Received: 3 December 2025 | Accepted: 19 May 2026

DOI: 10.1111/2041-210x.70337

Methods in Ecology and Evolution

RESEARCH ARTICLE

Challenging and diagnosing structured population models by testing predictions from stochastic demography

Stephen P. Ellner[1] | Robin E. Snyder[2] | Daniel T. Blumstein[3,4] | Dylan Z. Childs[5] | Joshua C. Fowler[6] | Christina M. Hernández[7] | Julien G. A. Martin[4,8] | María Paniw[9] | Yngvild Vindenes[10]

[1]Cornell University, Ithaca, New York, USA; [2]Case Western Reserve University, Cleveland, Ohio, USA; [3]University of California, Los Angeles, Los Angeles, California, USA; [4]Rocky Mountain Biological Laboratory, Crested Butte, Colorado, USA; [5]Sheffield University, Sheffield, UK; [6]University of Miami, Coral Gables, Florida, USA; [7]University of Oxford, Oxford, UK; [8]University of Ottawa, Ottawa, Ontario, Canada; [9]Estación Biológica de Doñana, Sevilla, Spain and [10]University of Oslo, Oslo, Norway

Correspondence
Robin E. Snyder
Email: res29@case.edu

Present address
Joshua C. Fowler, University of Colorado, Boulder, Colorado, USA

Christina M. Hernández, Old Dominion University, Norfolk, Virginia, USA

Funding information
Division of Environmental Biology, Grant/Award Number: DEB-1933497 and DEB-1933612; Ministerio de Ciencia, Innovación y Universidades, Grant/Award Number: PID2022-141004OA-I00; National Science Foundation, Postdoctoral Fellowship, Grant/Award Number: 2410282; European Social Fund Plus, Grant/Award Number: RYC2021-033192-I

Handling Editor: Res Altwegg

Abstract

1. Structured population models are parameterized to accurately project expected population sizes, stage/state distributions and population growth rates, but they also predict the variation in outcomes among individuals, such as the variance and skewness of lifetime reproductive output (LRO) and lifespan, the probability of never reproducing, and many other life-history metrics.
2. Testing such predictions about individual outcomes can be very useful model 'stress tests', because they depend on how components of the model (e.g. submodels for survival and fecundity) interact over multiple time steps, not just on the accuracy of the submodels. Because data on among-individual variation is rarely used to parameterize the models, models will not automatically pass the tests.
3. We present case studies (including zooplankton, plants and mammals) to demonstrate how structured population models can be tested by comparing individual-level predictions from existing models against individual-level data.
4. Some general themes emerge: (i) We often detect unmodelled individual heterogeneity, (ii) Unmodelled senescence can affect higher moments of lifespan even when lower moments and LRO are predicted well. (iii) Fitting one parametric model to multiple clones, species, locations, etc. can lead to poor predictions about groups for which the model is insufficiently flexible.
5. The ways in which model predictions fail can help to identify what the problems are, help us decide whether the problems are important for the model's intended purpose, and guide efforts to fix them. Structured population models are 'workhorses' for ecology: tests based on predictions from stochastic demography can help ensure their reliability.

KEYWORDS
Daphnia, endophytes, grass, marmots, model testing, rotifers, Soay sheep, stochastic demography

1 | INTRODUCTION

Ecologists often rely on models to predict things that we want to know but cannot observe directly. Structured population models have become workhorses of population and conservation ecology because they can project unobservables such as long-term population growth or decline rates, risk of extinction or quasi-extinction, and the sensitivity of those and other properties to potential management interventions or possible future changes in the environment. Because the predictions of interest are usually unobservable, models are typically 'validated' by assessing the adequacy of the submodels for individual demographic rates (survival, growth, fecundity, etc.). Is survival just a function of size, or does age also matter (Caswell, 2001, Ch. 3)? Does the growth model adequately capture how initial size affects the mean, variance, skewness and kurtosis of annual growth (Miller & Ellner, 2025)? There is no cut-and-dried approach for these assessments, but there are familiar customs and lore, and helpful general statistical tools for model selection.

However, what the full model predicts about the population also depends on how the submodels are connected, and on how they interact over multiple years. Most of the all-too-common errors in model specification discussed in Kendall et al. (2019) and Che-Castaldo et al. (2020) have to do with putting the pieces together incorrectly, such as omitting or incorrectly including dormant life stages, or confusing pre- versus post-breeding census. Even if you accurately model adult and neonate survival, the model is in trouble if you put them in the wrong place, or in both right and wrong places. Separate tests of fecundity and survival regressions will not reveal if survival–growth correlation (or its absence) is correctly modelled.

Additionally, re-use of published models for new purposes is increasingly popular, as attested by the scale and scope of the COMPADRE and COMADRE databases of plant and animal matrix projection models (MPM, Caswell, 2001; Jones et al., 2022; Salguero-Gomez et al., 2015; Salguero-Gomez et al., 2016) and the PADRINO database of integral projection models (IPM, Ellner et al., 2016; Levin et al., 2022): over 12,000 site-, year- or treatment-specific MPM matrices for over 1200 unique species and over 250 IPM kernels for more than 50 unique species. Even if the original study worked hard on model validation, every model is created for a purpose, and how the model was constructed and validated will have reflected that purpose. A new purpose calls for a new validation.

The long-term consequences of the interactions among demographic submodels are less straightforward to test than the individual submodels. Ideally, the overall model should be tested by comparing full-model predictions to data that were not used in fitting the model.

The predictions of a structured population model are the aggregate consequence of the fates of individuals: how many live or die, grow or shrink, breed or not. Consequently, those models can also predict the life course of individuals one at a time, either by stochastic simulations or using analytic methods from Markov Chain theory. By interpreting a structured population model as a recipe for each individual's random path through potential state transitions, we can predict (for example) the mean, variance and skewness of lifespan and lifetime reproductive output (LRO), the probability of breeding at least once, mean age of individuals in each life stage and many other life history metrics (e.g. Broekman et al., 2020; Caswell, 2001; Cochran & Ellner, 1992; Ellner et al., 2016; Snyder et al., 2021; Snyder & Ellner, 2016, 2018; Steiner et al., 2010; Steiner & Tuljapurkar, 2012; Tuljapurkar et al., 2009, 2020, 2021; van Daalen & Caswell, 2017).

In this paper, we argue that predicting among-individual variation, starting with predicting the variance and skewness of lifespan and LRO, is a particularly useful 'stress test' for many of the purposes for which structured population models are used. At a minimum, the higher moments of LRO and lifespan rely on the combined effect of multiple demographic processes, and among-individual variation is not normally used in fitting a model. But beyond this, many of the questions we seek to answer with population models will be affected by the among-individual variability in LRO and lifespan. For example, in many of our cases studies, the 'stress tests' reveal unmodelled individual heterogeneity. Within-population individual heterogeneity has been a major research theme in modern population ecology (Gamelon et al., 2025), so we now recognize that it can affect many population-level properties such as the magnitude of demographic stochasticity, extinction risk in small populations, metapopulation viability, population growth rate and rate of spatial spread (e.g. Fox & Kendall, 2002; Gibert, 2016; Kendall & Fox, 2002; Sorel et al., 2024; Steiner et al., 2021; Stover et al., 2014; Vindenes et al., 2012; Vindenes & Langangen, 2015). In epidemiological models, individual variation in contact rates is a key determinant of the pathogen net reproduction rate R_0 and therefore of disease persistence versus die-off (e.g. Antonovics, 2017). For studies involving any of those, unmodelled individual heterogeneity may well be a problem, depending on its magnitude. Other combinations of model goals and problems are less well studied, and case-specific analyses may be needed.

Our goal here is to demonstrate through case studies how comparisons of individual-level predictions against individual-level data can be used to test a structured population model and how the ways in which predictions fail can help us diagnose the model's deficiencies, help us to determine whether or not the identified problems are important for the model's intended applications and if the problems are important, guide our efforts to fix them. If a model does not successfully predict moments of lifespan and LRO, the other life-history calculations suggested above (such as the fraction of individuals

breeding at least once) can be useful as further diagnostics. As with testing demographic submodels, we cannot offer a cut-and-dried approach, but we hope to build up some customs and lore. If a model fails to predict some of the moments of LRO or lifespan, there are several ways to investigate potential causes, which we summarize in Box 1. We use all of these approaches in our case studies. We begin with detailed models for two laboratory zooplankton populations and for seven grass species in outdoor experiments, and then consider models for observational studies of several natural populations. The choice of case studies was driven by the availability of individual-level lifetime data, existence of a published population model with posted code (to avoid introducing errors when re-implementing a model) and model creators being open to having their published work challenged.

Major model mis-specifications (Che-Castaldo et al., 2020; Kendall et al., 2019) may already affect population-level predictions that can be compared with data, such as population growth rates or size/stage distributions during the study period. When used as a supplement to such tests, those we present here are more likely to reveal problems with the simplifying assumptions and compromises of model construction, such as omitting senescence. But population-level predictions can only be tested if you have population-level data. If only a small fraction of the population was marked and monitored, tests based on individual-level predictions may be the only way to check for even the most gross errors in model construction.

2 | GENERAL METHODS

For all datasets, for testing predictions about lifetime outcomes, we removed individuals who were born before the study started or who were still alive at the end of the study, because we were unable to observe their entire life's course. We also removed any individual who was not observed in the first lifestage—for example, we removed marmots who were not observed as juveniles and grasses that were first recorded as some age older than 0. We removed any pairs of individuals who were given the same ID. Empirical moments of life outcomes such as LRO were computed using the standard formulas as provided by the R functions `mean, var, skewness`; the same was done for simulated individuals.

Whenever possible, exact theoretical predictions for life-history traits (e.g. moments of LRO and lifespan) were generated using analytic results on Markov Chains with rewards from the papers cited in the Introduction. This was possible for laboratory studies with a constant environment (rotifers, *Daphnia*) and field studies where the population model was time-invariant (Soay sheep). The 'input' to these methods is a set of matrices R_1, R_2, R_3, that contain the first, second and third moments, respectively, of the reward associated with making each possible transition, including the final transition to death. For example, for calculating moments of LRO, the entry in row 5, column 2 of R_2 is the the expected value of $(\text{clutch size})^2$ for an individual that transitions from state 2 to state 5. For calculating moments of lifespan, every entry in all the reward matrices is 1: age goes up by exactly 1 every year. The outputs are the predicted first, second and third moments of variation in LRO and lifespan across individuals in the population. For further details, see van Daalen and Caswell (2017).

For models that included year-specific environmental parameters (grasses, marmots), we generated theoretical predictions by simulating the stochastic individual-based models (IBMs) implied by the transition matrix or kernel and any clutch size distributions not specified by the model. For example, if a matrix model has a survival rate of 60% for juveniles in year 3 of the study, then in the associated IBM, each juvenile individual survives from year 3 to year 4 with probability 0.6, simulated with an independent random 'coin toss' for each individual (using pseudo-random numbers). Unless stated otherwise in the case studies below, individual reproduction was assumed to follow a Poisson distribution if mean clutch size was above one, and a Bernoulli distribution if mean clutch size was below one. However, to most closely simulate the empirical study, in each simulated year, we added to the population a number of newborns equal to the observed number born that year, and recorded their lifespans and reproductive history. We only follow the life courses of their offspring if those individuals were also in the study dataset. As with the empirical data, we removed any individual who was still alive at the end of our simulated study. We simulated multiple replicate populations (100 unless otherwise noted), and if model parameters are given as a joint posterior distribution rather than point estimates (e.g. as in our marmot case study based on Paniw et al., 2020), we made a new draw from the posterior distribution to generate the parameter vector used for each replicate population. We calculated the mean, variance and skewness of LRO and lifespan for each population as well as the standard deviations of these across replicate simulations.

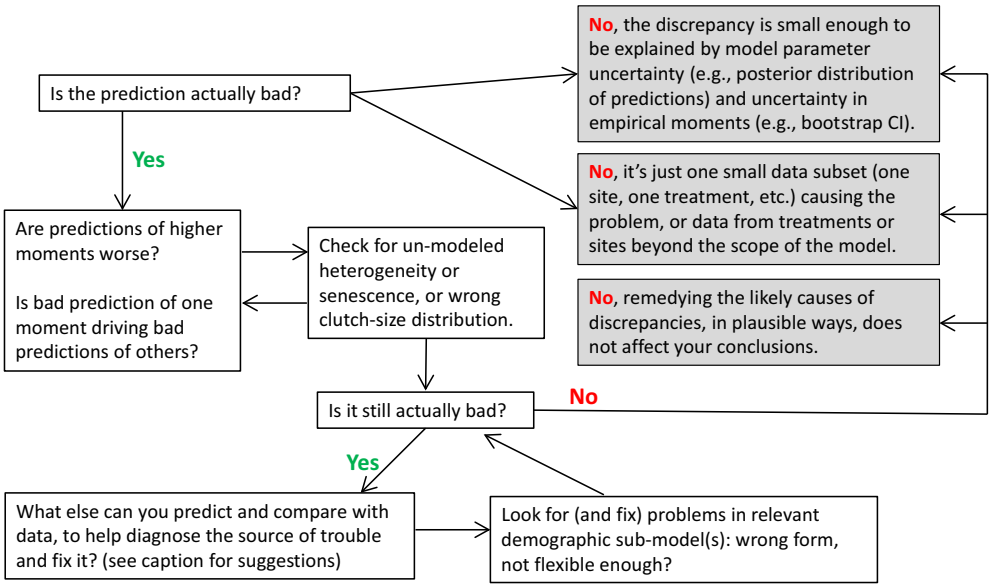
Our archived code `simulators.R` includes functions for simulating IBMs based on a list of yearly transition matrices, expected fecundity as a function of state and the number of births each year. We describe this code in Section S.4. Note that while all of our case studies use density-independent models, it is possible to apply these methods to density-dependent models by calculating moments via simulation.

3 | ROTIFERS, *Brachionus manjavacas*

3.1 | Data

The data for this case study are from a laboratory study of the marine monogonont rotifer *Brachionus manjavacas* (Bock et al., 2019; Hernández et al., 2020b). Female neonates of known maternal age of 3, 5, 7 or 9 days (and uniform grandmaternal and great-grandmaternal age) were grown individually in seawater, in wells of 24-well plates with chlorophyte algae for food. Survival and the number of live offspring were recorded once daily for all individuals, and the female was then transferred to a new well with fresh seawater and algae. Data were treated as right-censored for individuals accidentally killed.

BOX 1 Suggested workflow for challenging, diagnosing and improving structured population models using stochastic demography predictions. We use all of the approaches described here in our case studies; italic text highlights such examples



1. Is the prediction actually bad?
   a. Empirical moments of lifespan and LRO are always uncertain due to sampling variability. Standard errors can be generated for empirical moments by bootstrapping from the data, and for theoretical moments by using the fitted model to simulate multiple runs of the data-generating experiment and sampling procedures; *Daphnia*
   b. Whenever possible, take into account uncertainty in *predicted* moments due to parameter uncertainty, for example by drawing parameters from posterior distributions; *Marmots*, *grasses*
   c. If your data comes from multiple sources (e.g. multiple treatments or locations), are a few of them outliers responsible for poor overall prediction (*Rotifers*, *marmots*) or outside the scope of situations to which the model was fitted? *Grasses*
   d. Having diagnosed the likely cause of the discrepancies, are your conclusions robust against plausible ways of changing the model to remove the discrepancies?
2. In what way are things going wrong?
   a. Are predictions for higher moments worse than those for lower moments? This suggests the model is not capturing the tail of some distribution, which can happen for multiple reasons.
   b. Are bad predictions for some moments responsible for the bad predictions for others? What is really the root of the problem? Are bad predictions for lifespan causing bad predictions for LRO? Or is a bad prediction for the mean driving a bad prediction for a higher moment? Plot the badly predicted variable(s) as a function of others (e.g. LRO vs. lifespan) to see if there is a tight relationship, so that poor predictions of one drive poor predictions of others. *Daphnia*, *marmots*
3. Examine the implicated potential problems
   a. Unmodelled heterogeneity: plot moments versus your suspected source of heterogeneity (e.g. location). Plotting moments against time in stage may reveal within-stage heterogeneity (i.e. the model has oversimplified the life cycle). Try re-fitting model components (e.g. survival, fecundity) with an individual random effect on the intercept. *Rotifers*, *marmots*
   b. Unmodelled senescence: Does your model allow for unrealistically long lifespans? Try removing any simulated individuals surviving longer than the oldest observed individual. *Soay sheep*, *marmots*
   c. Wrong clutch size distribution:
      - Is the actual clutch size distribution more zero-inflated than you assume? *Marmots*
      - Try calculating moments via Markov Chains with Rewards (see Section 2) using the observed second and third moments for age, size or state-dependent clutch size instead of those for an assumed clutch size distribution. Use the observed mean too, if you suspect the fecundity model is wrong. *Rotifers*

4. Is there still a problem? If so, dig deeper.
   a. Is the model form wrong, or insufficiently flexible? Try plotting empirical moments of fecundity or survival or growth versus stage or size, and compare with model predictions. *Soay sheep*.
   b. What else can you predict and compare with available data to help identify the underlying problem(s)? This is case-specific, as it depends on what data are available and open-ended, because any life-history attribute can be predicted by individual-based simulations, if not analytically. Consider predicting:
      - the probability of dying in the first year (lifespan = 1); *Marmots*
      - the probability of never having offspring (LRO = 0); *Marmots*
      - mean size or mean clutch size versus age, Figure 8e, Figures S-6, S-7, S-9
      - moments and distribution of age and size at death (Figure 8c,d)
      - various moments conditional on surviving the first year, Figure S-3
      - demographic submodel predictions (e.g. fecundity or survival as a function of size; Figures S-4, S-5 and S-8)
      - complete distributions of LRO or longevity.

   Some of these can be calculated analytically (see Snyder & Ellner, 2016) for conditional transition matrices, Ellner et al. (2016, ch. 3) for size, age or number of offspring at death, Tuljapurkar et al. (2020) and Tuljapurkar et al. (2021) for complete LRO/longevity distributions). Otherwise, these can be calculated from IBM simulations of the population, using the functions in `simulators.R` in our code archive; see Section S.4 for a description of the code and how to use it. Also, see the archived code associated with our Supporting Information figures for examples.

3.2 | Model

Hernández et al. (2020a) constructed a matrix model (Caswell, 2001) with individuals cross-classified by age and maternal age. Maternal age is a static trait fixed at birth, and age transitions are, tragically, entirely predictable. Survival probability and mean fecundity were estimated by fitting a Weibull survival model and a Coale–Trussell fertility model to the complete experimental data set. To construct the matrix model, Hernández et al. (2020a) extrapolated these demographic models to all maternal ages from 1 to 16 days. For full details, see Hernández et al. (2020a). For our analyses, we regarded maternal age as four distinct treatments and analysed them separately.

3.3 | Results

Using the separate fitted matrix models for each maternal age, theoretical predictions for the mean LRO and for the mean, standard deviation and skewness of lifespan are all quite good (Figure 1a,d–f; for lifespan skewness, note that the prediction errors are all numerically small). Predictions for the standard deviation and skewness of LRO were much less successful (note that the 1:1 line is outside the plotted region in panel B). Note, a negative value for the fraction of variance explained by the predictions means that the mean square error of the predictions is larger than the variance of the observed values.

The predictions about LRO variation in Figure 1 rely on the parametric statistical models for fecundity fitted by Hernández et al. (2020a), and also on the typical assumption of Poisson-distributed clutch sizes. To see if those two assumptions might underlie the poor predictions, we repeated the predictions about LRO two ways (Figure 2). The first retained the Poisson clutch size assumption but used age-specific mean clutch size values calculated from the data without any smoothing or model-fitting (Figure 2a–c). This slightly improved predictions of mean LRO, but predictions about standard deviation and skewness remained bad. The second way eliminated the Poisson assumption by using age-specific second and third moments of clutch sizes (as well as the first moment) calculated directly from individual-level observations (Figure 2a–c). We then used these measured moments in the reward matrices instead of Poisson-derived moments. This slightly improved skewness predictions but made predictions about standard deviation even worse.

The most pronounced theory-data discrepancy is substantial under-prediction of the among-individual variation in LRO: predictions for the mean are good but predictions for higher moments are terrible. This suggests the presence of among-individual differences not explained by age or by maternal age. To examine this hypothesis we fitted statistical models to the experimental data on individual life histories, which give the daily clutch size of each individual from birth to death. We modelled the clutch size distribution for each maternal age as a zero-inflated Poisson, where the amount of zero-inflation and the mean conditional on a positive value were both specified as spline functions of age, with individual random effects on the two intercepts. The model was fitted using the `ziP` regression family in the `gam` function in the **mgcv** library (Wood, 2017).

Both individual random effects were highly significant ($p < 0.001$ from `anova.gam` for all maternal ages) and they generated large among-individual differences in fitted mean clutch size (Figure 3). Estimated random effect standard deviations for maternal ages 7 and 9 were similar, and roughly double those for maternal ages 3 and 5. This aligns with maternal ages 7 and 9 having higher observed LRO standard deviations, and even larger differences between observed

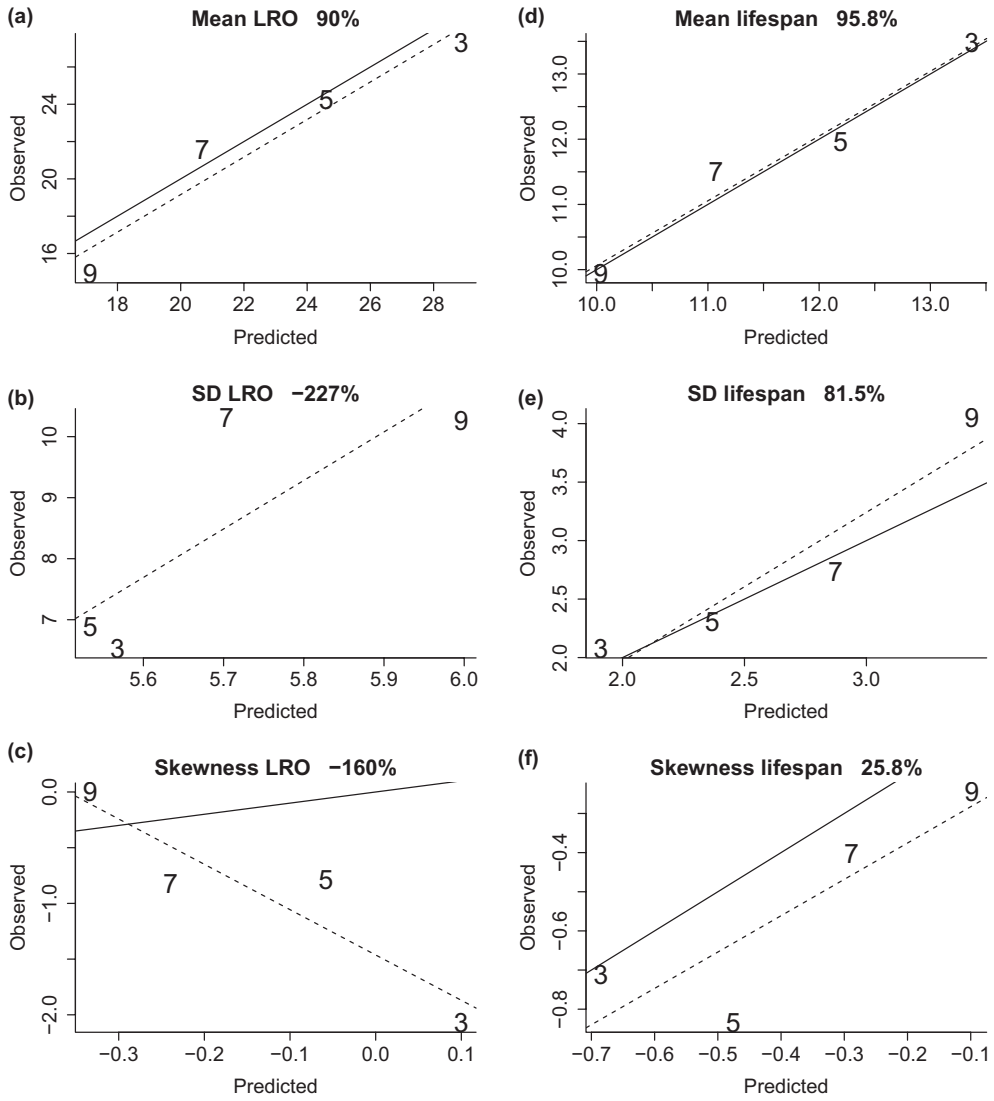


FIGURE 1 Predicted and observed trends across maternal age (3, 5, 7 or 9) in the mean, standard deviation and skewness of lifetime reproductive output (panels a,b,c) and lifespan (panels d,e,f) in *B. manjavacas* rotifers. Solid lines in each panel are the 1:1 line, and the dashed line is a fitted linear regression. Panel headings give R^2, the fraction of variance in observed values explained by the theoretical predictions (R^2 = mean of (observed-predicted)^2 divided by the variance of the observed moments). E. Empirical data analysis and theoretical predictions were generated by scripts `rotifer_moms/make_rotifer_maternalAge_matrices.R` and `rotifer_moms/calculate_rotifer_moments.R` and the figure generated by `rotifer_moms/compare_LRO_lifespan_rotifer.R`.

and predicted LRO standard deviations, than maternal ages 3 and 5 (Figure 2b,e). In contrast, in the *Daphnia magna* experiments described in the next section, individual random effects (while also present) were small enough to omit from the matrix model (Figure S-1).

Thus, comparing predicted with observed LRO variability has revealed the presence of demographic heterogeneity among individuals that the Hernández et al. (2020a) model does not include. Hernández et al. (2020a) were mostly concerned with effects of maternal age on mean demographic rates, so this does not call into question their conclusions. van Daalen et al. (2022) partitioned the among-individual variance in LRO into two components: variance due to differences in maternal age (estimated to be 26% of the total under laboratory conditions), and the remaining variance which they ascribed to individual stochasticity, variation 'arising from the random outcome of the same probabilities operating on identical individuals' (van Daalen et al. (2022), p. 603). Here, we found that the individual stochasticity implied by the fitted population model actually cannot explain much of the observed LRO variance within maternal age groups (Figure 2b,e). Our fitted clutch size models (Figure 3) imply that, instead, there is substantial persistent individual heterogeneity in fertility unrelated to age or maternal age. A more complete partition of LRO variance, with two different forms of individual heterogeneity, could be accomplished by constructing a model with a three- or four-way classification of individuals (age, maternal age, and one or two individual random effects modelling fertility variation unrelated to age or maternal age). Such a model could again be tested for whether it fully accounts for observed LRO variation. We do not attempt that here, having accomplished our present

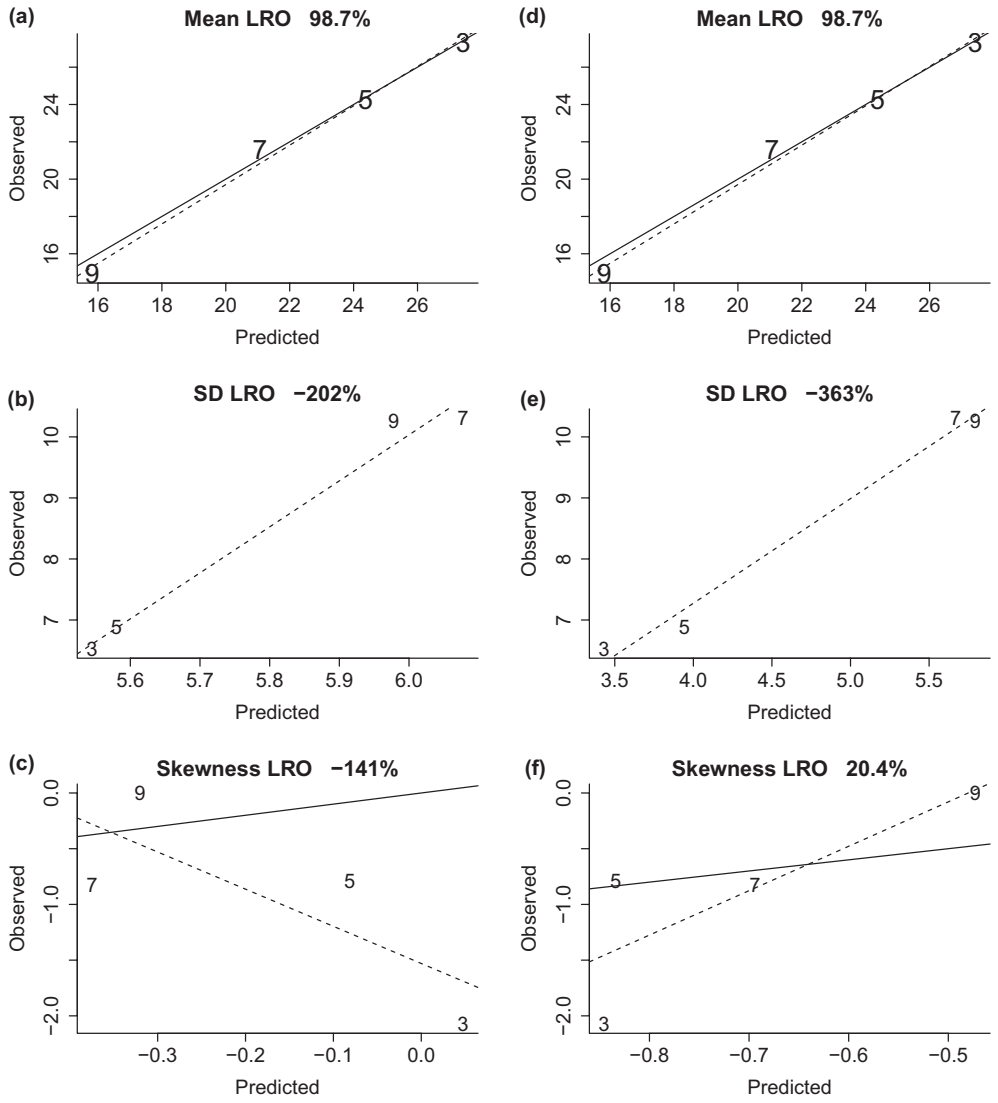


FIGURE 2 Predicted and observed trends across maternal age (3, 5, 7 or 9) in the mean, standard deviation and skewness of lifetime reproductive output in *B. manjavacas* rotifers. Left column (panels a,b,c): Using raw age-specific average clutch size estimates without any smoothing or modelling, while retaining the assumption of Poisson-distributed clutch sizes. Right column (panels d,e,f): Additionally eliminating the Poisson assumption, by using the raw age-specific second and third moments of clutch size without any smoothing or modelling. Solid lines in each panel are the 1:1 line, and the dashed line is a fitted linear regression. Panel headings give the fraction of variance in observed values explained by the theoretical predictions. Empirical data analysis and theoretical predictions were generated by scripts `rotifer_moms/make_rotifer_maternalAge_matrices.R` and `rotifer_moms/calculate_rotifer_moments.R`, and the figure generated by `rotifer_moms/compare_LRO_lifespan_rotifers.R`.

goal of illustrating how testing predictions from stochastic demography can detect and diagnose limitations of an existing model.

4 | WATER FLEAS, *Daphnia magna*

4.1 | Data

The data for this case study comes from a laboratory study of the freshwater zooplankton *Daphnia magna* (Vindenes et al., 2025b). Complete life histories were recorded for 10 individuals in each of 24 treatments: four clonal lines, raised at six constant temperatures running from 5 to 30°C. Individuals were raised in individual 100 mL jars at constant temperatures with food ad libitum and were checked daily from birth to death for survival, development (moulting), maturation to adulthood and reproduction.

4.2 | Model

A matrix model was constructed for each treatment, with individuals classified by phase (juvenile vs. adult), stage within phase (determined by the number of prior moults) and age within stage (Vindenes et al., 2025a). All individuals are characterized by survival

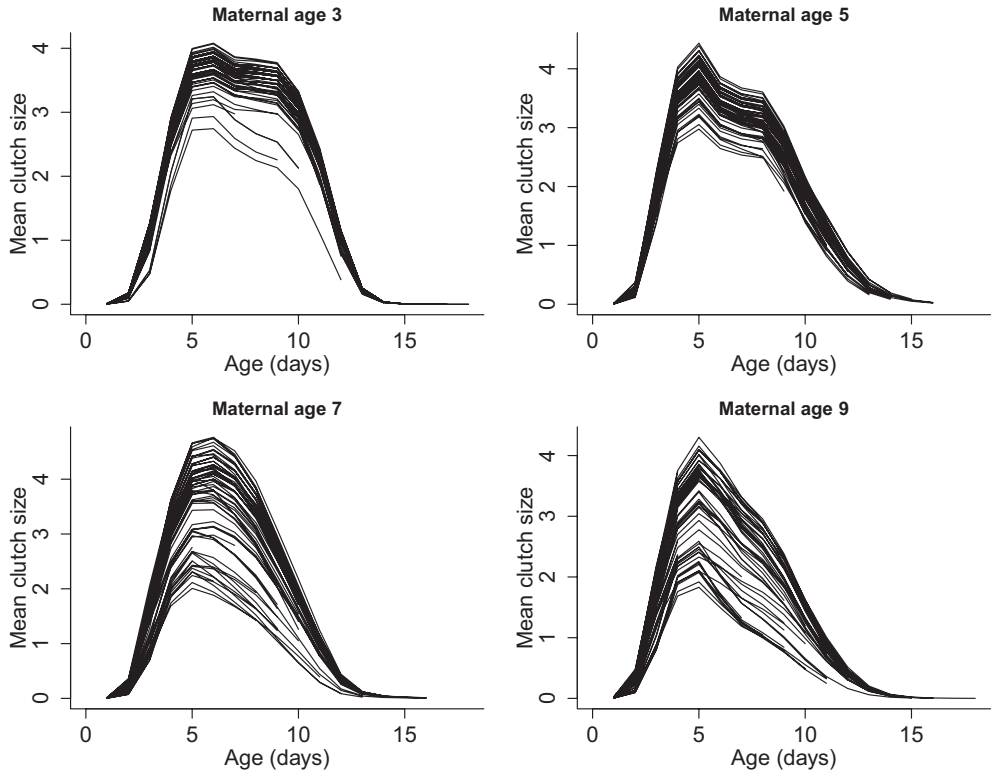


FIGURE 3 Predicted values of mean clutch size as a function of age for each individual *B. manjavacas* rotifers, for maternal ages 3, 5, 7 and 9. Each line represents one individual in the dataset, ending when that individual died. The fitted models were generalized additive models with zero-inflated Poisson response. The amount of zero-inflation and mean clutch size conditional on a positive value were both specified as a smoothing spline function of age with an individual random effect on the intercept (Wood, 2017). Modelling and plotting were performed by script `rotifer_moms/rotifer_random_effects_poisson.R`.

and moulting probabilities; juveniles are also characterized by a probability of maturation (producing a first clutch of eggs) and thus entering the first adult stage. These probabilities were all estimated by fitting statistical models to the combined data from all treatments accounting for censoring, with temperature, clone, phase, stage and time since last moult as potential covariates, with individual random effects where relevant, and separate moulting probability models for juveniles and adults. The matrix models were parameterized by setting individual random effects to zero, as predicting individual-level variation was not a goal of the study.

In good conditions, adults produce one parthenogenetic clutch per moult, and expected fecundity was assumed to depend on stage. We assumed that clutch sizes were Poisson-distributed with stage-specific mean.

4.3 | Results

The experiments were conducted to characterize trends with respect to temperature, and how those vary among clones and therefore had many treatments (24 temperature-clone combinations) with few individuals per treatment. Treatment-specific life-history metrics are therefore estimated imprecisely (if based on independent estimates in each treatment), so treatment-by-treatment model-data comparisons would have limited value. Our tests of the model therefore focus on predicted patterns of variation across treatments, the level at which the model was designed to make predictions.

For LRO, the model predicts the trends in mean extremely well, does less well with standard deviation and does poorly with skewness (Figure 4). The same pattern, only more pronounced, is found for the trends in lifespan across treatments. Based on mean squared error, the matrix model predictions of lifespan skewness for each treatment are substantially less accurate than simply using the mean of all observed values as the (constant) prediction for all treatments.

To try to identify why LRO skewness predictions are poor, we first compared the theoretical and empirical relationship between lifespan skewness and LRO skewness. Both theoretically and empirically, LRO and lifespan skewness are approximately equal (Figure 5): points lie close to the 1:1 line, and the linear regression line is very near the 1:1 line. Successful prediction of lifespan skewness would therefore entail successful prediction of LRO skewness. The question thus becomes: why is lifespan skewness poorly predicted?

Before seeking a mechanistic explanation, we first need to ask if sampling variability alone can account for poor overall prediction of lifespan skewness ($R^2 = -0.36$). We did so by simulating the sampling variability in lifespan skewness, as follows. First, we used the fitted matrix models to simulate 5000 random individual life histories for each treatment. Second, to obtain one draw from the sampling distribution

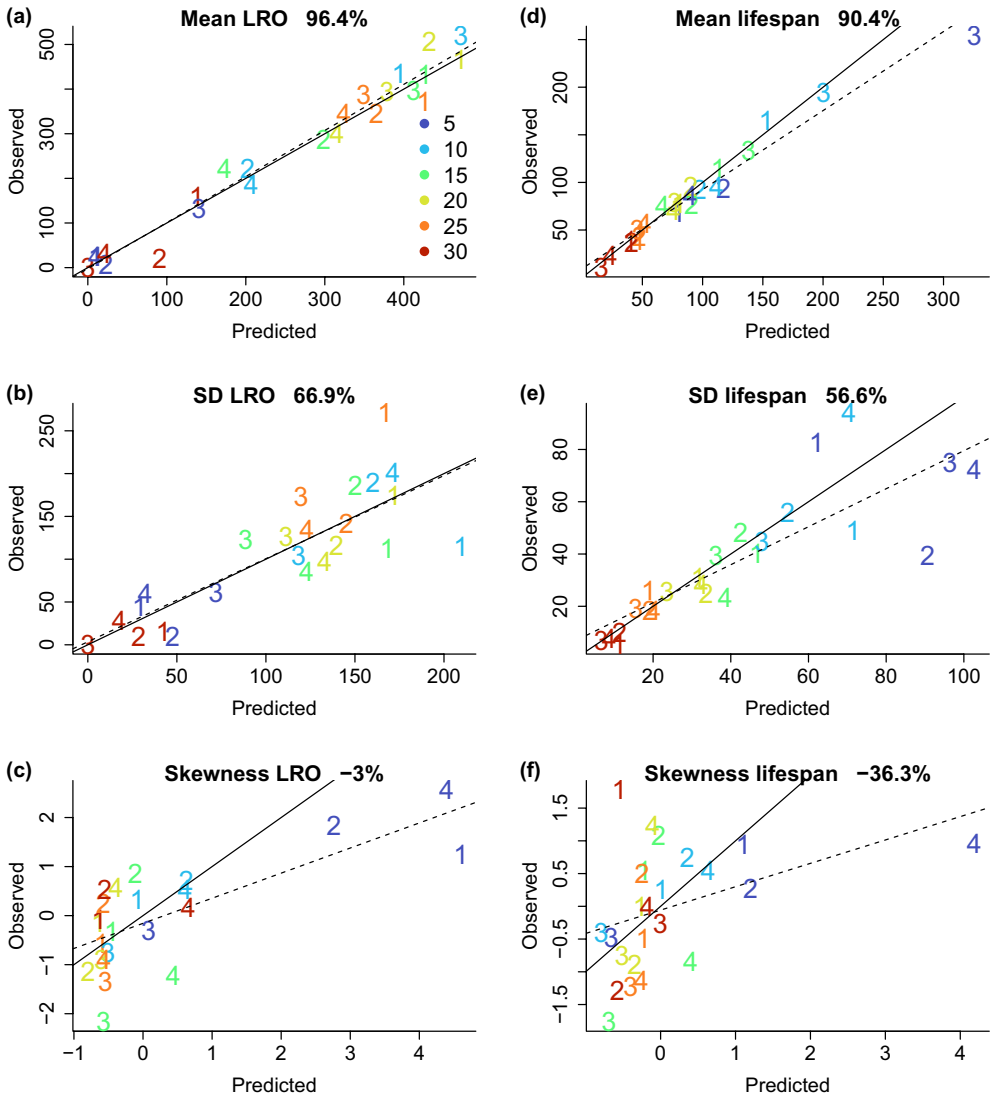


FIGURE 4 Predicted and observed trends across treatments in the mean, standard deviation and skewness of lifetime reproductive output (panels a,b,c) and lifespan (panels d,e,f) in *Daphnia*. Colours indicate rearing temperature (degrees C, shown in panel a) and numbers indicate clonal line (1 from Morocco, 2 from Sicily, 3 and 4, respectively, from near Uppsala and Lund, Sweden). The solid line in each panel is the 1:1 line, and the dashed line is a fitted linear regression. Panel headings give the fraction of variance in observed values explained by the model (defined as one minus the ratio between the mean of (observed-predicted)^2 and the variance of the observed values). Empirical data analysis, theoretical predictions and plots were generated by script `daphnia/compare_LRO_lifespan_daphnia.R`.

of lifespan skewness, we simulated the full experiment by drawing 10 individuals for each treatment at random from the 5000 that had been simulated, and computed lifespan skewness for each treatment. We repeated the second step 5000 times to approximate the sampling distribution of lifespan skewness. Across those 5000 simulations of the full experiment with sampling variability, only a fraction $p = 0.004$ of the simulations had a lower R^2 for prediction of observed lifespan skewness for each treatment. (R script: `sampling_lifespan_skewness_Daphnia.R`). Thus, more than sampling variability is involved.

However, if we omit the worst outlier in Figure 4f, clone 4 (SE-BY) at 5°C, then the results are very different (Figure 5c,d). The linear regression (dashed) line is then close to the 1:1 line indicating that the trend across treatments is predicted well, and the fraction of variance explained by the theoretical predictions is consistent with sampling variability (Monte Carlo $p = 0.85$ testing the null hypothesis that the differences between observed and predicted lifespan skewness are due to sampling variability).

So overall, 23 out of the 24 treatment-specific matrix models survive scrutiny with respect to stochastic variation in lifespan and LRO. The outlier, clone SE-BY at 5°C, is the one treatment for which there is no senescence in the fitted mortality rate, across the ages/stages prior to 100% mortality in the experiment (Vindenes et al., 2025a). This produces a small (in total probability) but very long tail in the predicted lifespan distribution: individuals that survive to be 'middle age' adults are predicted to have an impossibly long remaining lifespan of over 300 days on average. That tail has little effect on predicted mean lifespan for that treatment, slightly inflates the predicted standard deviation and greatly inflates the

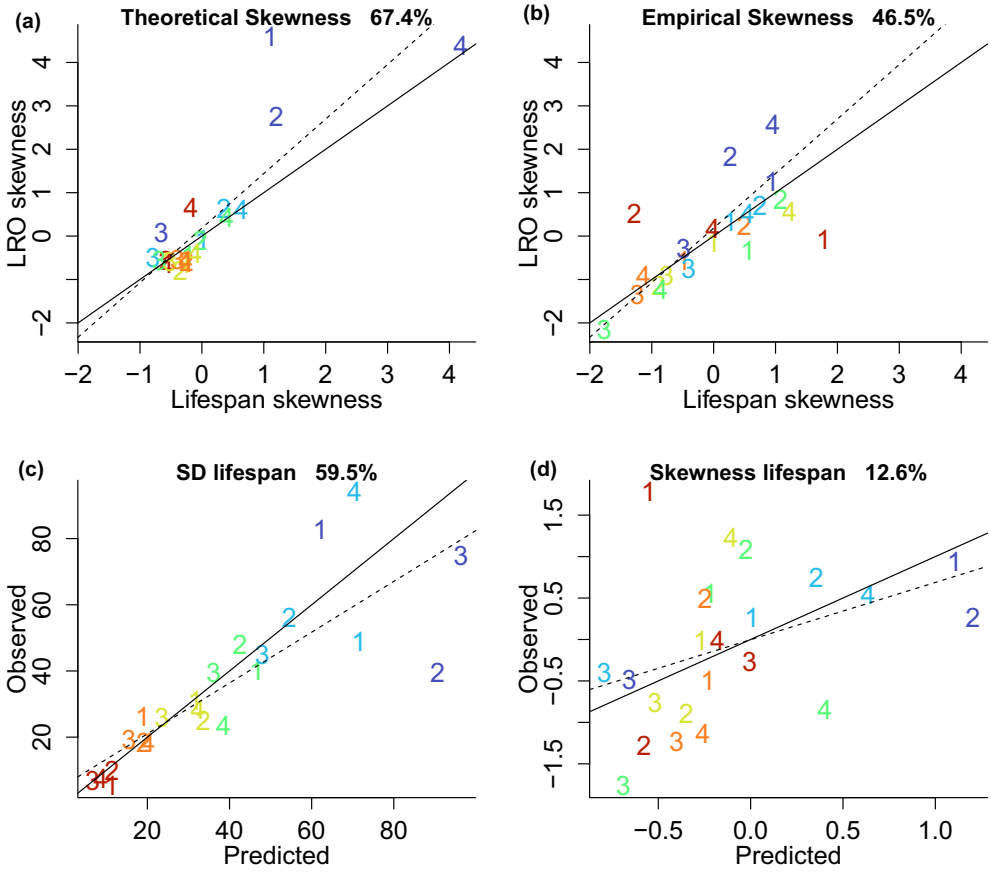


FIGURE 5 (a) Relationship between lifespan skewness and LRO skewness predicted from the matrix models for *Daphnia*. (b) Relationship between lifespan skewness and LRO skewness calculated directly from the individual lifetime data. The panel headings give the fraction of variance in LRO skewness explained by the prediction that LRO skewness equals lifespan skewness. (c, d) Predicted and observed trends across treatments in the standard deviation and skewness of lifespan, omitting clone 4 (SE-BY) raised at (5c). In all panels, colours indicate rearing temperature and numbers indicate clonal line, as in Figure 4. The solid line is the 1:1 line, and the dashed line is a fitted linear regression. Empirical data analysis, theoretical predictions and plots were generated by script `daphnia/compare_LRO_lifespan_daphnia.R`.

predicted skewness (Figure 4d–f). Similarly, across-treatment variation in the mean and standard deviation of number of moults as an adult are predicted very well (94.2% and 71.3% of cross-treatment variance) while the skewness is predicted very badly (−886% of variance explained; R script `compare_moults_daphnia.R`).

5 | GRASS POPULATIONS AND THE EFFECT OF ENDOPHYTES

5.1 | Data

The data for this case study come from experimental plots at Lilly-Dickey Woods in Indiana, USA (Miller et al., 2023). In 2007, the plots were planted with seven grass species (*Agrostis perennans*, *Elymus villosus*, *Elymus virginicus*, *Festuca subverticillata*, *Lolium arundinaceum*, *Poa alsodes* and *Poa sylvestris*), with half the plots containing endophyte-infected plants and half containing plants for which a heat treatment had been used to eliminate endophytes.

In addition to our general data curation procedures, we removed all individuals that were greenhouse-reared and transplanted, keeping only individuals who recruited in the field. We furthermore removed the four individuals that recruited in 2007, as there was some ambiguity about whether these were surviving original plants or recruits that emerged in the same place as their parent. We removed individuals whose size at the beginning of a time step was listed as NA. Finally, we removed any individuals that were recorded as dying in the same year they recruited but had non-zero lifetime reproduction, because new recruits are not actually able to reproduce.

5.2 | Model

A matrix projection model structured by individual size and stage (new recruit vs. older plants) was fit for each year from 2009 to 2020, for plants with and without endophytes (Fowler et al., 2024). Demographic rates were fitted as generalized linear mixed models in a Bayesian framework, with all rate functions depending on individual size (treated as a

continuous variable), endophyte presence or absence, a random plot effect (shared across species) and random year effects specific to each species/endophyte status combination. Size 1 represents non-reproductive seedlings. Larger sizes are defined by number of tillers, and size is capped in the model at the 97.5th percentile of observed sizes for each species. Annual reproductive output is defined as the number of inflorescences. For each year and endophyte status, we used transition rates that were the mean over 500 draws from the relevant posterior distribution of model parameters. Like Fowler et al. (2024) we used the average model across plots (i.e. plot effect was set to zero in all vital rate models).

5.3 | Results

Figures 6, 7 compare the moments of LRO and lifespan with and without endophytes, respectively. Agreement is reasonably good for most species, but (similar to *Daphnia*) some poor R^2 values are driven by a small number of notable outliers. We will discuss outliers in predictions for mean LRO and lifespan, as these are easiest to understand.

The model substantially overpredicts mean LRO for AGPE (*A. perennans*) in the presence of endophytes and underpredicts it in the absence of endophytes. Fowler et al. (2024) noted that endophytes reduce AGPE's mean flowering probability and recruitment, but it appears that the fitted model does not mirror this effect: the model predicts that the presence of endophytes increases mean LRO, which is the opposite of what is observed (Figure S-2). Visually, the data suggest that the endophytes may have opposite effects on transplanted and field-recruited plant fecundity, and a single endophyte effect was fitted for transplanted and field-recruited plants combined while the 'observed' effect only includes field-recruited plants.

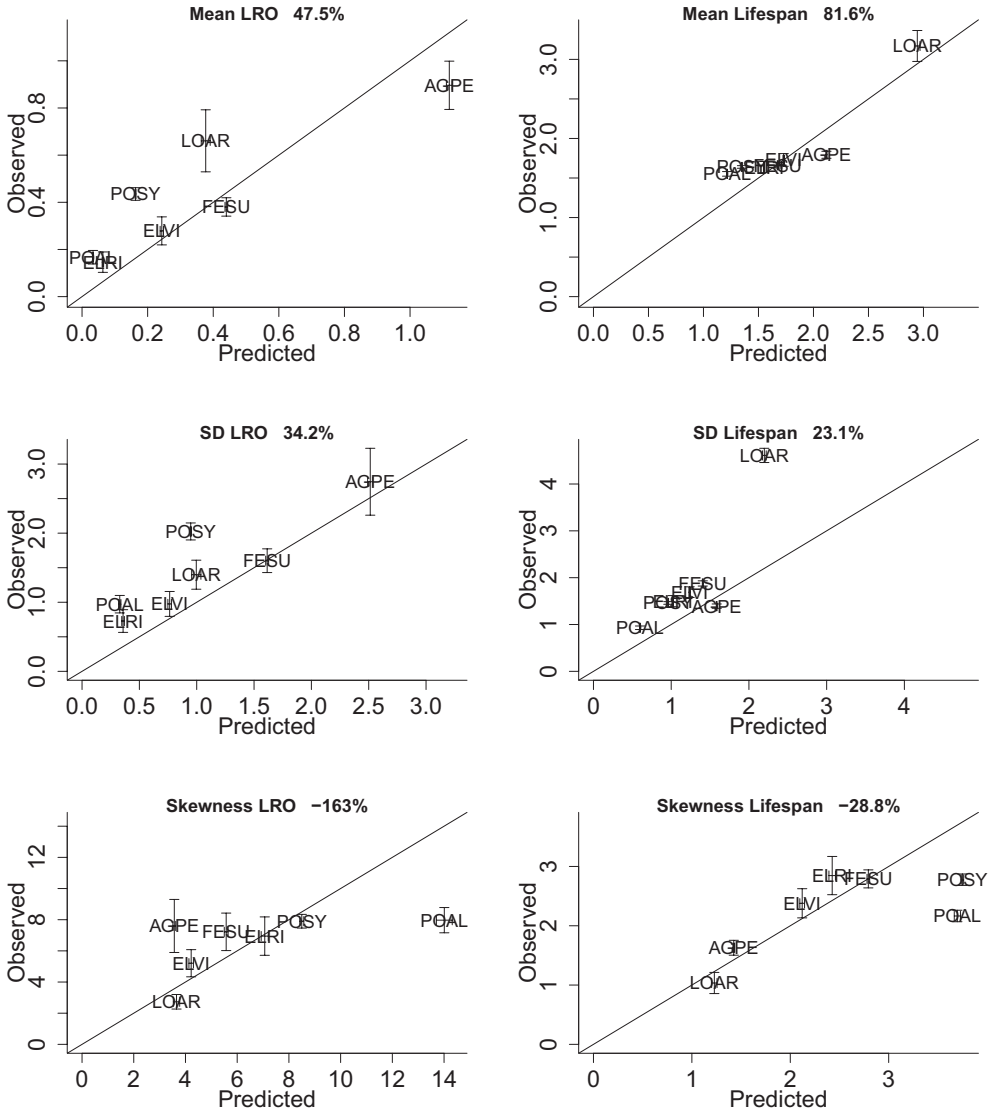


FIGURE 6 Grass–endophyte system with endophytes present: Theoretical values (from simulations) versus observed values for the mean, SD and skewness of LRO and lifespan for all grass species. Plots, data analysis and simulations performed by script `grassEndophyte/grassCompareMoments.R`.

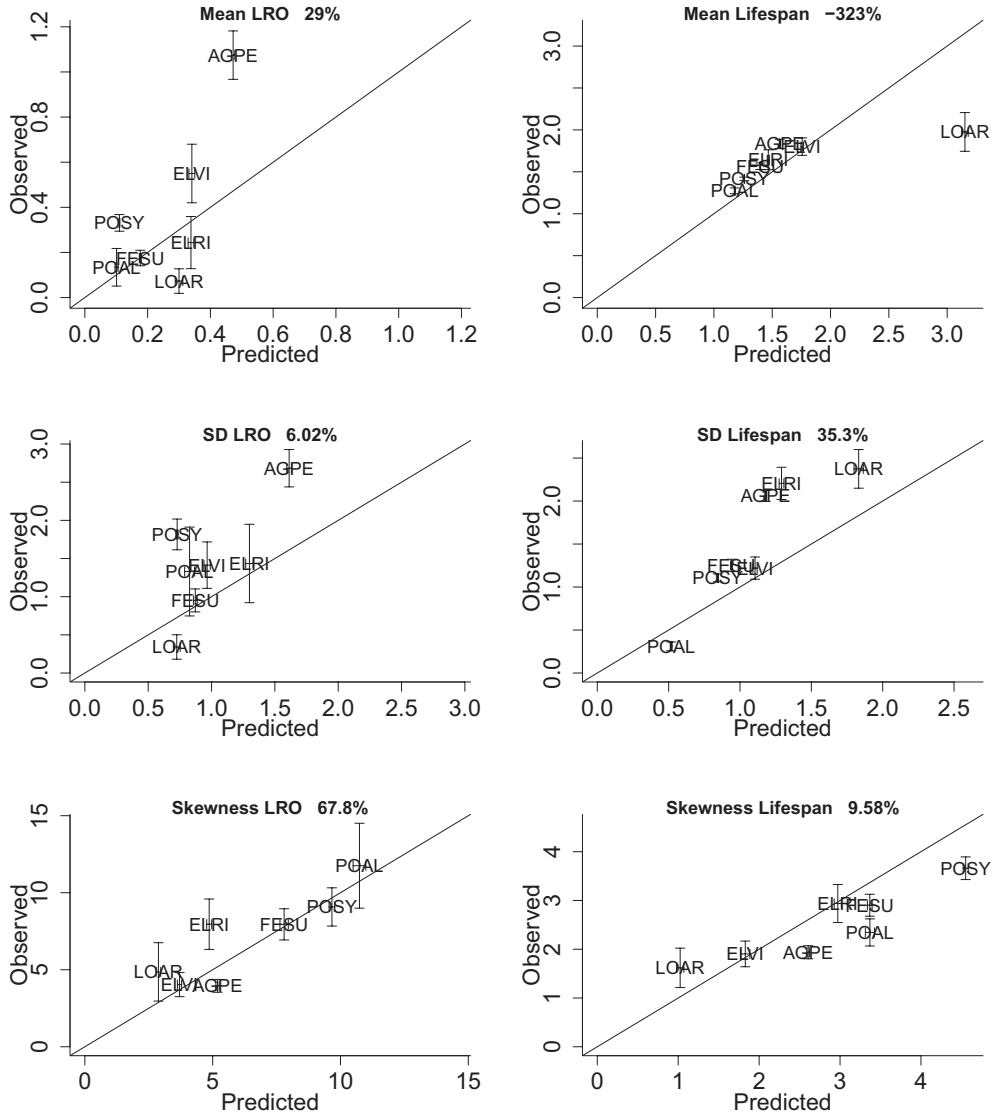


FIGURE 7 Grass–endophyte system with endophytes absent: Observed values versus theoretical values (from simulations) for the mean, SD and skewness of LRO and lifespan for all grass species. Plots, data analysis and simulations performed by script `grassEndophyte/grassCompareMoments.R`.

The model also strikingly overpredicts mean lifespan and LRO for LOAR (*L. arundinaceum*) in the absence of endophytes. The poor lifespan prediction is likely due to a lack of data: of the 41 LOAR individuals without endophytes, almost half recruited in 2009, and a little over half did not survive their first year. As a result, the fitted model is dominated by data on transplants, which are not included in the 'observed' life-history measures. In line with this conjecture, if we compute lifespan conditional on surviving the first year, the prediction becomes much better (Figure S-3). However, fixing mean lifespan in this way does not rescue the prediction for mean LRO in the absence of endophytes (Figure S-3). Rather, the problem appears to be that the model overestimates the speed of growth in the absence of endophytes (Figure S-7).

It is not clear why the model underpredicts mean LRO for LOAR in the presence of endophytes, given that it does a good job of predicting mean lifespan, mean fecundity as a function of size (Figure S-4) and mean size as a function of age (Figure S-6).

Although the species-specific errors are all small, it is striking that the SD of LRO and lifespan are underpredicted in nearly all species, both with and without endophytes present. This strongly suggests, once again, the presence of unmodelled heterogeneity (Vaupel et al., 1979). In this case, we know that plot effects were present in the real system but were omitted from the model used for all analyses both here and in the original publication (Fowler et al., 2024).

6 | SOAY SHEEP

6.1 | Data

The data for this case study comes from the long-running study of Soay sheep (*Ovis aries*) in the Village Bay catchment of the Island of Hirta in the Scottish St. Kilda archipelago (Clutton-Brock &

Pemberton, 2004). The model is fit to data from 1985 to 2009, so we include only individuals whose lives were contained by that interval. The model tracks only female individuals, so in our data analyses only female offspring are counted towards LRO. In determining the moments of lifespan, we used only individuals with a known year of death. After curation, we had observations of LRO for 2128 individuals and lifespan for 2200 individuals.

6.2 | Model

The model, from Childs et al. (2011), is an IPM in which individuals are classified by total body mass and by life-history stage, lamb (first year of life) or ewe (all later ages). The transition kernel takes the population from one summer to the following summer, but because lambs are born in the spring, transition rates integrate over a number of seasonal processes. For example, the probability that a lamb this summer produces a lamb that survives until next summer is the probability that the potential parent survives the winter times the probability that they reproduce (integrated over possible birth weights of their offspring), times the probability that the single or twin lambs survive until the following summer. More details can be found in the Supporting Information for Childs et al. (2011).

As the model has no time-varying parameters or year effects, exact theoretical predictions were calculated using the Markov Chain with Rewards methods (see the Section 2 above). Appendix S.2 describes in detail how we calculated the reward matrices for this model.

6.3 | Results

The model does an excellent job of predicting the mean and standard deviation of LRO and only slightly overpredicts skewness, even though it substantially overpredicts all moments of lifespan (Figure 8a,b). In particular, the predicted mean lifespan is 3.71 years, while the observed mean is 2.8 years.

The model gets LRO right despite overpredicting lifespan because its predictions for age at death are excellent out to age 16, the age of the oldest observed sheep (Figure 8c). Lifespan predictions are pulled too high by the long, thin tail of the calculated age at death distribution: 5% of the distribution lies in ages older than 16. If we truncate the calculated distribution at age 16 and renormalize it to 1, the new predicted mean is 2.74, quite close to the observed mean.

The prediction for LRO skewness is a little too high because the model predicts a roughly age-independent average yearly number of offspring for ewes over 2 years old, while in reality, fecundity drops off sharply for the oldest sheep (Figure 8e). Because very few individuals live long enough to see a fall off in fecundity and because the model slightly underpredicts yearly number of offspring for ages 3–6, the absence of reproductive senescence in the model does not much affect its predictions for the mean and standard deviation of LRO. However, it does stretch out the tail of the LRO distribution by enough to affect skewness because higher moments are more sensitive to distribution tails at large values.

The model is less good at predicting size and stage at death: it predicts too many individuals dying as lambs (and therefore too few dying as ewes), and of those that die as lambs, it predicts death at somewhat smaller sizes than observed (Figure 8d). Both discrepancies appear to be a consequence of a difference between observed and predicted size-dependent survival at the smallest sizes, possibly as a result of the fitted logistic regression model not being flexible enough to accurately predict survival at all sizes (Figure S-8).

7 | YELLOW-BELLIED MARMOTS

7.1 | Data

The data for this case study are from a study of yellow-bellied marmots (*Marmota flaviventer*) conducted for over 40 years near the Rocky Mountain Biological Laboratory in Gothic, Colorado USA (Armitage, 2014).

Because the matrix model was based on data from 1976 to 2016, we excluded individuals who did not live their entire lives within this interval. The matrix model furthermore only used individuals from 'the bench, boulder, cliff, gothic, marmot meadow, picnic, river and stonefield colonies', and so we did likewise. After this curation, 1181 individuals remained for whom we had complete life data. Also after curation, there were 1902 pups born, and these formed the basis of our simulated lives with those extending past 2016 excluded from model-data comparisons.

7.2 | Model

The model, from Paniw et al. (2020), is an IPM that classifies individuals by both body mass and life history stage (juvenile, yearling, non-reproductive adult and reproductive adult). In order to model imperfectly correlated interannual variation in survival, growth and recruitment, the demographic models include both a year-specific latent variable affecting all demographic rates, which is associated with winter severity, and separate year effects in survival, growth and recruitment functions. There are separate winter and summer functions for survival and growth.

7.3 | Results

The model does a mostly good job of predicting the moments of LRO and lifespan, but substantially underestimates the standard deviation of LRO (Figure 9). The model also modestly overpredicts the skewness in lifespan and somewhat overestimates the probability of a marmot dying in their first year (lifespan = 1): the observed fraction is 0.50, and the mean simulated fraction is 0.65, with a standard deviation of 0.02.

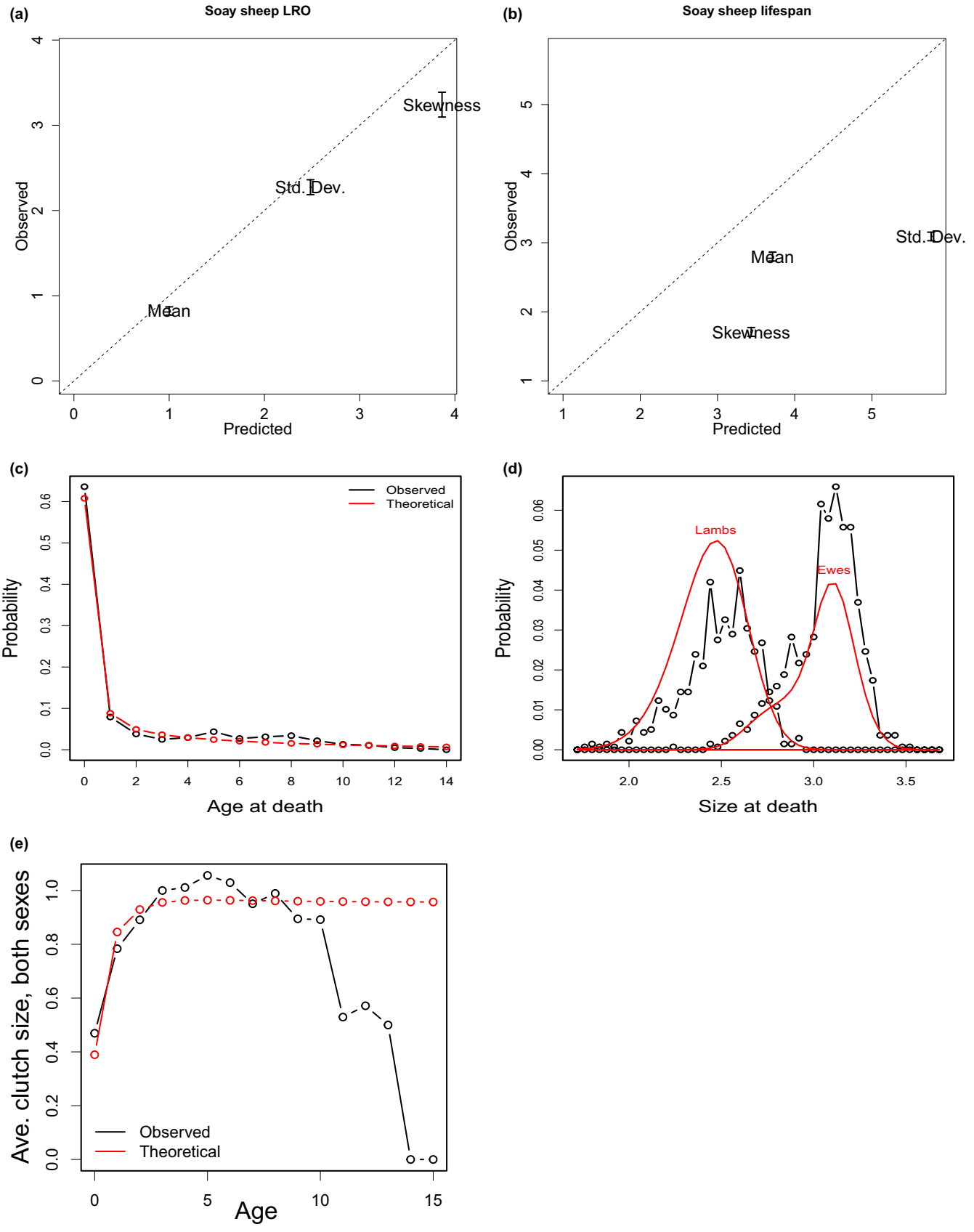


FIGURE 8 Soay sheep. (a, b) observed values versus theoretical values for the mean, SD and skewness of LRO and lifespan. Generated by `sheep/sheepCompareMoments.R`. (c, d) Observed and theoretical distributions of age and size at death. Generated by `sheep/compareAgeAndSizeAtDeath.R`. (e) Observed and calculated mean clutch size versus age. Figure generated by `sheep/compareClutchSizeProb.R`.

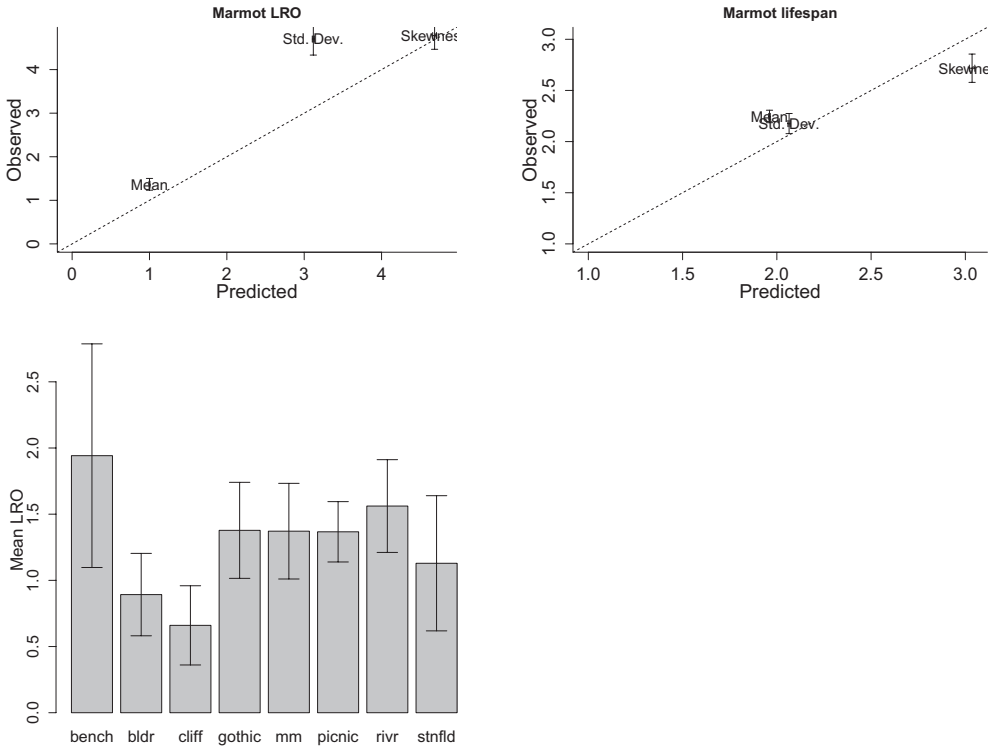


FIGURE 9 TOP: Yellow-bellied marmots: Simulated versus observed central moments of lifetime reproductive output. Plot and empirical analysis generated by `marmots/marmotLifespanLRO.R`, with simulated data generated by `marmots/simulateIndividuals3.R`. Bottom: The observed mean LRO differs noticeably between some colonies. Plot generated by `marmots/marmotLifespanLRO.R`.

The overprediction of lifespan skewness appears to be caused by the model omitting senescence. Maximum lifespan is not predicted badly: the maximum observed lifespan is 15, while the mean simulated maximum lifespan is 18.9 with a standard deviation of 3.6. However, when we limit the lifespan of simulated individuals to 15 years, the lifespan skewness prediction becomes accurate (Figure S-10). As in other case studies, higher moments are sensitive to small errors in distribution tails.

Our first hypothesis for the discrepancy in SD of LRO was that beyond the distinction between reproductive and non-reproductive adults each year, there might be zero-inflation in the clutch size distribution for reproductive adults (i.e. they bred but the litter failed), increasing LRO variance above model predictions. However, the observed and predicted proportions of individuals who never had offspring were very close: 86.1% of observed individuals never had offspring, while in the simulations, the proportion was 82.8% with a standard error of 0.2%.

Second, we noted that although it is visually less apparent, mean LRO is also underpredicted: simulated mean LRO is 73% of observed, while simulated SD LRO is 66% of observed—perhaps error in SD is a consequence of the error in mean? There is fairly good agreement between predicted and observed fecundity as a function of age over the range of likely marmot lifespans (Figure S-9), but the underprediction of first year survival is by itself enough to account for the low mean LRO (survival 0.35/0.5 = 0.7). However, if we artificially increase first-year survival in the model to the observed value, mean LRO is matched but the SD is still far off (77% of observed).

These results suggest that the error in first-year survival accounts for error in mean LRO, but there is some unmodelled source of heterogeneity in reproductive success. We found that mean LRO varies noticeably between some colonies (Figure 9). While this contributes to the overall SD of LRO, the among-colony variance is minute (≈ 0.15) relative to the discrepancy between the predicted and observed variances ($\approx 3.5^2$ vs. $\approx 4.5^2$). Other possibilities include spatial heterogeneity and effects of social interactions, both absent from the model but potentially contributing to variation in LRO. The fitted model also might underestimate variation in LRO because (to avoid over-stating the impact of environmental variation) several year effects were omitted despite having > 80% posterior support (Paniw et al., 2020, table S1.1).

8 | DISCUSSION

This paper has presented a series of 'stress tests' of published structured population models. In nearly all cases, the tests push the models beyond their original purposes. The point of the tests is to identify underlying potential problems in the model that might have significant impacts on the predictions that the model was built to make. Testing becomes particularly important when existing models are re-used by others. As the popularity of databases like COMADRE and COMPADRE shows, we are often eager to re-use models in ways unanticipated by their authors (e.g. transient dynamics Capdevila

et al., 2022; moments of lifespan and LRO Hernández et al., 2024; pace-of-life life-history strategy Salguero-Gómez et al., 2015; resilience to environmental fluctuations Cant et al., 2023).

The 'stress tests' we propose, involving among-individual variation in LRO and lifespan, can provide new insights because the empirical values are rarely used to parameterize the models and because the predictions result from the model as a whole. Thus, the tests may uncover issues that would not be found by evaluating demographic rate models one-by-one against the data that were used to parameterize them. Any uncovered issues may or may not matter, depending on the questions the model is intended to answer. When they do matter, the nature of the discrepancy can suggest what modifications are needed.

Our case studies show that our proposed stress tests can turn up model deficiencies that had gone unnoticed by experienced modellers (i.e. by us and other authors of the original papers). However, exactly what kinds of model mis-specifications can and cannot be discovered by testing individual-level predictions, and how large a problem must be before it is detected, remain to be determined.

What commonalities come through in our case studies? Our analyses often found that unmodelled individual heterogeneity was present, revealed by errors in predicted higher moments. Unmodelled among-individual variation in rotifer fertility created substantial errors in the predicted standard deviation and skewness of LRO, implying that prior estimates for the contribution of pure luck to LRO variation (van Daalen et al., 2022) were far from the mark. In marmots, the unmodelled between-colony variation can explain only a small fraction of the discrepancy between observed and predicted LRO variance, suggesting the presence of unmodelled among-individual heterogeneity.

Models that neglect senescence may still correctly predict higher moments in LRO, though they would not do so for lifespan. The Soay sheep model predicts an unrealistically long tail in the lifespan distribution, which throws off predictions for higher moments of lifespan. It also predicts roughly constant fecundity beyond the age of 2, whereas in reality, fecundity falls off sharply for the oldest sheep. However, the tail in the predicted lifespan distribution is thin enough that most simulated individuals die before the age at which fecundity predictions diverge, so the lack of senescence in the model does not affect predictions for the moments of LRO. The best-fitting model for one *Daphnia* clone at the lowest temperature had negative senescence, causing poor predictions of lifespan variance and skewness. Here, and probably very often, issues with distribution tails will have stronger and stronger effects on higher and higher moments. Similar to our findings, Caswell (2011) found closer agreement between predictions and observations for the variance of LRO than for the skewness, in both species for which that comparison was made (*Caenorhabditis elegans*, *Streblospio benedicti*).

Another general message is that choosing a single functional form for multiple species, locations or clones has costs as well as benefits. The benefits can be substantial. Parameter estimation can borrow strength across categories (which often makes biological sense, such as common temperature responses independent of size). We can more easily compare results across categories. When all data are used to fit one model, the larger sample size aids model selection. However, choosing a common model involves selecting among many potentially variable interactions, and researchers often make simplifying choices that can obscure features specific to a particular species, location, or clone. The one outlier *Daphnia* treatment ended up with a very long tail in lifespan because the chosen functional form was flexible enough to represent every clone but that one. It is likely that the grass–endophyte model predicts the wrong sign for the effect of endophytes on the LRO of *A. perennans* because the common model used for all species assumed that transplanted and field-recruited individuals had equal endophyte effects.

A final message is that a problem may not need fixing, depending on the goal of the study. It is hard to give general advice about this because the answer depends on the nature of the problem and on the goals of the study. As noted in the Introduction, for example, unmodelled within-population individual heterogeneity is likely to affect predictions about extinction risk, metapopulation viability, the rate of spatial spread and whether or not a disease introduced in a susceptible population becomes an epidemic or dies out, among other issues. In contrast, for our rotifer case study, it is unlikely that the model's original use (characterizing average maternal age effects and their consequences, Hernández et al., 2020a) was compromised by unmodelled heterogeneity within maternal age groups. Of course, accounting for individual heterogeneity (when the data permit) is generally important in fitting demographic rate models to maximize the accuracy of parameter estimates, even if including it in the population model is not necessary for the model's purposes.

Ideally, every newly identified problem that might compromise a study's conclusions would be resolved by building a better model. But given finite time, and no more data than you had yesterday, that may not be possible. Nonetheless, it may often be feasible to resolve the problem a different way, by showing that the study's conclusions are robust against the departures from reality suggested by the 'stress test' results. In studies where a general model is found to not fit all populations or treatments (such as our *Daphnia* and grass–endophyte studies), it is straightforward to see if any conclusions change or weaken if the ill-fitted instances are dropped, or if the worst predictions are replaced by values that are close to what actually happened, or close to the overall regression line relating observed and predicted values. In our marmot case study, we imposed senescence crudely to resolve one discrepancy, but this could also be done more carefully. If the effects are trivial, then unmodelled senescence is unlikely to be worth worrying about. Similarly, the other major issue identified in the marmot model is some unmodelled heterogeneity among marmots affecting LRO, most likely unmodelled variation in fecundity. In Section S.3, we add to the model enough heterogeneity in fecundity to make up for all of the 'missing' Var(LRO), and show that some of the conclusions of the original study are unaffected.

We also emphasize that the stress tests we suggest here are possible only by going back to the original data. As much as possible, researchers should endeavour to make the raw data and a wealth of

metadata available to enable the careful archiving and re-use of the models (Gascoigne et al., 2023).

We close by urging population researchers to follow the fate of individuals from the beginning of your study through the end, whenever this is possible. Even in a short-term study without complete lifetime observations, the fitted model can still be used to simulate individual lives over the years of the study, using the observed recruitment each year. Predicting individual lives and comparing with the data may reveal problems (and suggest solutions) that would not be found by examining the statistical fit of each demographic rate model one at a time. Does your model correctly predict the mean, standard deviation and skewness of reproductive output across individuals over the span of the study? Does it correctly predict the number of individuals who never bore offspring, and the number who had offspring in every year of the study? You can probably think of other synthetic measures that will be most informative in your system, for challenging the population model based on what it predicts about the lives of individuals one at a time.

AUTHOR CONTRIBUTIONS

Stephen Ellner and Robin Snyder conceived of the paper, performed the analyses and wrote the first draft. Daniel Blumstein, Dylan Childs, Joshua Fowler, Christina M. Hernández, Julien Martin, María Paniw and Yngvild Vindenes provided data and/or model code, and consulted on implementation and interpretation of case studies. All authors contributed to manuscript revisions.

ACKNOWLEDGEMENTS

We are grateful to Tom Miller for his generosity in answering our questions about the grass-endophyte model and data. We thank Jennifer Rudgers, Kenneth Whitney and the numerous researchers who contributed to data collection for the experimental study of grass–endophyte symbiosis. We thank all those involved in the long-term study of Soay sheep on St. Kilda, and are grateful to the National Trust for Scotland for granting permission to conduct research on the islands. We are grateful to the late Ken Armitage for starting the marmot study and the many previous and current marmoteers who collected these data over the past 64 years. This research was supported by National Science Foundation grants DEB-1933497 (SPE) and DEB-1933612 (RES), a National Science Foundation Postdoctoral Fellowship (2410282) (JCF), by MCIN/AEI/10.13039/501100011033 and 'ERDF A way of making Europe' grant PID2022-141004OA-I00 (PW), by the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Social Fund through the Ramón y Cajal Program grant RYC2021-033192-I (PW). Over the years, the marmot project has been supported by the National Science Foundation, the Rocky Mountain Biological Laboratory, the University of California Los Angeles, the National Sciences and Engineering Council of Canada, the University of Ottawa and the National Geographic Society.

CONFLICT OF INTEREST STATEMENT

No authors have conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are publicly available at https://doi.org/10.6084/m9.figshare.29453003 (Snyder et al., 2026).

ORCID

Stephen P. Ellner https://orcid.org/0000-0002-8351-9734
Robin E. Snyder https://orcid.org/0000-0002-6111-0284
Daniel T. Blumstein https://orcid.org/0000-0001-5793-9244
Christina M. Hernández https://orcid.org/0000-0002-7188-8217
Julien G. A. Martin https://orcid.org/0000-0001-7726-6809

REFERENCES

Antonovics, J. (2017). Transmission dynamics: critical questions and challenges. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *372*, 20160087.

Armitage, K. (2014). *Marmot biology: Sociality, individual fitness, and population dynamics*. Cambridge University Press.

Bock, M. J., Jarvis, G. C., Corey, E. L., Stone, E. E., & Gribble, K. E. (2019). Maternal age alters offspring lifespan, fitness, and lifespan extension under caloric restriction. *Scientific Reports*, *9*, 3138.

Broekman, M. J., Jongejans, E., & Tuljapurkar, S. (2020). Relative contributions of fixed and dynamic heterogeneity to variation in lifetime reproductive success in kestrels (*Falco tinnunculus*). *Population Ecology*, *62*, 408–424.

Cant, J., Capdevila, P., Beger, M., & Salguero-Gómez, R. (2023). Recent exposure to environmental stochasticity does not determine the demographic resilience of natural populations. *Ecology Letters*, *26*, 1186–1199.

Capdevila, P., Stott, I., Cant, J., Beger, M., Rowlands, G., Grace, M., & Salguero-Gómez, R. (2022). Life history mediates the trade-offs among different components of demographic resilience. *Ecology Letters*, *25*, 1566–1579.

Caswell, H. (2001). *Matrix population models: Construction, analysis and interpretation* (2nd ed.). Sinauer Associates.

Caswell, H. (2011). Beyond R0: Demographic models for variability of lifetime reproductive output. *PLoS One*, *6*, e20809.

Che-Castaldo, J., Jones, O. R., Kendall, B. E., Burns, J. H., Childs, D. Z., Ezard, T. H., Hernandez-Yanez, H., Hodgson, D. J., Jongejans, E., Knight, T., Merow, C., Ramula, S., Stott, I., Vindenes, Y., Yokomizo, H., & Salguero-Gómez, R. (2020). Comments to "persistent problems in the construction of matrix population models". *Ecological Modelling*, *416*, 108913.

Childs, D., Coulson, T., Pemberton, J., Clutton-Brock, T., & Rees, M. (2011). Predicting trait values and measuring selection in complex life histories: reproductive allocation decisions in Soay sheep. *Ecology Letters*, *14*, 985–992.

Clutton-Brock, T., & Pemberton, J. (2004). *Soay Sheep: Dynamics and selection in an Island population*. Cambridge University Press.

Cochran, M. E., & Ellner, S. (1992). Simple methods for calculating age-specific life history parameters from stage-structured models. *Ecological Monographs*, *62*, 345–364.

Ellner, S. P., Childs, D. Z., & Rees, M. (2016). *Data-driven modeling of structured populations: A practical guide to the integral projection model*. Springer.

Fowler, J. C., Ziegler, S., Whitney, K. D., Rudgers, J. A., & Miller, T. E. (2024). Microbial symbionts buffer hosts from the demographic costs of environmental stochasticity. *Ecology Letters*, *27*, e14438.

Fox, G., & Kendall, B. (2002). Demographic stochasticity and the variance reduction effect. *Ecology*, *83*, 1928–1934.

Gamelon, M., Morimoto, J., & White, H. (2025). Special feature: Intraspecific variation in ecology & evolution. *Journal of Animal Ecology*, *94*, 262–267.

Gascoigne, S. J. L., Rolph, S., Sankey, D., Nidadavolu, N., Stell Pičman, A. S., Hernández, C. M., Philpott, M. E. R., Salam, A., Bernard, C., Fenollosa, E., Lee, Y. J., McLean, J., Hetti Achchige Perera, S., Spacey, O. G., Kajin, M., Vinton, A. C., Archer, C. R., Burns, J. H., Buss, D. L., … Salguero-Gómez, R. (2023). A standard protocol to report discrete stage-structured demographic information. *Methods in Ecology and Evolution*, *14*, 2065–2083.

Gibert, J. (2016). The effect of phenotypic variation on metapopulation persistence. *Population Ecology*, *58*, 345–355.

Hernández, C. M., Ellner, S. P., Snyder, R. E., & Hooker, G. (2024). The natural history of luck: A synthesis study of structured population models. *Ecology Letters*, *27*, e14390.

Hernández, C. M., van Daalen, S. F., Caswell, H., Neubert, M. G., & Gribble, K. E. (2020a). A demographic and evolutionary analysis of maternal effect senescence. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 16431–16437.

Hernández, C. M., van Daalen, S. F., Caswell, H., Neubert, M. G., & Gribble, K. E. (2020b). Supplemental data to accompany Hernández et al. 2020 PNAS. https://github.com/chrissy3815/rotifer-moms

Jones, O. R., Barks, P., Stott, I., James, T. D., Levin, S., Petry, W. K., Capdevila, P., Che-Castaldo, J., Jackson, J., Römer, G., Schuette, C., Thomas, C. C., & Salguero-Gómez, R. (2022). 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. *Methods in Ecology and Evolution*, *13*, 770–781.

Kendall, B., & Fox, G. (2002). Variation among individuals and reduced demographic stochasticity. *Conservation Biology*, *16*, 109–116.

Kendall, B. E., Fujiwara, M., Diaz-Lopez, J., Schneider, S., Voigt, J., & Wiesner, S. (2019). Persistent problems in the construction of matrix population models. *Ecological Modelling*, *406*, 33–43.

Levin, S., Compagnoni, A., Childs, D., Evers, S., Potter, T., Salguero-Gomez, R., & Knight, T. (2022). Rpadrino: an R package to access and use PADRINO, an open access database of integral projection models. *Methods in Ecology and Evolution*, *13*, 1923–1929.

Miller, T., Fowler, J., Ziegler, S., Rudgers, J., Whitney, K., & Sheehan, M. (2023). Demographic data from long-term symbiont removal experiments with grasses and Epichloë fungal endophytes ver 2. Environmental Data Initiative https://doi.org/10.6073/pasta/ea7db07a578fb030a173f37f76596b62

Miller, T. E. X., & Ellner, S. P. (2025). My, how you've grown: A practical guide to modeling size transitions for integral projection model (IPM) applications. *Ecology*, *106*, e70088.

Paniw, M., Childs, D. Z., Armitage, K. B., Blumstein, D. T., Martin, J. G., Oli, M. K., & Ozgul, A. (2020). Assessing seasonal demographic covariation to understand environmental-change impacts on a hibernating mammal. *Ecology Letters*, *23*, 588–597.

Salguero-Gomez, R., Jones, O. R., Archer, C. R., Bein, C., Buhr, H., Farack, C., Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Ruoff, T., Sommer, V., Wille, J., Voigt, J., Zeh, S., Vieregg, D., Buckley, Y. M., Che-Castaldo, J., … Vaupel, J. W. (2016). COMADRE: a global data base of animal demography. *Journal of Animal Ecology*, *85*, 371–384.

Salguero-Gomez, R., Jones, O. R., Archer, C. R., Buckley, Y. M., Che-Castaldo, J., Caswell, H., Hodgson, D., Scheuerlein, A., Conde, D. A., Brinks, E., de Buhr, H., Farack, C., Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Runge, J., Ruoff, T., … Vaupel, J. W. (2015). The COMPADRE Plant Matrix Database: an open online repository for plant demography. *Journal of Ecology*, *103*, 202–218.

Salguero-Gómez, R., Jones, O. R., Jongejans, E., Blomberg, S. P., Hodgson, D. J., Mbeau-Ache, C., Zuidema, P. A., de Kroon, H., & Buckley, Y. M. (2015). Fast–slow continuum and reproductive strategies structure plant life-history variation worldwide. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 230–235.

Snyder, R. E., & Ellner, S. P. (2016). We happy few: using structured population models to identify the decisive events in the lives of exceptional individuals. *American Naturalist*, *188*, E28–E45.

Snyder, R. E., & Ellner, S. P. (2018). Pluck or luck: Does trait variation or chance drive variation in lifetime reproductive success? *American Naturalist*, *191*, E90–E107.

Snyder, R., Ellner, S., Blumstein, D. T., Childs, D. Z., Fowler, J. C., Hernandez, C. M., Martin, J. G. A., Paniw, M., & Vindenes, Y. (2026). Code and data for "*Challenging and diagnosing structured population models by testing predictions from stochastic demography*," Ellner et al., 2025 (Version 1) [Dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.29453003.V1

Snyder, R. E., Ellner, S. P., & Hooker, G. (2021). Time and chance: Using age partitioning to understand how luck drives variation in reproductive success. *The American Naturalist*, *197*, E110–E128.

Sorel, M. H., Jorgensen, J. C., Zabel, R. W., Scheuerell, M. D., Murdoch, A. R., Kamphaus, C. M., & Converse, S. J. (2024). Incorporating life history diversity in an integrated population model to inform viability analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, *81*, 535–548.

Steiner, U., Tuljapurkar, S., & Roach, D. (2021). Quantifying the effect of genetic, environmental and individual demographic stochastic variability for population dynamics in *Plantago lanceolata*. *Scientific Reports*, *11*, 23174.

Steiner, U. K., & Tuljapurkar, S. (2012). Neutral theory for life histories and individual variability in fitness components. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 4684–4689.

Steiner, U. K., Tuljapurkar, S., & Orzack, S. H. (2010). Dynamic heterogeneity and life history variability in the kittiwake. *Journal of Animal Ecology*, *79*, 436–444.

Stover, J., Kendall, B., & Nisbet, R. (2014). Consequences of dispersal heterogeneity for population spread and persistence. *Bulletin of Mathematical Biology*, *76*, 2681–2710.

Tuljapurkar, S., Steiner, U. K., & Orzack, S. H. (2009). Dynamic heterogeneity in life histories. *Ecology Letters*, *12*, 93–106.

Tuljapurkar, S., Zuo, W., Coulson, T., Horvitz, C., & Gaillard, J. M. (2020). Skewed distributions of lifetime reproductive success: beyond mean and variance. *Ecology Letters*, *23*, 748–756.

Tuljapurkar, S., Zuo, W., Coulson, T., Horvitz, C., & Gaillard, J. M. (2021). Distributions of LRS in varying environments. *Ecology Letters*, *24*, 1328–1340.

van Daalen, S. F., & Caswell, H. (2017). Lifetime reproductive output: Individual stochasticity, variance, and sensitivity analysis. *Theoretical Ecology*, *10*, 355–374.

van Daalen, S. F., Hernández, C. M., Caswell, H., Neubert, M. G., & Gribble, K. E. (2022). The contributions of maternal age heterogeneity to variance in lifetime reproductive output. *The American Naturalist*, *199*, 603–616.

Vaupel, J., Manton, K., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*, 439–454.

Vindenes, Y., Broch, C., Andersen, T., Hessen, D., & Ergon, T. (2025a). Understanding the role of ageing in the thermal responses of life history and fitness in *Daphnia magna*. *Proceedings of the Royal Society B-Biological Sciences*, *292*, 20250430.

Vindenes, Y., Broch, C., Andersen, T., Hessen, D. O., & Ergon, T. (2025b). Supplementary materials for the article 'Understanding the role of ageing in the thermal responses of life history and fitness in daphnia magna'. *Figshare*. Online resource. https://doi.org/10.6084/m9.figshare.26197589.v1

Vindenes, Y., & Langangen, Ø. (2015). Individual heterogeneity in life histories and eco-evolutionary dynamics. *Ecology Letters*, *18*, 417–432.

Vindenes, Y., Sæther, B. E., & Engen, S. (2012). Effects of demographic structure on key properties of stochastic density-independent population dynamics. *Theoretical Population Biology*, *82*, 253–263.

Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC.

SUPPORTING INFORMATION

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

Figure S-1. Predicted values of mean clutch size as a function of age for each *Daphnia magna* individual, for clones one through four (left to right) and all experimental temperatures. Fitted models were linear mixed models with Gaussian response, fitted using `lme` from the **nlme** package in R.

FigureS-2. Observed versus simulated change in mean LRO and lifespan for the grasses: with endophytes minus without. Red species names indicate that the predicted change has the wrong sign. Generated by `compareChangeInMeanLROAndLifespanMultSpecies2.R`.

Figure S-3. Comparing predicted and observed LOAR mean LRO and lifespan, with and without conditioning on surviving the first year. Conditioning on surviving the first year dramatically improves the prediction for mean lifespan in the absence of endophytes. Generated by `compareCondAndUncondMeans.R`.

Figure S-4. Observed grass fecundity versus expected fecundity in the presence of endophytes. Point colours represent size, with reds and oranges representing smaller sizes and blues and purples representing larger sizes. Point size represents number of observations. The solid line is a 1-to-1 line, while the dashed line is a linear regression, with 95% confidence intervals indicated in blue. Generated by `checkFecundityBinnedTomAllSpecies.R`.

Figure S-5. Observed grass fecundity versus expected fecundity in the absence of endophytes. Point colours represent size, with reds and oranges representing smaller sizes and blues and purples representing larger sizes. Point size represents number of observations. The solid line is a 1-to-1 line, while the dashed line is a linear regression, with 95% confidence intervals indicated in blue. Generated by `checkFecundityBinnedTomAllSpecies.R`.

Figure S-6. Observed and simulated mean size versus age with endophytes present. Red lines represent model predictions while black lines show observed values. Generated by `compareSizeVsAge.R`.

Figure S-7. Observed and simulated mean size versus age with endophytes absent. Red lines represent model predictions while black lines show observed values. Generated by `compareSizeVsAge.R`.

Figure S-8. Predicted and observed lamb survival as a function of size. Circle size indicates the amount of data available at that size. The model (red line) underestimates survival at smaller sizes. Generated by `compareAgeAndSizeAtDeath.R`.

Figure S-9. Marmot mean number of offspring versus age. Error bars represent one standard error. The error bars for the simulation results are based on simulating 100 replicated populations. The error bars for the empirical results are the result of bootstrapping. Plot generated by `marmotLifespanLRO.R`.

Figure S-10. Yellow-bellied marmots: simulated versus observed central moments of lifetime reproductive output and lifespan when no simulated individual was allowed to live past 15 years old. Adding this crude form of senescence fixes the prediction for lifespan skewness but makes little change to the moments of LRO. Plot and empirical analysis generated by `marmots/marmotLifespanLRO.R`, with simulated data generated by `marmots/simulateIndividuals3.R`.

Figure S-11. Yellow-bellied marmots: empirically observed age-dependent survival probability $p(x)$ for all years and colonies used in constructing the model and in our model-data comparisons. Values are not extended past age 13 because of very small sample size (two surviving marmots). Survival of young marmots increases with age because of growth. The decreasing survival at older ages is best explained as senescence. Figure generated by `marmots/marmotLifespanLRO.R`.

Figure S-12. Replicating the results in the left half of Paniw et al. (2020) fig. 4, showing the 'probability of quasi-extinction (i.e. <4 non-juveniles in the population) of yellow-bellied marmots under different scenarios of environmental change. The scenarios consisted of projecting population dynamics for 50 years fixing a different mean (μ) and standard deviation (σ) of environmental quality Q in all demographic processes'. Points are the mean estimate and lines extend from the 5th to 95th percentile of quasi-extinction estimates across parameter draws from the posterior distribution. Simulations were generated by R script `marmot_PrExt_sims_quality.R` based on `build_marmot_IPM.R` from the Paniw et al. (2020) code archive, and the plotting was done with script plot `replication.R`.

How to cite this article: Ellner, S. P., Snyder, R. E., Blumstein, D. T., Childs, D. Z., Fowler, J. C., Hernández, C. M., Martin, J. G. A., Paniw, M., & Vindenes, Y. (2026). Challenging and diagnosing structured population models by testing predictions from stochastic demography. *Methods in Ecology and Evolution*, *00*, 1–19. https://doi.org/10.1111/2041-210x.70337