysis beyond simply connecting model behavior to parameter values to describing model output via tangible ecological processes.

For ease of communication, we called our ecological processes (L:D ratio, γ_{12}^{*}, and γ_{2F}^{*}) "factors." Using these factors, we detailed how demographic and trophic rates interact to drive ontogenetically mediated consumer-resource dynamics via a consistent set of ecological mechanisms. These factors were also able to predict model output with comparable accuracy to that of the random forest. Furthermore, despite being found from categorically tasked random forests (stable or unstable), these same factors performed well on continuous predictions (maximum eigenvalues) given the relatedness of each variable.

In expanding upon our work to higher dimensionalities found in food web/network models, several considerations come to mind. First, random forests facilitate intuitive feature selection which allows researchers to deal with higher parameter counts by focusing on the most important simulation model parameters (via accuracy decreases, see "Methods"). Additionally, subsetting simulation data (as we did above by focusing on specific herbivory allocations) simplifies analysis and further facilitates finding mechanistic drivers of model dynamics. In cases where parameterization is systematically controlled and not varied (e.g., the Allometric Trophic Network), analysis of model behavior can easily be refocused onto quantitative network properties via random forest inputs.

Finally, while our analysis used equilibria and linear stability metrics, research at the level of food web analysis typically focuses on other metrics of model behavior. This is partially because finding and expressing equilibria in high dimensional models can be difficult. However, it has been shown to be tractable[21] when using root finding algorithms which do not rely on explicit parametric expressions[38], but instead provide the numerical values of equilibria which would be sufficient for our purposes. On the other hand, other metrics frequently used at the food web/network scale, such as species extinctions[6], temporal variation[39], biomass production[14] can readily fit the random forest framework. Additionally, by establishing threshold cutoffs of interest akin to our dual use of categorical stability and maximum eigenvalues, researchers can couple classification and regression tasks to maximize their ability to discover mechanistic drivers of model behavior via the processes detailed here.

We do not claim our exact methodology will be entirely applicable for all models or questions. Instead, we aimed to present an example process extending the analytical power of random forests from prediction to mechanistic explanation in complex parameter space. Indeed, there are global sensitivity analysis techniques aimed

at analyzing model sensitivity across broad spectrums of parameter space that researchers can consider at the onset of model analysis. Random forests do carry some inherent advantages given their ability to implicitly deal with correlation and high dimensional data, identify informative inputs with relatively fast permutation based variable importance indices, and handle interactions between input parameters. Furthermore, in our case, the interpretability tools available via random forests (and general machine learning) were also key to our ability to develop our ecologically based understanding of our model's output. For comprehensive comparisons between global sensitivity analyses and random forest techniques, we refer the reader to these references[40,41]. In the case of stage-structured ontogeny, our process revealed that complex interactions amongst parameters could be consolidated into key ecological mechanisms. These mechanisms explained how trophic interactions mediated by plant ontogeny can produce disparate results compared to traditional model structure. The extent of these differences exposes the need to further integrate unique sources of ecological complexity to better understand drivers of community dynamics. Such work also presents opportunities to expand the role of machine learning in ecology, both for random forests and machine learning algorithms with more advanced causal properties.

Data availability

All COMPADRE data used in our model formulation is available on the COMPADRE platform and the exact specifications are reproducible using Supplementary File 1: CompadreAnalysisRMD.html (download zipped file locally, extract files, and open in any internet browser). Simulation data used in our analysis is available as a zipped file online https://drive.google.com/drive/folders/1hjU3fe0IEpthVNEDkpQL8DHp2K7sMaDk.

Code availability

The simulation model (Eq. 1) as well as R scripts for all random forest analysis, main text figures/analysis, and supplementary figures are available at https://github.com/prglaum/PlantOntogenyDynamics. Additionally, random forest analysis is available as an R Markdown file in Supplementary File 1 (randomForestCode.html; download zipped file locally, extract files, and open in any internet browser) and as an online interactive app for readers at https://prglaum.shinyapps.io/simData-randomForestRMD/.

Received: 17 August 2022; Accepted: 21 February 2023
Published online: 08 March 2023

References

1. McCann, K. S. The diversity–stability debate. *Nature* **405**, 228–233 (2000).
2. Namba, T. Multi-faceted approaches toward unravelling complex ecological networks. *Popul. Ecol.* **57**, 3–19 (2015).
3. Allesina, S. & Tang, S. The stability–complexity relationship at age 40: A random matrix perspective. *Popul. Ecol.* **57**, 63–75 (2015).
4. McCann, K., Hastings, A. & Huxel, G. R. Weak trophic interactions and the balance of nature. *Nature* **395**, 794–798 (1998).
5. Kondoh, M. Foraging adaptation and the relationship between food-web complexity and stability. *Science* **299**, 1388–1391 (2003).
6. Brose, U., Williams, R. J. & Martinez, N. D. Allometric scaling enhances stability in complex food webs. *Ecol. Lett.* **9**, 1228–1236 (2006).
7. Pimm, S. L. & Lawton, J. H. On feeding on more than one trophic level. *Nature* **275**, 542–544 (1978).
8. Hale, K. R. S., Valdovinos, F. S. & Martinez, N. D. Mutualism increases diversity, stability, and function of multiplex networks that integrate pollinators into food webs. *Nat. Commun.* **11**, 2182 (2020).
9. Grilli, J., Barabás, G., Michalska-Smith, M. J. & Allesina, S. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017).
10. Prowse, T. A. A. *et al.* An efficient protocol for the global sensitivity analysis of stochastic ecological models. *Ecosphere* **7**, e01238 (2016).
11. Lucas, T. C. D. A translucent box: Interpretable machine learning in ecology. *Ecol. Monogr.* **90**, e01422 (2020).
12. Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. & Hartig, F. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.* **11**, 281–293 (2020).
13. Wardeh, M., Blagrove, M. S. C., Sharkey, K. J. & Baylis, M. Divide-and-conquer: Machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations. *Nat. Commun.* **12**, 3954 (2021).
14. Ceulemans, R., Guill, C. & Gaedke, U. Top predators govern multitrophic diversity effects in tritrophic food webs. *Ecology* **102**, e03379 (2021).
15. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?': Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016). https://doi.org/10.1145/2939672.2939778.
16. Valdovinos, F. S. *et al.* A bioenergetic framework for aboveground terrestrial food webs. https://doi.org/10.22541/au.165836510.07094278/v1 (2022).
17. Miller, T. E. X. & Rudolf, V. H. W. Thinking inside the box: Community-level consequences of stage-structured populations. *Trends Ecol. Evol.* **26**, 457–466 (2011).
18. von Euler, T., Ågren, J. & Ehrlén, J. Environmental context influences both the intensity of seed predation and plant demographic sensitivity to attack. *Ecology* **95**, 495–504 (2014).
19. Glaum, P. & Vandermeer, J. Stage-structured ontogeny in resource populations generates non-additive stabilizing and de-stabilizing forces in populations and communities. *Oikos* **130**, 1116–1130 (2021).
20. Rudolf, V. H. W. & Lafferty, K. D. Stage structure alters how complexity affects stability of ecological networks. *Ecol. Lett.* **14**, 75–79 (2011).
21. de Roos, A. M. Dynamic population stage structure due to juvenile–adult asymmetry stabilizes complex ecological communities. *Proc. Natl. Acad. Sci.* **118**, e2023709118 (2021).
22. Jones, O. R. *et al.* Rcompadre and Rage—Two R packages to facilitate the use of the COMPADRE and COMADRE databases and calculation of life history traits from matrix population models. https://doi.org/10.1101/2021.04.26.441330 (2021).
23. Futuyma, D. J. & Gould, F. Associations of plants and insects in deciduous forest. *Ecol. Monogr.* **49**, 33–50 (1979).
24. Forister, M. L. *et al.* The global distribution of diet breadth in insect herbivores. *Proc. Natl. Acad. Sci.* **112**, 442–447 (2015).
25. Fowler, S. V. Differences in insect species richness and faunal composition of birch seedlings, saplings and trees: The importance of plant architecture. *Ecol. Entomol.* **10**, 159–169 (1985).
26. Kearsley, M. J. C. & Whitham, T. G. Developmental changes in resistance to herbivory: Implications for individuals and populations. *Ecology* **70**, 422–434 (1989).

27. Basset, Y. Communities of insect herbivores foraging on saplings versus mature trees of Pourouma bicolor (Cecropiaceae) in Panama. *Oecologia* **129**, 253–260 (2001).
28. Stiegel, S. & Mantilla-Contreras, J. Environment vs. plant ontogeny: Arthropod herbivory patterns on European beech leaves along the vertical gradient of temperate forests in Central Germany. *Insects* **9**, 9 (2018).
29. Quintero, C. & Bowers, M. D. Plant and herbivore ontogeny interact to shape the preference, performance and chemical defense of a specialist herbivore. *Oecologia* **187**, 401–412 (2018).
30. Coley, P. D. & Barone, J. A. Herbivory and plant defenses in tropical forests. *Annu. Rev. Ecol. Syst.* **27**, 305–335 (1996).
31. Cutler, D. R. *et al.* Random forests for classification in ecology. *Ecology* **88**, 2783–2792 (2007).
32. Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform.* **20**, 492–503 (2019).
33. Breiman, L., Cutler, A., Liaw, A. & Wiener, M. Random Forest: Breiman and Cutler's random forests for classification and regression. 4.6–7 (2022).
34. Friedman, J. H. & Popescu, B. E. Predictive learning via rule ensembles. *Ann. Appl. Stat.* **2**, 916–954 (2008).
35. Molnar, C., Casalicchio, G. & Bischl, B. iml: An R package for interpretable machine learning. *JOSS* **3**, 786 (2018).
36. Greenwell, B. M., Boehmke, B. C. & McCarthy, A. J. A simple and effective model-based variable importance measure. http://arxiv.org/abs/1805.04755 (2018).
37. Box, G. E. P. Science and statistics. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).
38. Soetaert, K. *et al.* Nonlinear Root Finding, Equilibrium and Steady-State Analysis of Ordinary Differential Equations (2009).
39. Glaum, P., Wood, T. J., Morris, J. R. & Valdovinos, F. S. Phenology and flowering overlap drive specialisation in plant–pollinator networks. *Ecol. Lett.* **24**, 2648–2659 (2021).
40. Antoniadis, A., Lambert-Lacroix, S. & Poggi, J.-M. Random forests for global sensitivity analysis: A selective review. *Reliab. Eng. Syst. Saf.* **206**, 107312 (2021).
41. Stein, B. V. *et al.* A comparison of global sensitivity analysis methods for explainable AI with an application in genomic prediction. *IEEE Access* **10**, 103364–103381 (2022).

Acknowledgements

This research was supported by National Science Foundation grants DEB-1834497 and DEB-2129757 to F.S.V. S.S. was partially supported by the UC Davis Grant for Advancing Sustainable Development Goals, Global Affairs as awarded to F.S.V and the Sustainable Oceans National Research Trainee Fellowship (NSF award # 1734999) as awarded to S.S. This research was supported in part through computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor.

Author contributions

S.M.S., P.G., F.S.V. conceived and conceptualized this project. S.M.S. and P.G. developed the simulation model. P.G. conceived and implemented the random forest analysis and extension. P.G. and S.M.S. developed and implemented the linear analysis. P.G. and S.M.S. wrote the first draft of the manuscript. S.M.S., P.G. and F.S.V. contributed to the writing of the following versions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30313-8.

Correspondence and requests for materials should be addressed to S.M.S. or P.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.