## Model checking for the statistical models

To evaluate the fit of our statistical models to the data, we performed model checks with posterior predictive checks and graphical diagnostics (Conn et al. 2018; Hobbs 2015). We used the posterior distribution to simulate replicate datasets based on the parameters of our model. We compared samples from the simulated datasets to the observed datasets using by calculating Bayesian p-values for test statistics calculated from the observed and simulated data, and with graphical, visual checks. In a sense, these checks form a similar part of the workflow as examining plots of residuals versus fitted values when fitting a linear regression.

The general idea behind posterior predictive checks is to compare properties of the observed data to properties of data simulated from the joint posterior of the model fit to the data (Conn et al. 2018; Hobbs 2015). Posterior predictive checks require the choice of a discrepancy function to quantify the lack of fit. To quantify the similarity of the distribution of the test statistic from simulated, we calculate a Bayesian p-value. Bayesian p-values measure the frequency with which the test statistics calculated from the simulated datasets exceeds the test statistic calculated from the observed dataset. Extreme Bayesian p-values (e.g. less than 0.05 or greater than 0.95) indicate lack of model fit while values around 0.5 indicate reasonable model fit. Graphical methods complement posterior predictive checks, particularly when it is not clear which discrepancy function is most appropriate for quantitatively assessing fit (Conn et al. 2018). Graphical presentations of the simulated data and the observed data can be used to identify mismatches that indicate issues with model fit.

For example, for the model of seedling survival to fruiting, we simulated replicate binomial trials for each plot in each year at each population. The simulations draw the population- and year-specific probability of survival from the model's joint posterior, and use number of seedlings in a plot as the number of trials to simulate numbers of fruiting plants. We thus used the fitted model to generate a series of simulations to which we compare the observed data. From this series of replicate datasets, we calculate the mean of each simulated dataset and compared this

to the mean of the observed dataset. The Bayesian p-value is then the frequency with which the means of the simulated datasets exceeds the mean of the observed dataset. A similar logic applies to calculating any relevant test statistic. To complement this assessment of model fit, we also plot the simulated numbers of fruiting plants. We can thus compare the simulated data, also known as the posterior predictive distribution, with the distribution of the observed data. For all datasets, we calculated Bayesian p-values for the mean value of observations for each combination of population and year. We calculated $\chi^2$ as the discrepancy function (Conn et al. 2018).

## *Seedling survival to fruiting*

For the model of seedling survival to fruiting, we used the posterior distribution of parameters from the model to simulate observations for each plot, in each year, at each population. The simulations used the population- and year-specific probability of survival from the joint posterior of the model, and took the observed number of seedlings in a plot as the number of trials with which to simulate numbers of fruiting plants. We calculated the mean and $\chi^2$ values as test statistics (Fig. 1). We summarized the Bayesian p-values from these test statistics to represent the distribution of p-values across populations and years.

We generally observed a good fit of the model to the mean number of fruiting plants in permanent plots, as indicated by the narrow distribution of the test statistics around 0.5 (Fig. 1A). In our study, we used the population-level estimate of seedling survival to fruiting to calculate per-capita reproductive success. A good fit between our model's mean and the observed mean suggests that we can use the model to make inferences about population-level patterns of seedling survival to fruiting in these populations. However, Bayesian p-values based on the $\chi^2$ test statistic suggested a poorer fit of the model to the data (Fig. 1B). We observed evidence of a poor fit of the model in at least some populations in all years across the study. Evidence for lack of fit based on the $\chi^2$ test statistic suggests that the model failed to represent the distribution of seedling survival to fruiting across all plots. Put another way, the models fit the mean well but

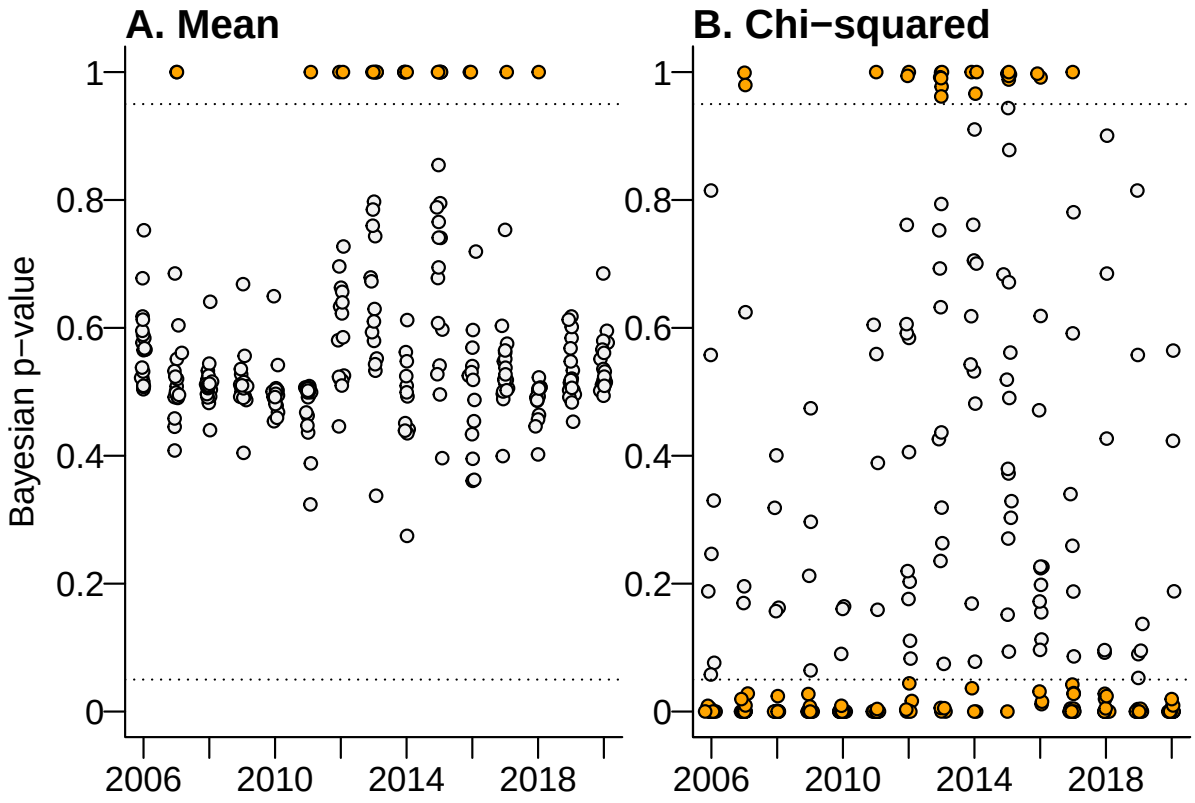


Figure 1: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of seedling survival to fruiting in all populations from 2006-2020. Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds at which the test statistic indicates lack of model fit. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

poorly represented the variance.

To examine cases with poor fit in more detail, we plotted samples from the posterior predictive distribution against the observations. Specifically, we plotted the posterior predictive distribution for all populations in each year. Here, we discuss 2007 as an example of the patterns that we observed across years (Fig. 2). In 2007, we observed models that (1) had reasonable p-values for both the mean and $\chi^2$, which we interpreted as fitting the observations well (gray lines). These models did a reasonable job of fitting both the mean and distribution across plots (e.g. LCE, DLW) We also observed models that (2) showed a poor fit for both the mean and

$\chi^2$ test statistic (orange lines). We primarily observed this lack of fit when all of the seedlings died at a population in a given year; no seedlings survived to become fruiting plants (dotted lines showing the observed number of fruiting plants at 0 for S22 and GCN). Because our model partially pooled year-level means to the population-level mean, estimates of seedling survival in these years was greater than observed (compare orange distribution and dotted lines). Finally, we observed models that (3) had reasonable p-values for the mean but showed a lack of fit based on the $\chi^2$ test statistic (purple lines).

To examine the patterns underlying these different model fits, we plotted the simulated data at the plot level for a handful of populations in 2007 (Fig. 3). These plots reveal why the models may predict average seedling survival to fruiting correctly but fail to provide a good fit to the overall distribution of the data. For example, consider LCW. The posterior predictive distribution for LCW (Fig. 2) showed what looks to be a reasonable fit to the data, but the p-values for the $\chi^2$ test statistic suggest a lack of fit. Examining the per plot distribution of simulated fruiting plant counts (Fig. 3) shows that the model sometimes incorrectly simulated high numbers of fruiting plants for some plots with few plants (observation number 17 - distribution of simulated values is around 80 but the observed value is close to 0), and sometimes incorrectly simulated low numbers of fruiting plants for some plots with many plants (observation number 4 - distribution of simulated values is around 20 but the observed value is close to 80). In addition, the distribution of simulated values is not always centered on the plot-level true observation.

One of the reasons that the plot-level simulated values do not provide a good fit to the plot-level true observations is that our hierarchical model partially pools estimates towards the population-level mean. For LCW, this means that more extreme estimates (i.e., observations of lower or higher survival than average) are pooled towards the mean. For example, this can be seen by examining the posterior predictive distribution at the plot level for LCW across all years in the dataset (Fig. 4). There are several years with observations of 0 seedlings surviving to fruiting (e.g. 2013, 2015, 2017), and these in turn influence the overall population-level mean, which then in turn influences the estimate in other years as well via partial pooling.

In this study, we use the estimates from these models to make inferences about population- and year-level seedling survival to fruiting. The models accurately estimate the mean across populations and years (Fig. 1A). We describe a lack of fit for the full distribution of seedling survival to fruiting (Fig. 1B), despite the distribution of seedling survival to fruiting generally matching the observed distribution (Fig. 2). We attribute this mismatch to the models not fitting the plot-level observations across plots (Fig. 3). However, we suggest that this mismatch arises in part as a consequence of partial pooling (Fig. 4), which is a desired property of the hierarchical models we use.

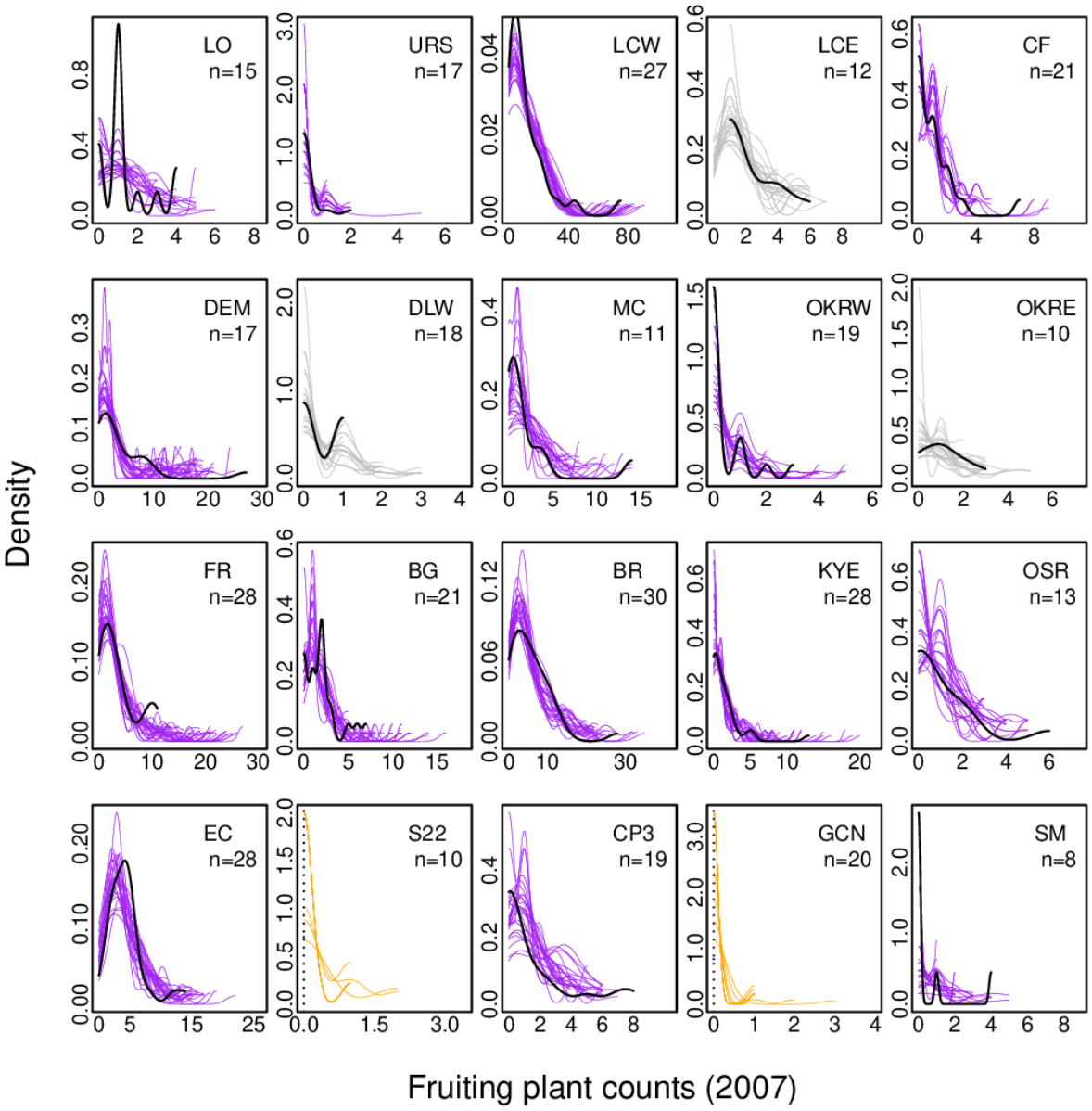


Figure 2: Distributions of counts of fruiting plants for 2007, based on the model for seedling survival to fruiting. Each orange or gray line is the distribution of counts from one *in silico* replicate of seedlings surviving in one year. We simulated 5000 datasets using the observed number of seedlings in permanent plots, and the estimated population-level probability of seedling survival to fruiting. We then randomly sampled 25 datasets for plotting. The plots with orange lines correspond to models that show lack of fit based on both the mean and the $\chi^2$ test statistic; the plots with the purple lines correspond to models that show lack of fit based on the $\chi^2$ test statistic but not the mean; the plots with gray lines correspond to models that show good fit based on both the mean and the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of fruiting plants per permanent plot from the field observations.

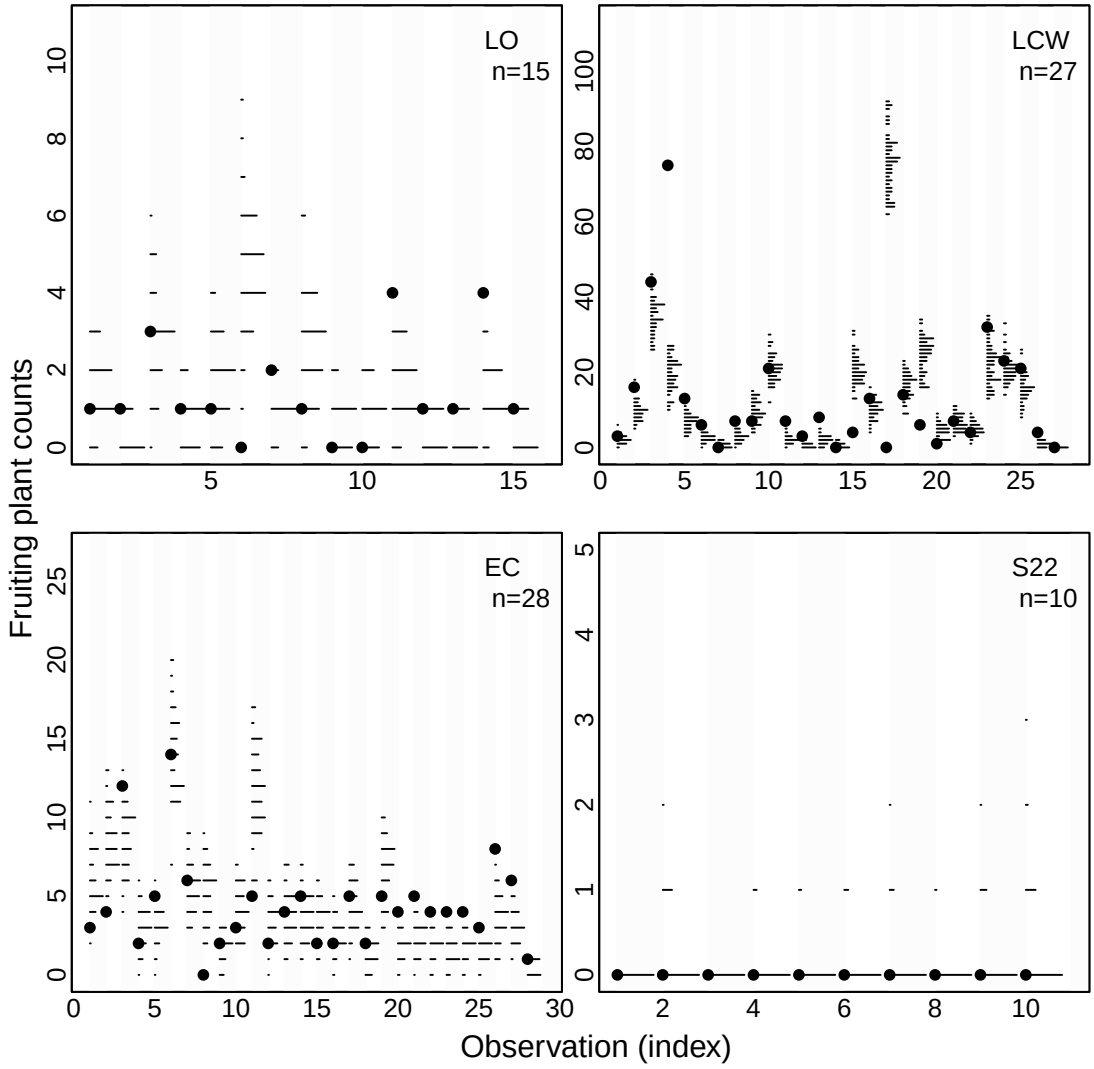


Figure 3: Distributions of simulated counts of fruiting plants for 2007 per plot, based on the model for seedling survival to fruiting. In each panel, the x-axis indexes individual observations (i.e., permanent plots) and the y-axis shows the number of fruiting plants. For each panel, the number of observations corresponds to the number of permanent plots that had seedlings this year. For each observation, we used the observed number of seedlings and the estimated population-level probability of seedling survival to fruiting to simulate 5000 values for fruiting plants. We then randomly sampled 50 values and plotted these as a discrete density plot. We repeated this for each observation in the population. The vertical length of the bars corresponds to how frequently the value was simulated. We then plotted the observed number of fruiting plants as a point.

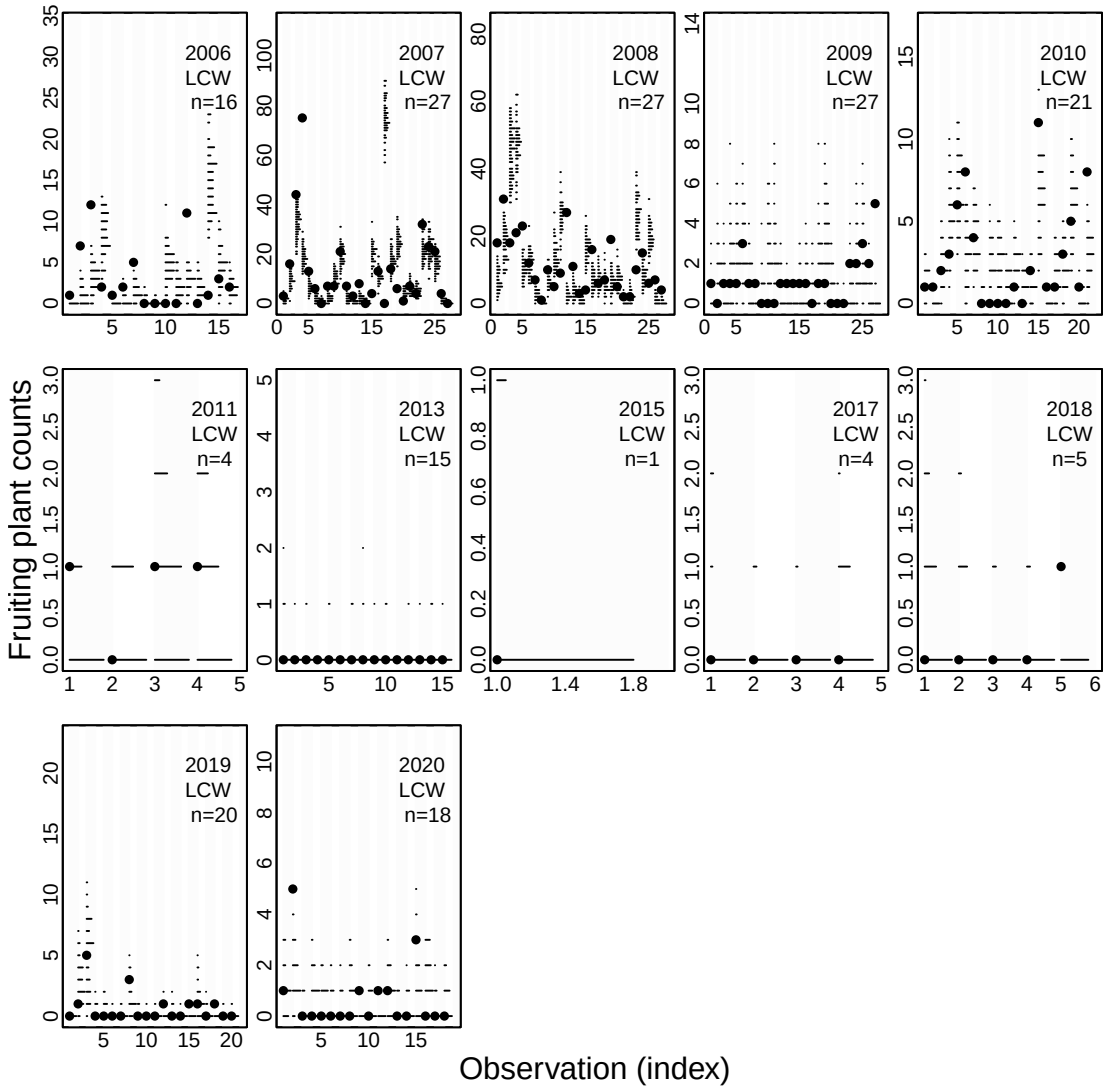


Figure 4: Distributions of simulated counts of fruiting plants for LCW for all years with observations, based on the model for seedling survival to fruiting. In each panel, the x-axis indexes individual observations (i.e., permanent plots) and the y-axis shows the number of fruiting plants. For each panel, the number of observations corresponds to the number of permanent plots that had seedlings that year. For each observation, we used the observed number of seedlings and the estimated population-level probability of seedling survival to fruiting to simulate 5000 values for fruiting plants. We then randomly sampled 50 values and plotted these as a discrete density plot. We repeated this for each observation. The vertical length of the bars corresponds to how frequently the value was simulated. We then plotted the observed number of fruiting plants as a point. There were no observations of seedling survival to fruiting at LCW in 2012, 2014, or 2016.

## *Fruits per plant*

For the model of fruits per plant, we used the posterior distribution of parameters from the model to simulate observations of total fruit equivalents per plant (2006-2012), total fruits per plant (2013-2020), and the proportion of fruits per plant that are damaged (2013-2020). The simulations used the population- and year-specific means from the models' joint posterior. For total fruit equivalents per plant and total fruits per plant, we calculated the mean and $\chi^2$ values as test statistics. We summarized the Bayesian p-values from these test statistics to represent the distribution of p-values across populations and years.

For total fruit equivalents per plant and total fruits per plant, we observed a good fit of the model to the mean estimate, as indicated by the narrow distribution of the test statistics around 0.5 (Figs. 5A, 6A). Bayesian p-values based on the $\chi^2$ test statistic generally indicated a reasonable fit of the model to the data (Figs. 5B, 6B). However, in some cases, we calculated Bayesian p-values close to 1, suggesting the model provided a poor fit to the data in some populations and years (orange points in Figs. 5B, 6B). In our study, we used the mean estimates of fruits per plant to calculate per-capita reproductive success. The absence of evidence for lack of fit suggests that we can use this model to make inferences about average patterns in fruit production by plants.

To examine cases with poor fit in more detail, we plotted samples from the posterior predictive distribution against the observations. Specifically, we compared populations in which the posterior predictive check indicated issues with model fit against populations in which we inferred good fit (orange vs. gray lines in Figs. 7, 8). For total fruit equivalents per plant, poor model fit was associated with an excess of 1s in the observed data, relative to the simulated data (Figs. 7 shows this pattern for observations from 2009).

For total fruits per plant, poor model fit was associated with small sample sizes and few fruits per plant in those observations. For example, in 2013 we were only able to count fruits on a single plant at S22 and this plant only had 2 fruits. Simulations from our model generated a broader distribution of possible values, and the larger values in the simulations led to larger $\chi^2$

values for the simulated values than the observed data (Figs. 7). In turn, the Bayesian p-value for the $\chi^2$ test statistic at this population in this year was one.

For damaged fruits per plant, we observed a good fit of the model to the mean estimate, as indicated by the narrow distribution of the test statistics around 0.5 (Fig. 9A). However, Bayesian p-values based on the $\chi^2$ test statistic were extreme (close to 0 or 1) for some populations across the years in the study, which indicated a poorer fit of the model to the data (Fig. 9B). We again compared populations in which the posterior predictive check indicated issues with model fit against populations in which we inferred good fit (orange vs. gray lines in Fig. 10). Extreme values of the $\chi^2$ test statistic suggested poor model fit and in 2013, LCE and OSR had high p-values while CP3 and OKRE had low p-values. High p-values were associated with an excess of 0s in the observed data, relative to the simulated values (Fig. 10). All counts of damaged fruits per plant at LCE and OSR in 2013 were zero, but we simulated 0-3 damaged fruits per plant. Low p-values were associated with observations that included several high counts of damaged fruits per plant, relative to the simulated values (Fig. 10). Both CP3 and OKRE had several observations of 7-8 fruits per plant, and the simulated values did not reproduce this small second peak of the distribution. Compare the cases with extreme p-values to those with less extreme p-values (populations with gray posterior predictive distributions in Fig. 10). In those cases, the distribution of observed data and simulated values overlaps more closely throughout the entire range of values.

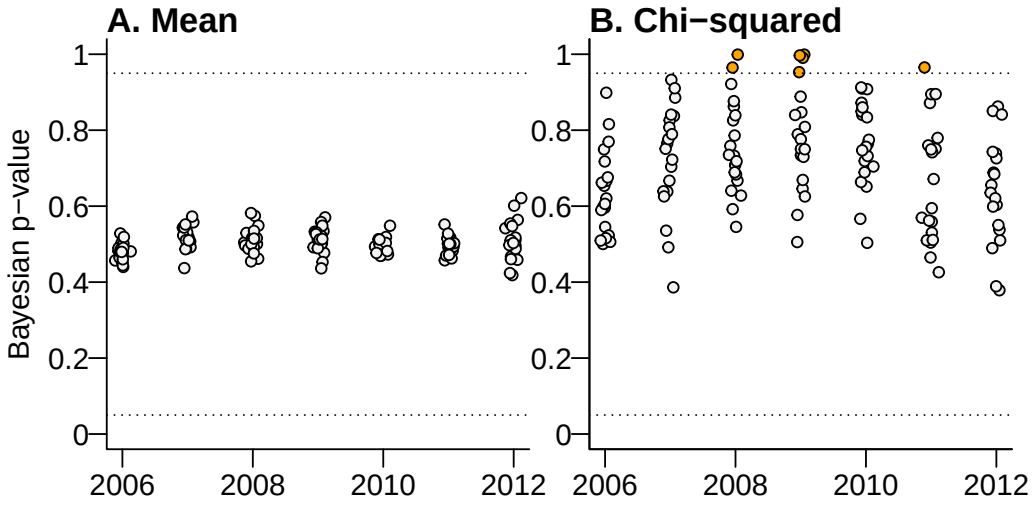


Figure 5: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of total fruit equivalents per plant (2006-2012). Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds at which the test statistic indicates lack of model fit. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

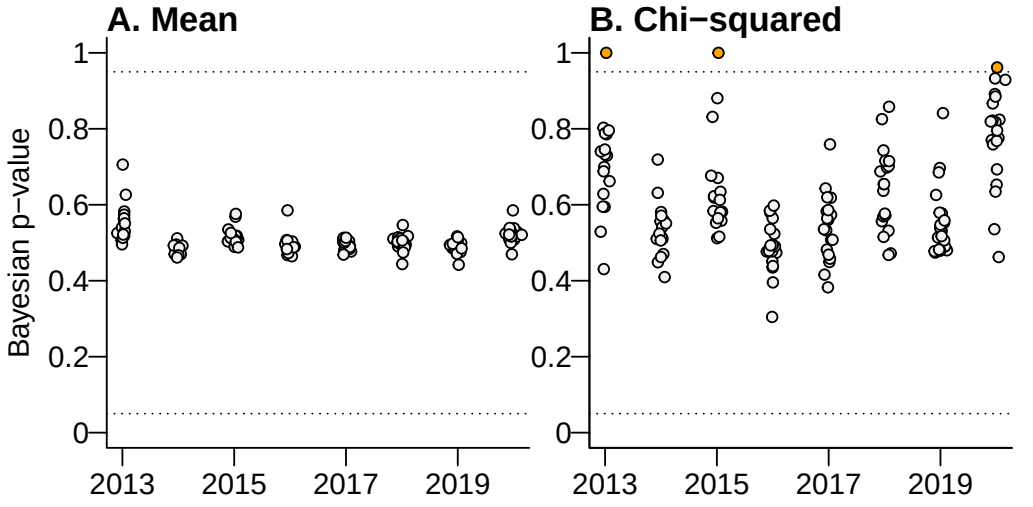


Figure 6: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of total fruits per plant (2013-2020). Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds at which the test statistic indicates lack of model fit. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

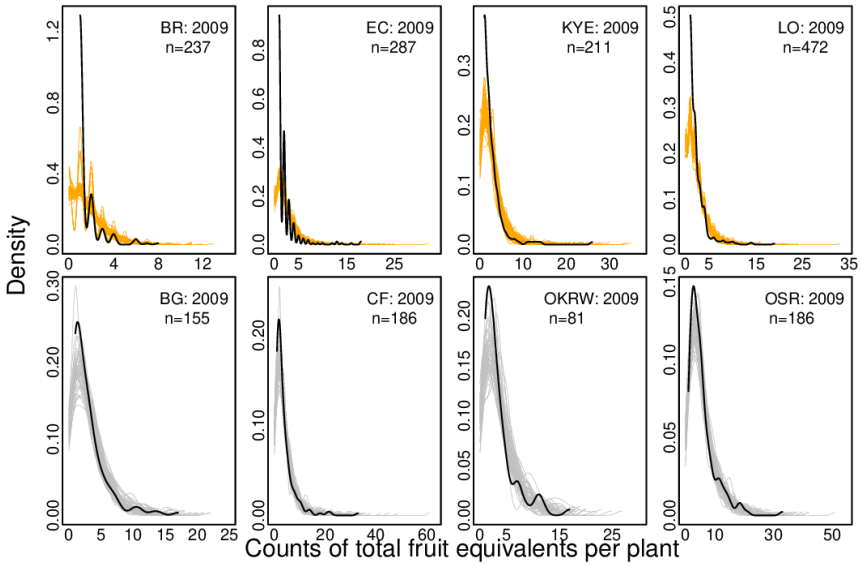


Figure 7: Distributions of counts for total fruit equivalents per plant (2006-2012). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line is the distribution of counts of damaged fruits per plant from the field observations.

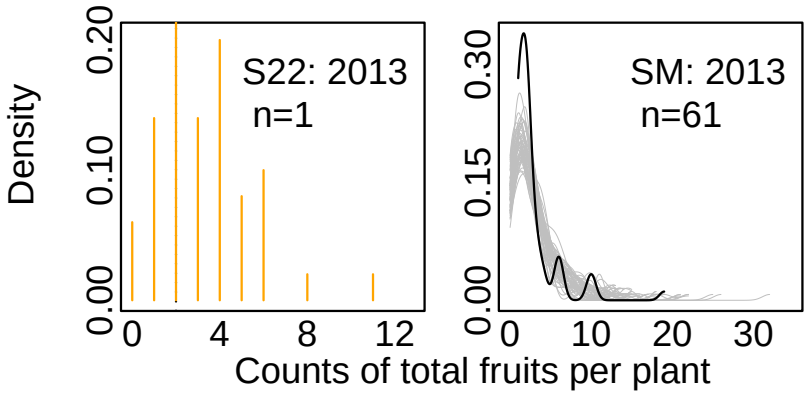


Figure 8: Distributions of counts for total fruits per plant (2013). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

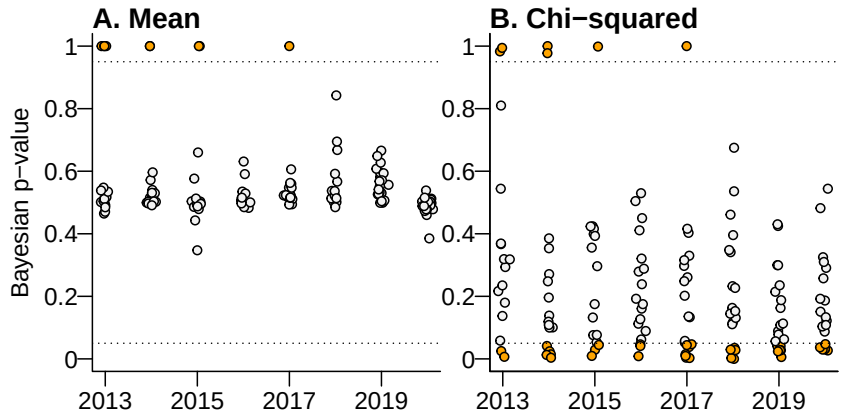


Figure 9: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of damaged fruits per plant (2013-2020). Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds at which the test statistic indicates lack of model fit. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

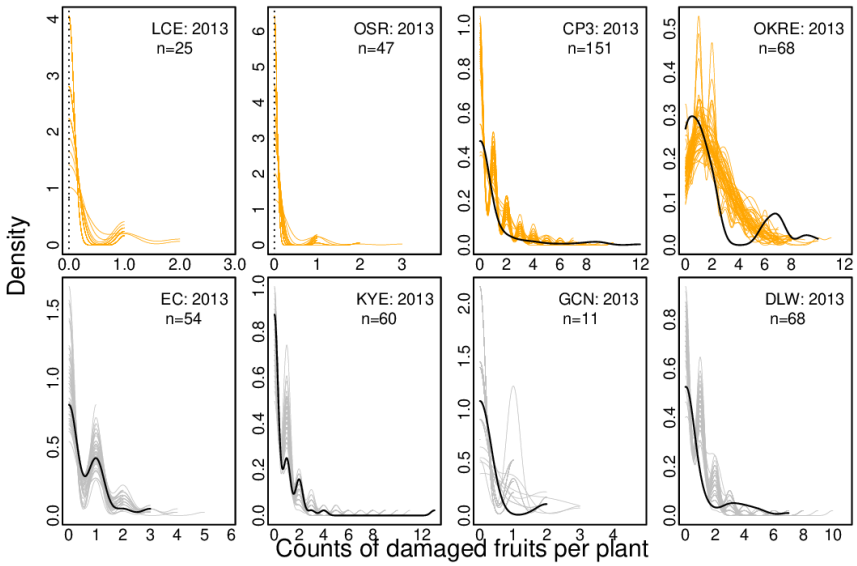


Figure 10: Distributions of counts for damaged fruits per plant (2013). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic. For LCE and OSR, the p-value for the $\chi^2$ test statistic was 1. For CP3 and OKRE, the p-value for the $\chi^2$ test statistic was 0. The plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

## *Seeds per fruit*

For the model of seeds per fruit, we simulated observations of seeds per undamaged fruit and seeds per damaged fruit for each year in each population. The simulations used the population- and year-specific means from the models' joint posterior. For seeds per undamaged fruit, we simulated values for fifteen years for each of the twenty populations. For seeds per damaged fruit, we simulated values for eight years for each of the twenty populations. For seeds per undamaged and seeds per damaged fruit, we calculated the mean and $\chi^2$ values as test statistics. We summarized the Bayesian p-values from these test statistics to represent the distribution of p-values across populations and years.

For seeds per undamaged fruit, we observed a good fit of the model to the mean estimate, as indicated by the narrow distribution of the test statistics around 0.5 (Fig. 11A). For seeds per damaged fruit, we also generally observed a good fit of the model. to the mean estimate, with a few exceptions in which p-values for the mean were 1 (Fig. 12A). For both seeds per undamaged fruit and seeds per damaged fruit, we also generally observed Bayesian p-values based on the $\chi^2$ test statistic that suggested a good fit of the model to the observations (Figs. 11B, 12B). In our study, we used the mean estimates of seeds per undamaged fruit and seeds per damaged fruit when we calculate per-capita reproductive success. The general absence of evidence for lack of fit suggests that we can use this model to make inferences about average patterns in seed set.

To examine cases with poor fit in more detail, we plotted samples from the posterior predictive distribution against the observations. Specifically, we compared populations in which the posterior predictive check indicated issues with model fit against populations in which we inferred good fit. For seeds per undamaged fruit, we examined 2010 and 2014 because these were the two years in which the $\chi^2$ test statistic was extreme for a population. In 2010, it appears that the model may underestimate the number of 0s relative to what was observed for population SM (Figs. 13). Based on samples from the posterior predictive distribution, it is likely that this is because simulations using model parameters generated fewer 0s than observed in the

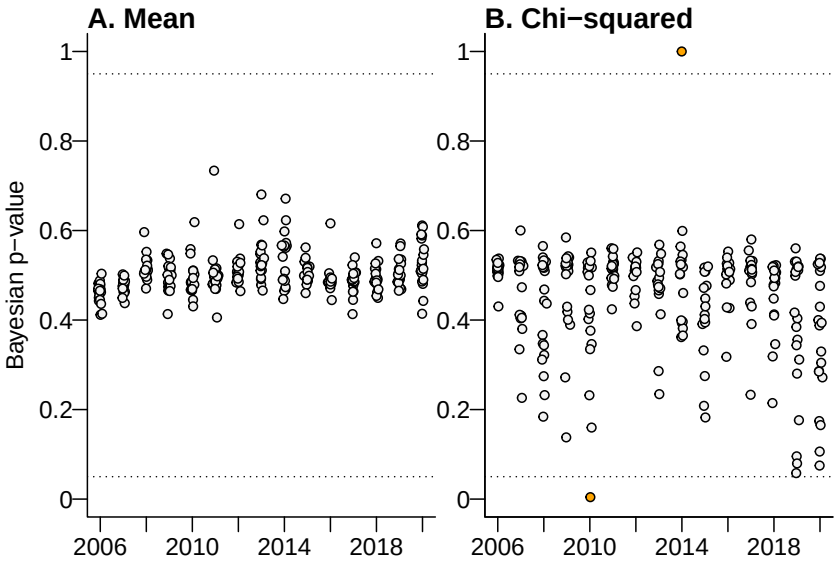


Figure 11: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of seeds per undamaged fruit (2006-2020). Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds at which the test statistic indicates lack of model fit. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

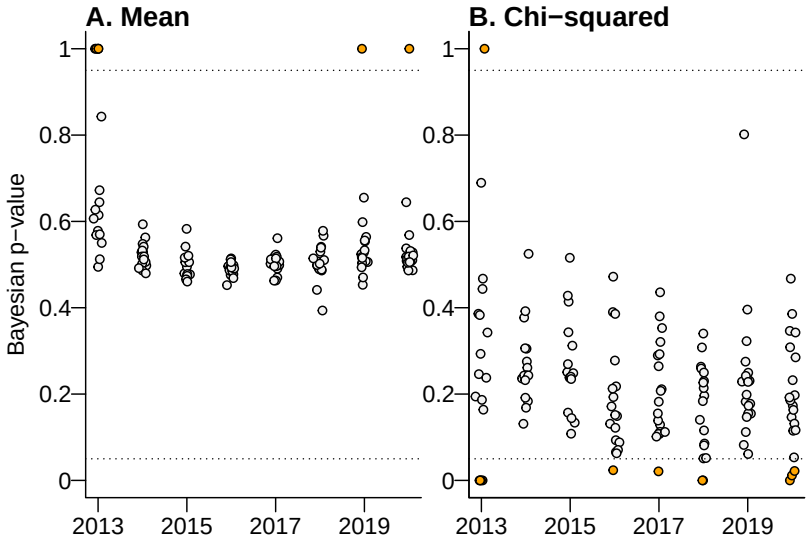


Figure 12: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of seeds per damaged fruit (2013-2020). Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

data. Individual samples from the posterior predictive distributions seem to be narrower than the observed distribution. In 2014, it seems likely that the poor fit is a result of the very small sample size at DLW (Figs. 14). Counts of seeds come from a single fruit, which likely means that the estimate of seeds per undamaged fruit for this year is more strongly pooled towards the population-level mean. In turn, this biases the simulated values towards the population-level mean and away from the mean in that particular year. This manifests as $\chi^2$ values that are larger for the simulated data than the observed data, which translates to a high Bayesian p-value.

For seeds per damaged fruit, several years showed extreme $\chi^2$ test statistics and we chose to examine 2013 because it had both high and low values. In general, lack of model fit was associated with very small sample sizes of a single fruit (compare orange vs. gray in Figs. 15, 16). We observed either p-values of zero or one in several cases where the single observation was zero seeds per damaged fruit. The posterior predictive distribution for the former (Fig. 15) and latter (Fig. 16) case showed very similar patterns: the distribution of simulated counts included zero but also some higher values. Whether the p-value is zero or one is simply a consequence of the extent to which the year-level estimate is pooled towards the population-level mean. When pooling is stronger, the distribution of simulated values is slightly greater, as in SM in 2013, and the $\chi^2$ test statistic for simulated values is greater than that for the observed values.

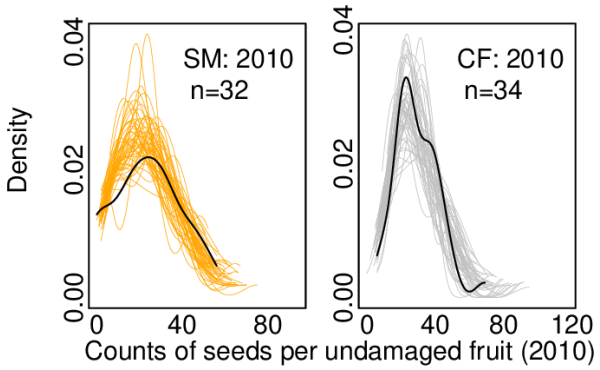


Figure 13: Distributions of counts for seeds per undamaged fruit (2010). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

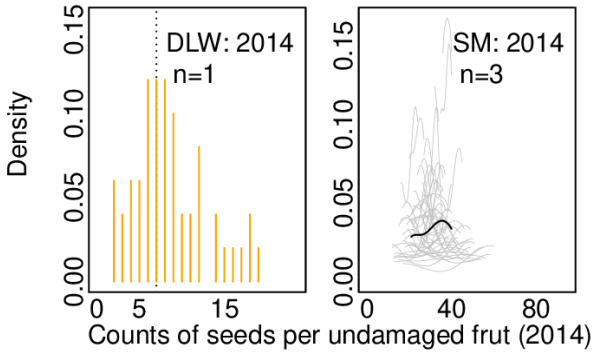


Figure 14: Distributions of counts for seeds per undamaged fruit (2014). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

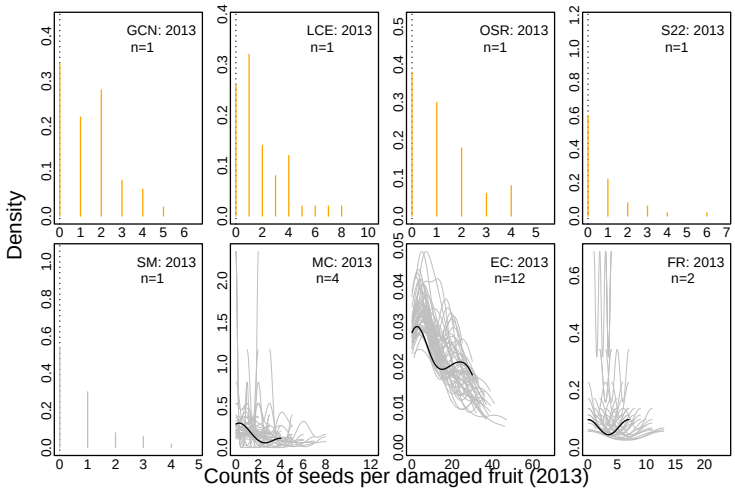


Figure 15: Distributions of counts for seeds per damaged fruit (2013). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic being less than 0.05; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

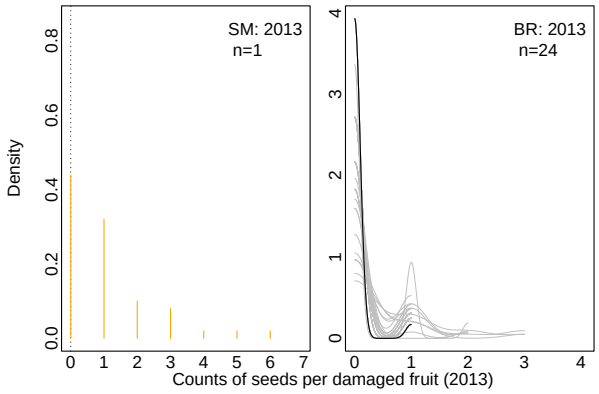


Figure 16: Distributions of counts for seeds per damaged fruit (2013). Each orange or gray line is the distribution of counts from one *in silico* replicate of the experiment. We simulated 5000 datasets, and randomly sampled 50 for plotting. The plots with orange lines correspond to models that show lack of fit based on the $\chi^2$ test statistic being greater than 0.95; the plots with gray lines correspond to a random set of models that show good fit based on the $\chi^2$ test statistic. The superimposed black line (for multiple observations) or vertical dotted line (for a single observation) is the distribution of counts of damaged fruits per plant from the field observations.

## *Belowground statistical models*

For the model of seed survival and germination, we simulated replicate binomial trials corresponding to germination and survival. In our simulations, we drew values for the parameters of the survival function and germination probabilities from the joint posterior of the model for the seed bag experiment. We used those values to simulate replicate binomial trials for the times at which we excavated seed bags and counted intact seeds or germinants. Conceptually, we repeated the seed bag experiments *in silico* using the estimated parameters to generate many replicate datasets and compare those to the observed dataset. We calculated the mean and $\chi^2$ values as test statistics. We summarized the Bayesian p-values from these test statistics to represent the distribution of p-values across populations and years.

Posterior predictive checks with the mean as a test statistic suggested that the model for the seed bag burial experiment adequately fit the data for germination (Fig. 17A). The majority of Bayesian p-values fell in the middle of the range $[0, 1]$. However, all Bayesian p-values for the $\chi^2$ test statistic were less than 0.05; this was true in all populations in all years (Fig. 17B).

To examine model fit in more detail, we plotted samples from the posterior predictive distribution and overlapped the distribution of observations. We examined the fit of the model to each population in all three years of the seed bag experiment (Figs. 18, 19, 20). In the first two years of the experiment, only one population showed evidence for lack of fit based on the mean test statistic (OKRW in 2006, BR in 2007). In the third year, five populations showed evidence for lack of fit. Examining plots of the posterior predictive distribution reveals that the models generally fit the mean and range of the data well. However, the individual posterior predictive distributions tended to be quite narrow and tended to not reproduce long tails of the observed data (e.g. GCN in 2006). The inability to match the variance likely reflects the relatively small sample size that we had for these observations; in each year, we had at most 11 seed bags in which we recorded observations. The mismatch between the distribution of the observations and the posterior predictive distribution explains the ubiquitous small $\chi^2$ test statistics (Fig. 17B). Poor

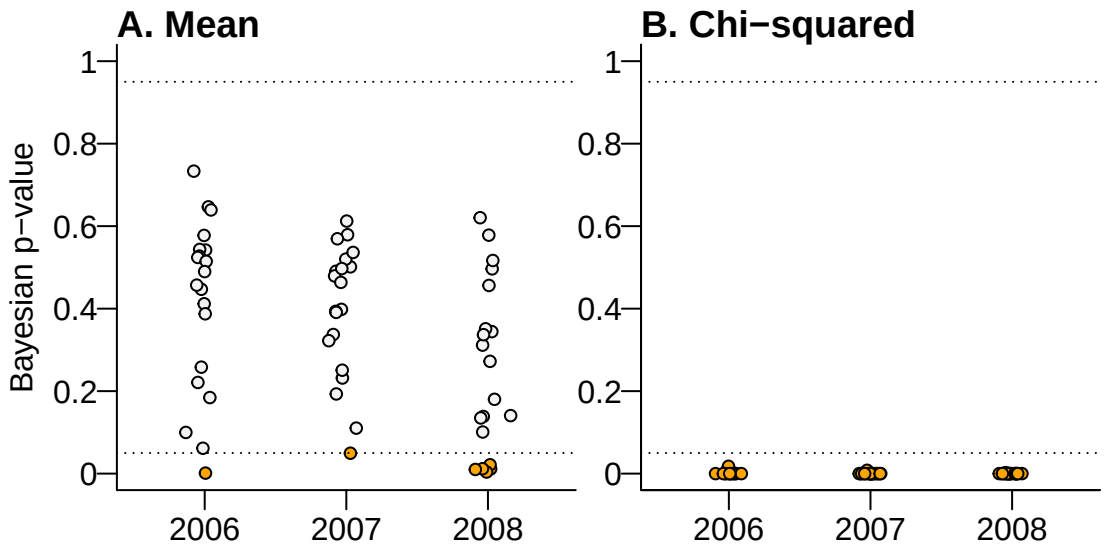


Figure 17: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of germinants from seeds in the seed bag burial experiment that survived to January of year $t+1$, for seeds produced in year $t$. Each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

model fit based on the $\chi^2$ test statistic was associated with more small values in the observed data, relative to the simulated data.

To understand why some population-year combinations also showed evidence for lack of fit based on the mean test statistic, we plotted the simulated data at the level of seed bags for OKRW across all three years (Fig. 21). These plots illustrate both why there is a lack of fit based on the mean test statistic in 2006, as well as why the $\chi^2$ test statistics are small. Although all three years had 10 seed bags, the years differed substantially in terms of how many seeds remained by January (i.e., the number of binomial trials). The first and third years had $\tilde{6}0$ seeds across 10 bags, while the second year had almost three times as many seeds. The year-level estimates are thus more influenced by information from the second year than from the first and third years. The distribution of simulated values in 2006, in particular, falls on the low end of below the observed values, especially for observations 1, 9, and 10. These also happen to be observations with many trials, which helps explain why the mean test statistic is small. In addition, the observed

germinant counts across all years tend to fall both below and above the simulated values. This means that the $\chi^2$ based on observations is greater than $\chi^2$ based on the simulated values, which leads to a small p-value. As before, while the model estimates average germination, across bags, sample sizes of 10 bags are likely too small to accurately infer the variability in germination across bags.

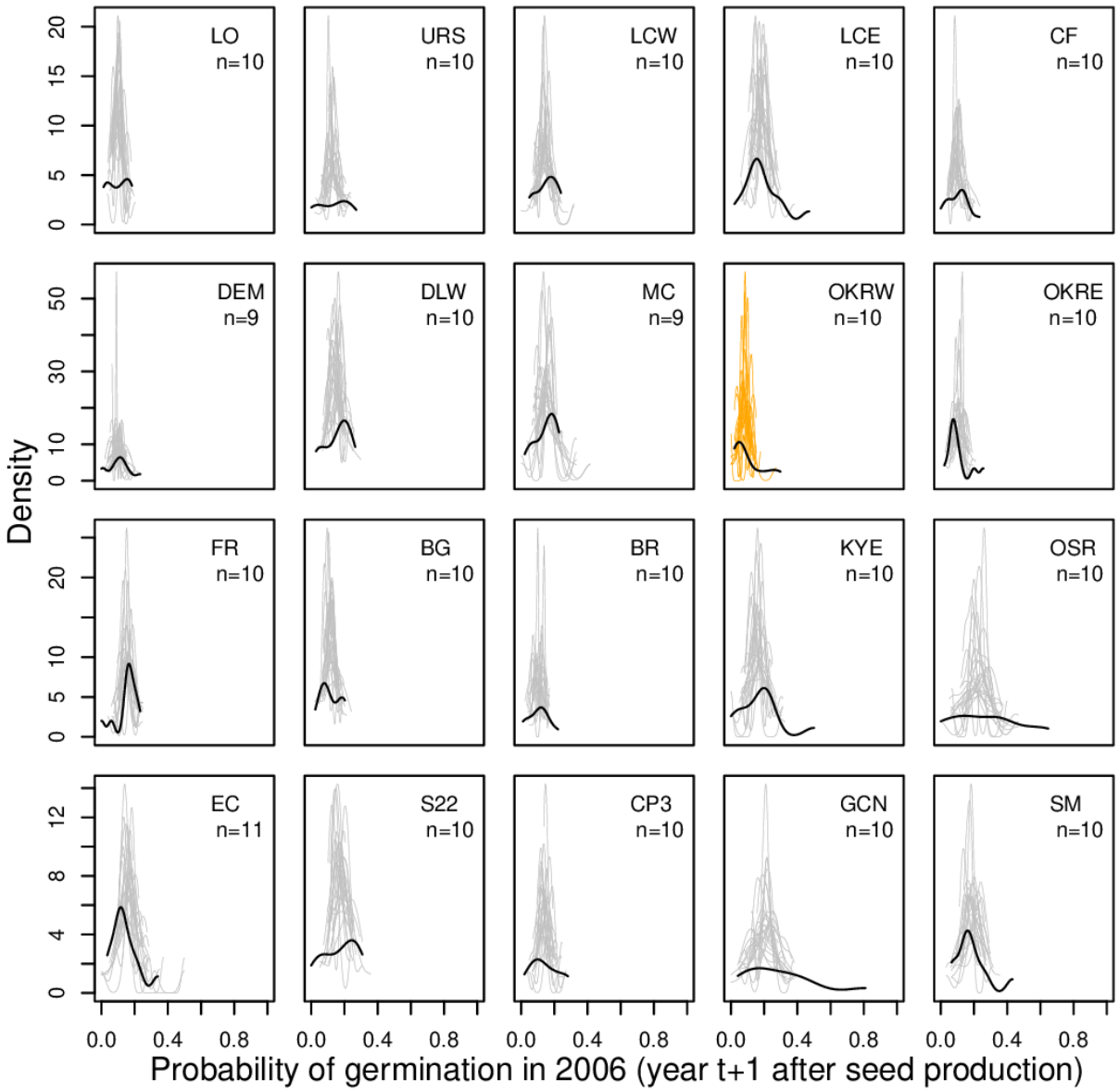


Figure 18: Distributions of germinants from January 2006 in the seed bag burial experiment for seeds that survived to January of year $t+1$, for seeds produced in year $t$. Each orange or gray line is the distribution of counts from one *in silico* replicate of germinant counts, based on the statistical model fit to the observations from the seed bag burial experiment. We simulated datasets using the observed counts of seeds intact in January as the number of binomial trials, and the estimated population-level probability of germination. We then randomly sampled 50 simulated datasets for plotting. The plots with orange lines correspond to models that show lack of fit based on the mean test statistic; the plots with gray lines correspond to models that show good fit based on the mean test statistic. The superimposed black line (for multiple observations) is the distribution of counts of germinants from the field experiment.

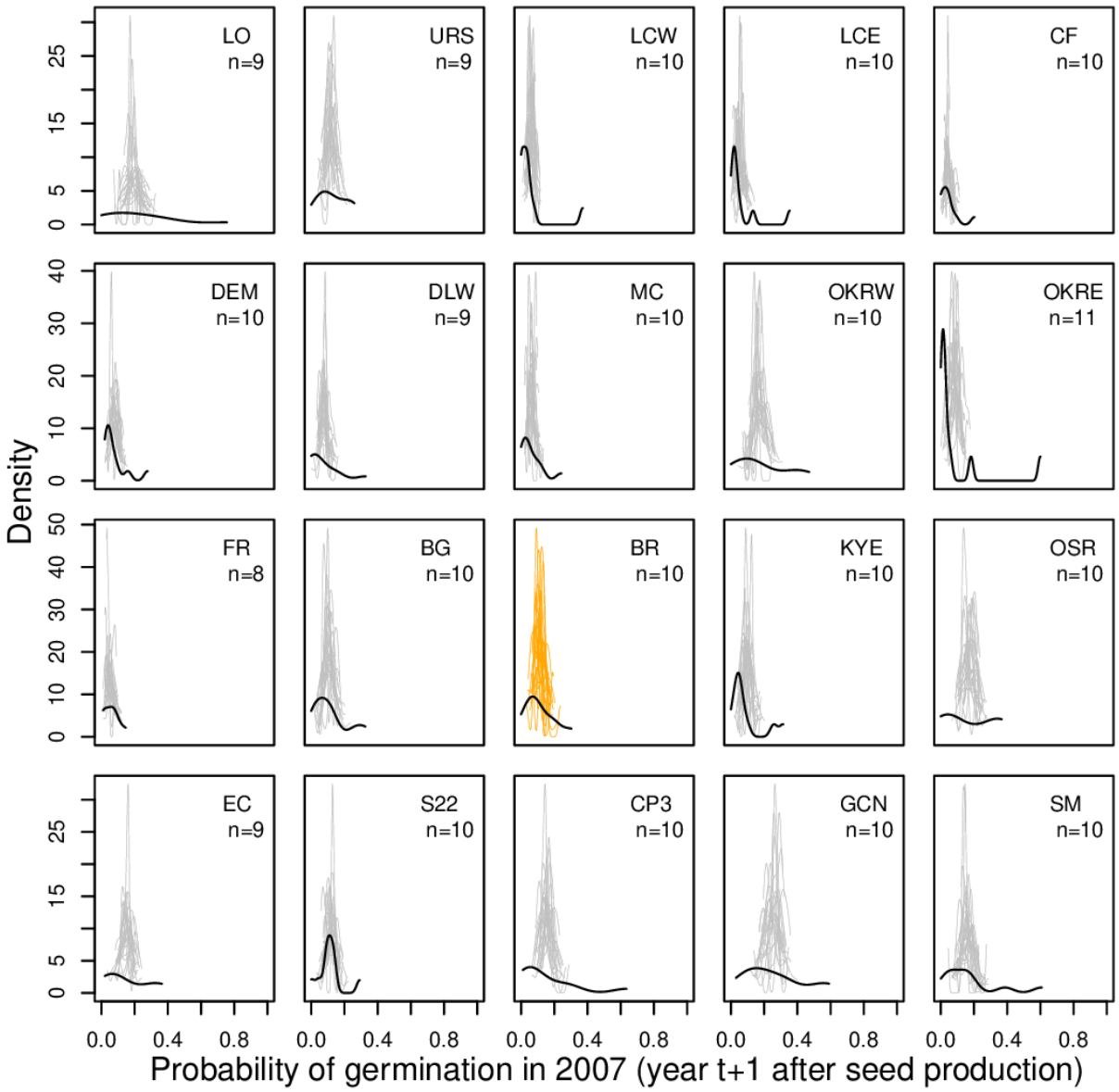


Figure 19: Distributions of germinants from January 2007 in the seed bag burial experiment for seeds that survived to January of year $t+1$, for seeds produced in year $t$. Each orange or gray line is the distribution of counts from one *in silico* replicate of germinant counts, based on the statistical model fit to the observations from the seed bag burial experiment. We simulated datasets using the observed counts of seeds intact in January as the number of binomial trials, and the estimated population-level probability of germination. We then randomly sampled 50 simulated datasets for plotting. The plots with orange lines correspond to models that show lack of fit based on the mean test statistic; the plots with gray lines correspond to models that show good fit based on the mean test statistic. The superimposed black line (for multiple observations) is the distribution of counts of germinants from the field experiment.

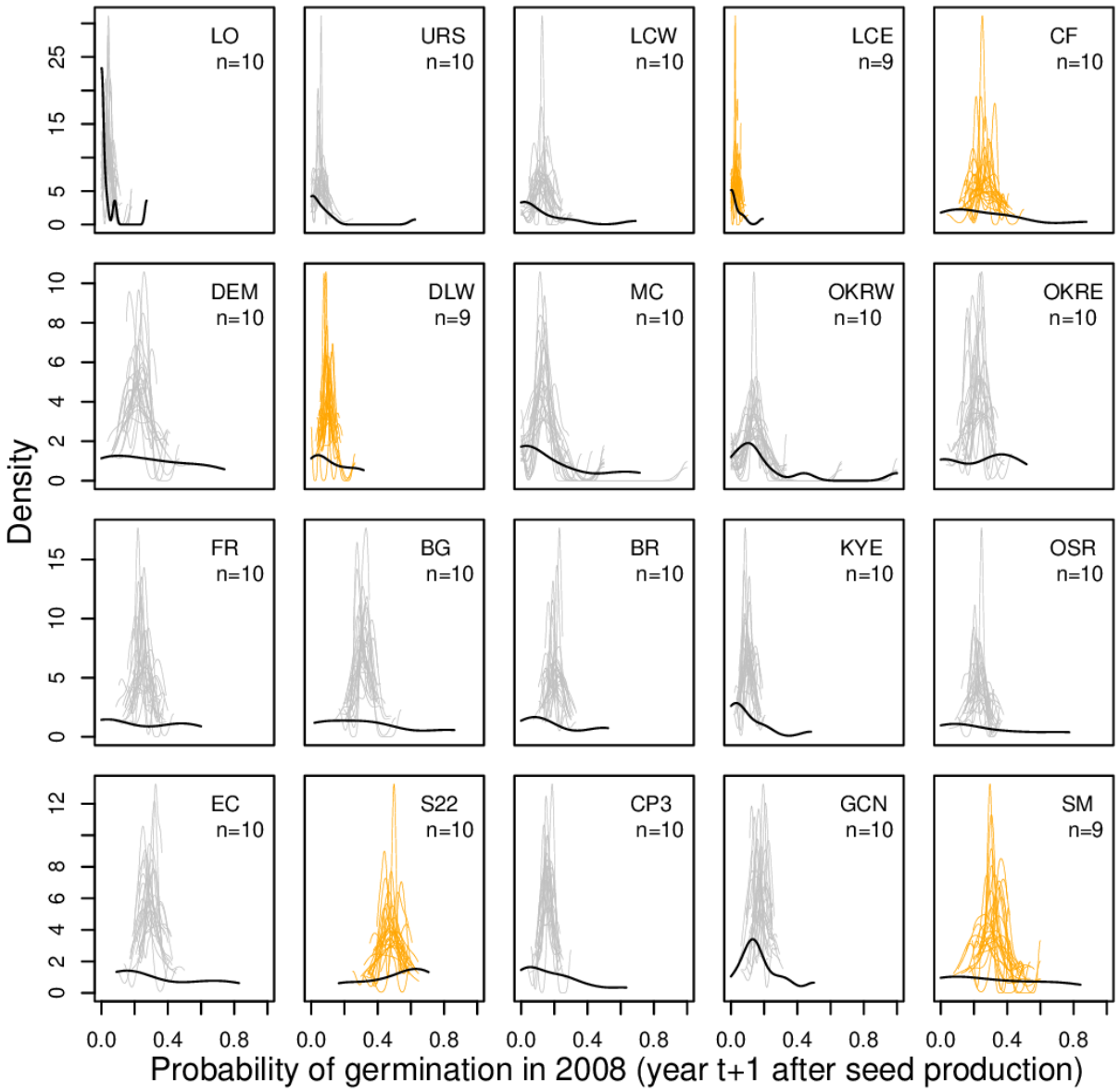


Figure 20: Distributions of germinants from January 2008 in the seed bag burial experiment for seeds that survived to January of year $t+1$, for seeds produced in year $t$. Each orange or gray line is the distribution of counts from one *in silico* replicate of germinant counts, based on the statistical model fit to the observations from the seed bag burial experiment. We simulated datasets using the observed counts of seeds intact in January as the number of binomial trials, and the estimated population-level probability of germination. We then randomly sampled 50 simulated datasets for plotting. The plots with orange lines correspond to models that show lack of fit based on the mean test statistic; the plots with gray lines correspond to models that show good fit based on the mean test statistic. The superimposed black line (for multiple observations) is the distribution of counts of germinants from the field experiment.

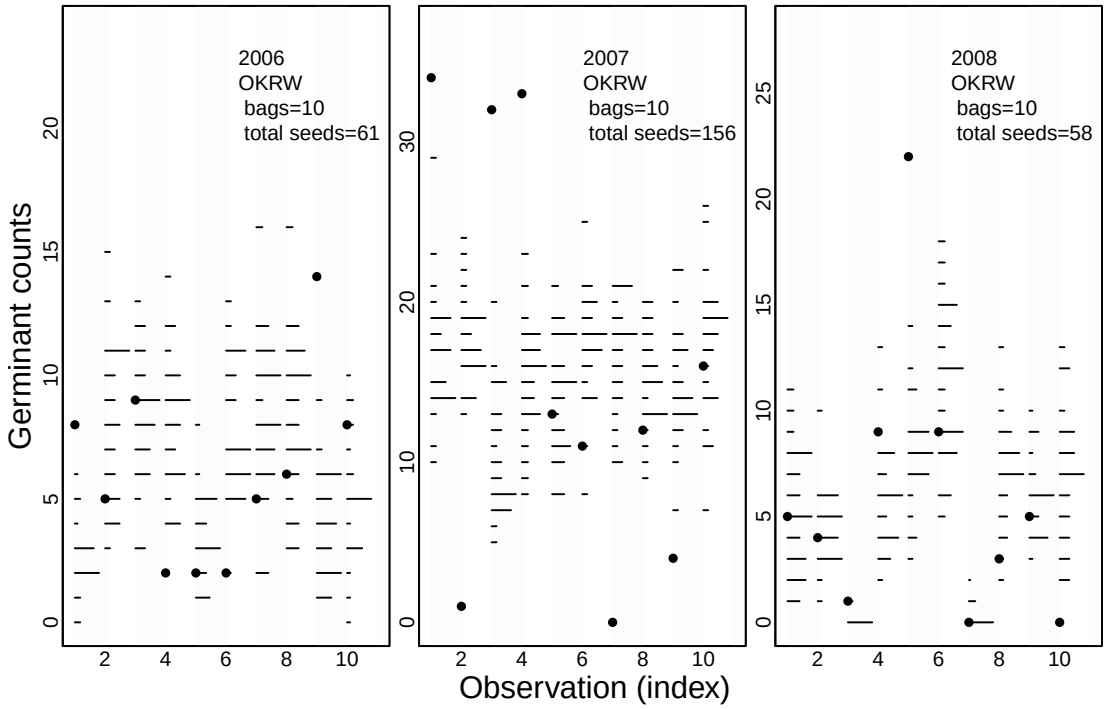


Figure 21: Distributions of germinants from January 2006, 2007, and 2008 for OKRW in the seed bag burial experiment for seeds that survived to January of year $t+1$, for seeds produced in year $t$. In each panel, the x-axis indexes individual observations (i.e., seed bags) and the y-axis shows the number of germinants. For each panel, the number of observations corresponds to the number of seed bags with observations that year. For each observation, we used the observed number of total seeds in January and the estimated population-level probability of germination to simulate values for the number of germniants. We then randomly sampled 50 values and plotted these as a discrete density plot. We repeated this for each observation. The vertical length of the bars corresponds to how frequently the value was simulated. We then plotted the observed number of germinants as a point.

Posterior predictive checks with the mean as a test statistic suggested that the model for the seed bag burial experiment generally fit the data for intact seeds (Fig. 22A, C E). For intact seeds in the first January of the experiment, the majority of Bayesian p-values fell in the middle of the range $[0, 1]$ for all years and populations (Fig. 22A). For intact seeds in the first October of the experiment, a greater number of p-values exhibited extreme values close to 1 ((Fig. 22C). By the second January of the experiment, p-values were fell at both extremes. With a single exception, all Bayesian p-values for the $\chi^2$ test statistic were less than 0.05 (Fig. 22B, D, F).

We plotted samples from the posterior predictive distribution and overlapped the distribution of observations. We focused on two populations in the first round of the seed bag burial experiment, for seed bags buried in October 2005 (Fig. 23). We show the probability of seed survival through time for one cohort of seed bags because the model is jointly fit to observations from the entire experiment. Examining plots of the posterior predictive distribution reveals that the models underfit the variability of the data. Individual posterior predictive distributions tended to be quite narrow and did not reproduce long tails of the observed data, both for models that had good fit to the mean (bottom row) and those that did not (top). The inability to match the variance of observations reflects the relatively small sample size that we had for these observations, as with the germinant observations. The mismatch between the observations and the posterior predictive distribution explains the ubiquitous small $\chi^2$ test statistics (Fig. 22B), as the greater variance in observations would lead to greater $\chi^2$ values for the observed data relative to the simulated data.

Use of a joint model that was fit to observations from the whole experiment also explained why we observed extreme p-values. Consider the example in the top row in Fig. 23. The observed distribution of seed survival probability increases from twelve to sixteen months. This is impossible–the cumulative probability that a seed remains intact can not increase–but because the seed bags are removed after each year, individual seed bags may follow different trajectories that produce these observations. The model 'resolves' this by splitting the difference; the estimated probability is higher than observed after twelve months but lower than observed after

sixteen months. This leads to a mismatch between the posterior predictive distribution and the observed distribution (compare the colored lines and the solid black line).

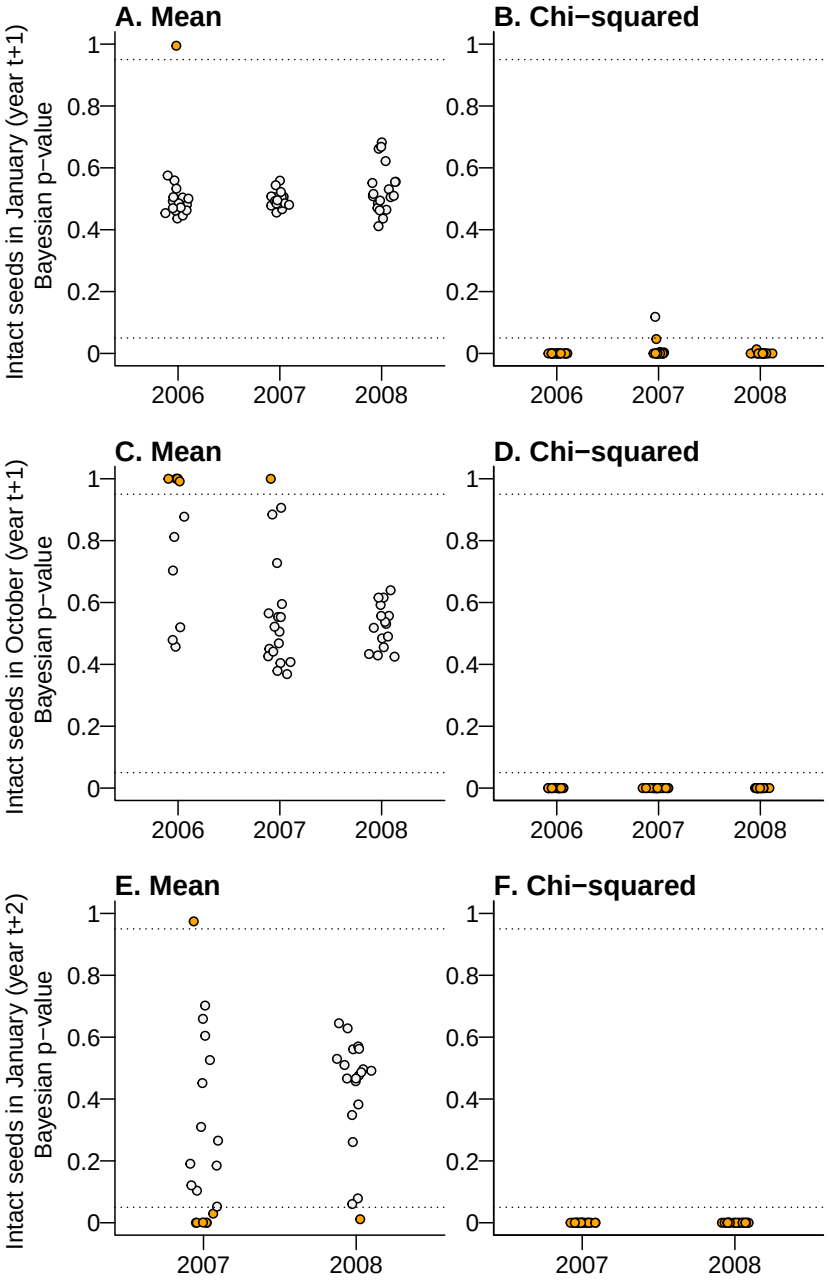


Figure 22: Bayesian p-values for the mean and $\chi^2$ value for observations of intact seeds in the seed bag burial experiment for seeds produced in year $t$. (A and B) Bayesian p-values for observations of seeds in January of year $t+1$. (C and D) Bayesian p-values for observations of seeds in October of year $t+1$. (E and F) Bayesian p-values for observations of seeds in January of year $t+2$. In all panels, each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

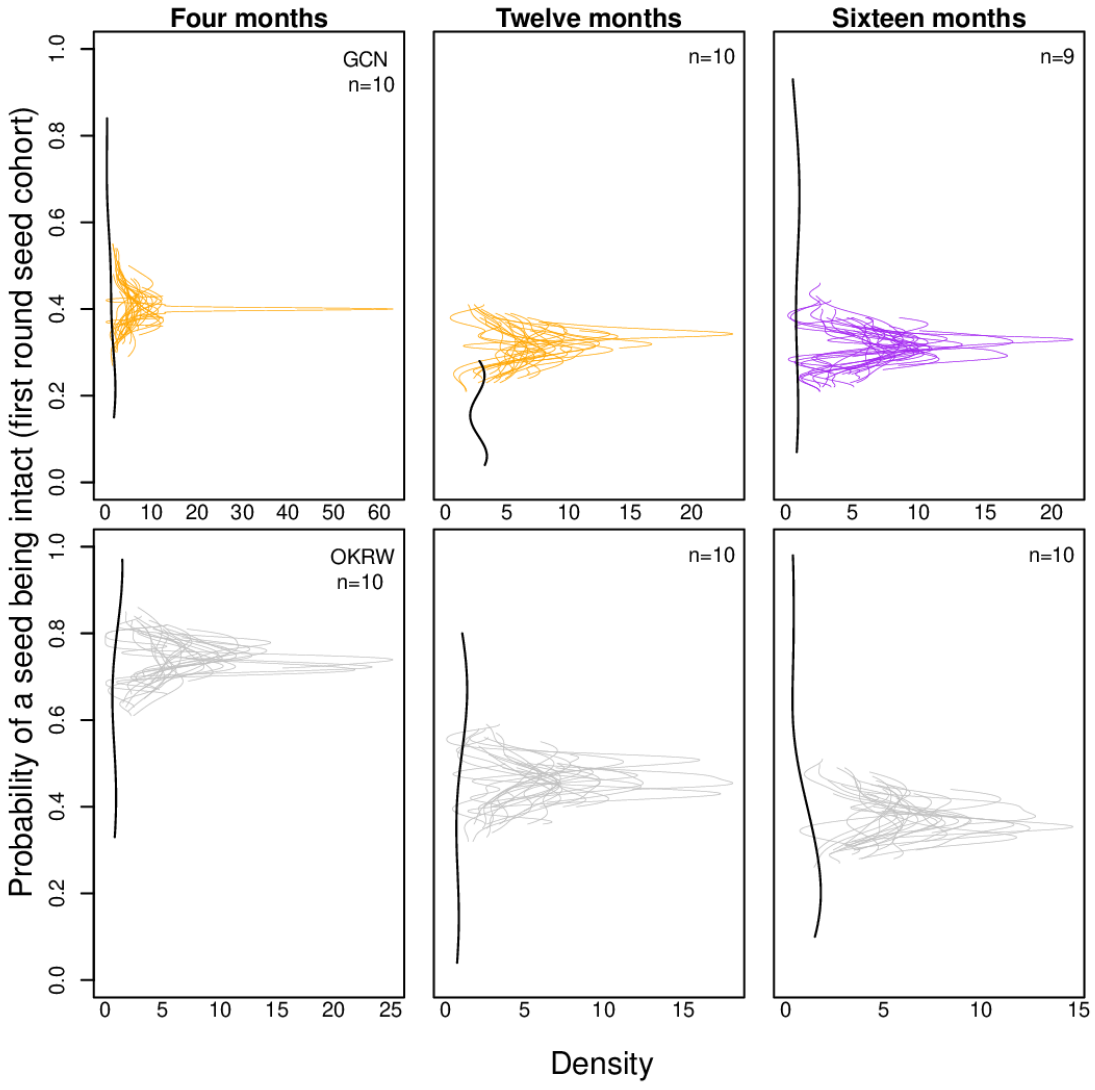


Figure 23: Distributions of probability that a seed remains intact in January 2006 (four months), October 2006 (twelve months), and January 2007 (sixteen months) in the seed bag burial experiment for seeds from the first round (seeds buried in October 2005). We simulated intact seed counts using the number of seeds starting the experiment, the estimated probability of germination, and the estimated probability of seed survival. We then randomly sampled 50 simulated datasets for plotting. The top row shows a population that exhibited extreme p-values of the mean test statistic; the bottom row shows a population that showed evidence of good fit based on the mean test statistic. Each orange, purple or gray line is the distribution of seed survival probability from one *in silico* replicate, based on the statistical model fit to the observations from the seed bag burial experiment. The plots with orange lines correspond to a lack of fit based on high values of the mean test statistic; the plots with the purple lines correspond to a lack of fit based on low values of the mean test statistic; the plots with gray lines correspond to models that show good fit. The superimposed black line (for multiple observations) is the distribution of seed survival probability from the field experiment.

Finally, we conducted posterior predictive checks for the number of seedlings emerging from plots. We only conducted model checks for seedling emergence in 2008, as this was the only years for which we could combine the observations from the seed bag burial experiment, seed rain, and seedling emergence to perform the model checks. Posterior predictive checks with the mean as a test statistic suggested a lack of fit, with Bayesian p-values being both high and low (Fig. 24A). Bayesian p-values for $\chi^2$ were low for all populations, though two populations were less extreme (gray points in Fig. 24B).

Individual posterior predictive distributions were more idiosyncratic than the observed data but generally matched the range of the observed data (see x-axis in Fig. 25). The mean of simulated values was typically greater than the observed mean (orange lines) in populations in which we had few observations of seedlings. Vice-versa, the mean of simulated values was typically smaller than the observed mean (purple lines) in populations in which we had many observations of seedlings. Because we summed across plots to obtain the number of seeds produced and seedlings emerging in a transect, we had relatively few samples for fitting this model even though the total number of seedlings in the plots could be quite large (n in the figure). Recall that the number of seeds emerging is also the product of many processes: seed survival over multiple time periods, germination (or absence thereof), and seed rain. Bias in estimating any of these could contribute to the mismatch we see between the posterior predictive distribution and the observed data.

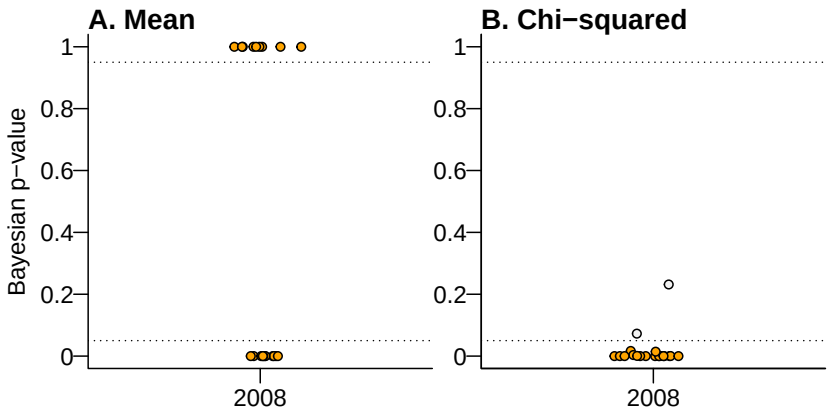


Figure 24: Bayesian p-values for the (A) mean and (B) $\chi^2$ value for observations of seedlings emerging in permanent plots in 2008. These model checks were calculated using values simulated using seed rain in permanent plots in summer 2006 and 2007, estimates of seed survival and germination from the seed bag burial experiment, and the observed number of seedlings in permanent plots. Each point corresponds to the p-value in of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

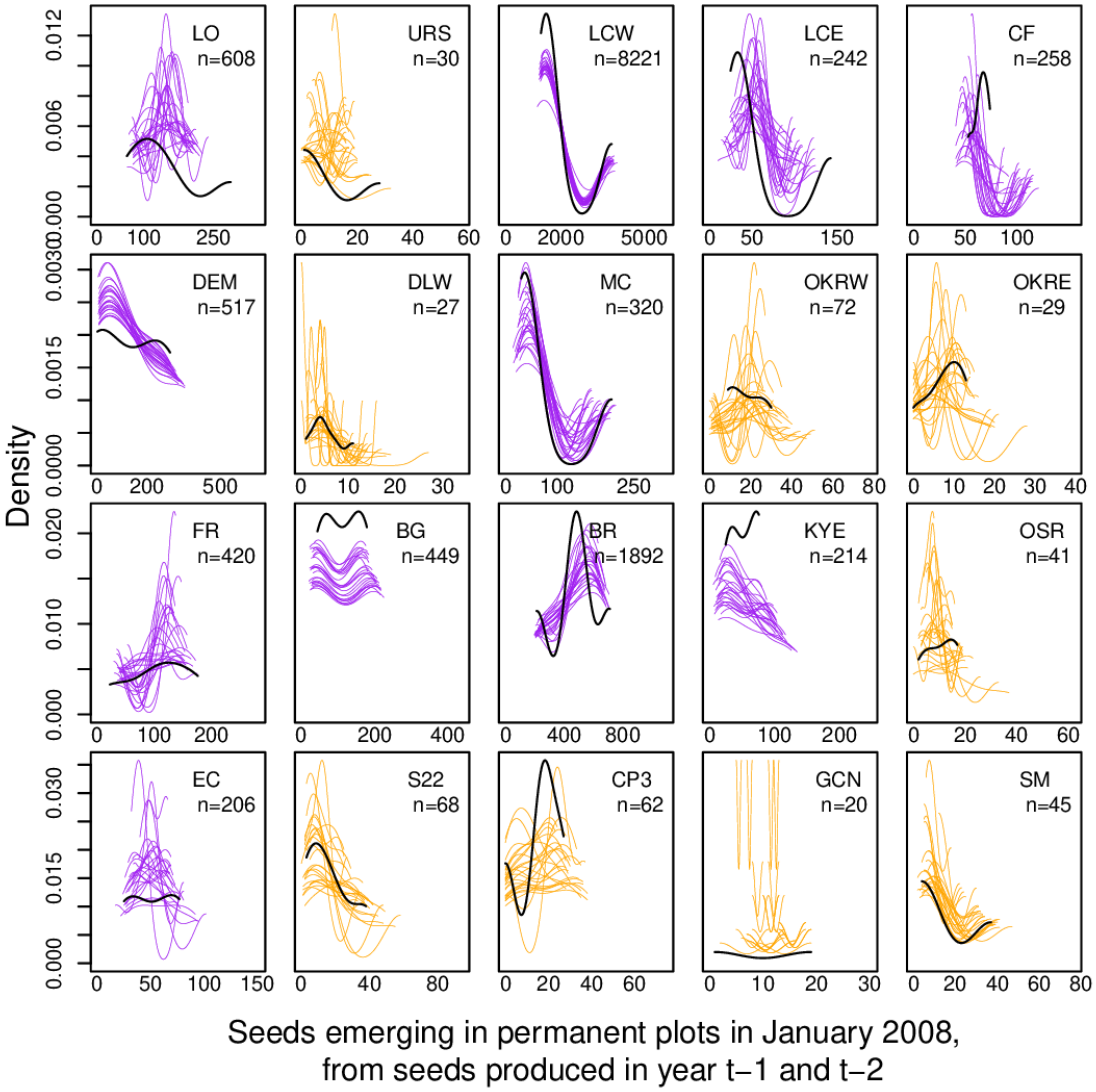


Figure 25: Distributions of number of seeds emerging in permanent plots in January 2008. We simulated intact seed counts using the seed rain in permanent plots, the estimated probability of germination, and the estimated probability of seed survival. We then randomly sampled 50 simulated datasets for plotting. Each orange, purple or gray line is the distribution of seed survival probability from one *in silico* replicate, based on the statistical model fit. The plots with orange lines correspond to a lack of fit based on high values of the mean test statistic, and the plots with the purple lines correspond to a lack of fit based on low values of the mean test statistic. The superimposed black line (for multiple observations) is the distribution of seed survival probability from the field experiment.

## *Viability trials*

For the models for germination and viability for observations from the viability trials, we simulated replicate binomial trials using the number of seeds starting each part of the viability trial. In our simulations, we drew values for germination and viability from the joint posterior of the model for the viability trials. We used those values to simulate replicate binomial trials. Conceptually, we repeated the viability trials *in silico* using the estimated parameters to generate many replicate datasets and compare those to the observed dataset. We calculated the mean and $\chi^2$ values as test statistics. We summarized the Bayesian p-values from these test statistics to represent the distribution of p-values across populations and years.

Posterior predictive checks with the mean as a test statistic suggested that the models for the viability trials adequately fit the data for germinants (Fig. 26A,C) and seeds staining viable (Fig. 27A,C). Models fit similarly for data from age 1 and age 2 seeds. Across all sets of observations, three-quarters to all of p-values for the $\chi^2$ test statistic were extreme, falling below 0.05; this was true for geminants (Fig. 26B,D) and seeds staining viable (Fig. 27B,D). Lack of fit is primarily the result of models failing to represent the variability in observations. This issue is very similar to the lack of fit we observed in the models for observations of germinants in the field seed bag burial experiment. The plots of the posterior predictive distribution and the observed data distribution (not shown here) are generally very similar to those for models of germinants from the seed bag burial experiment in the field. With few samples (there were always 11 or fewer seed bags that were tested in the viability trials), it is challenging to estimate both the mean and the variance. The lack of fit based on the $\chi^2$ test statistic mainly reflects the models' inability to match the distribution.

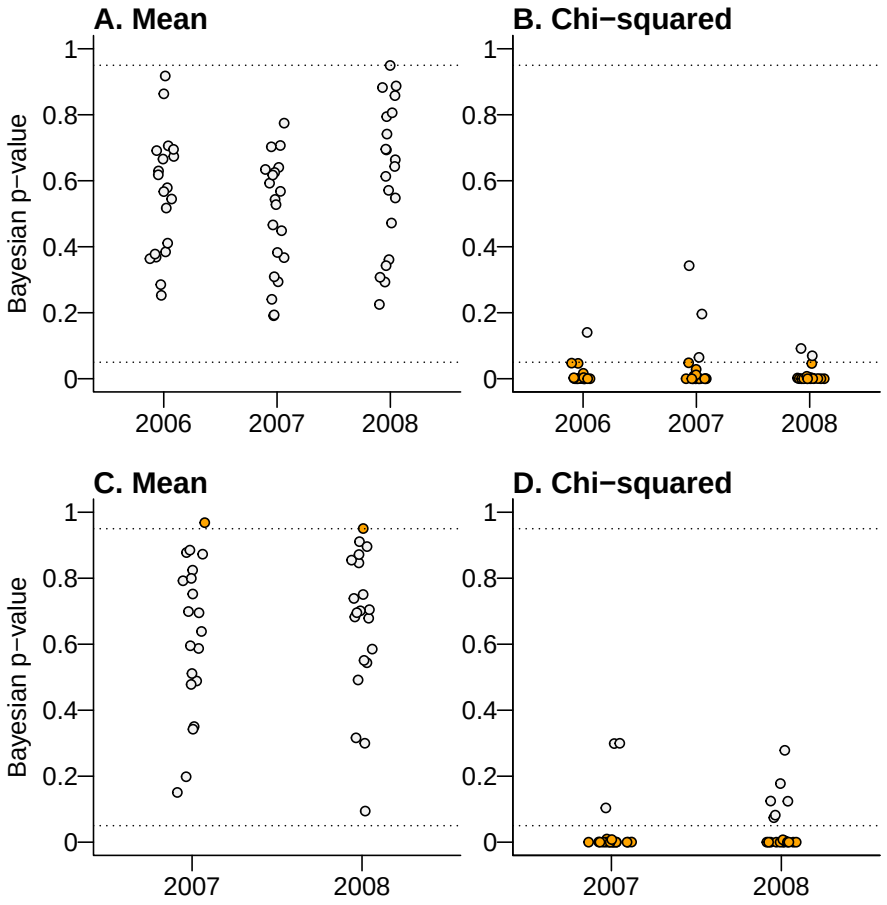


Figure 26: Bayesian p-values for the mean and $\chi^2$ value for observations of germinants in the viability trials. (A and B) Bayesian p-values for observations for age 1 seeds (seeds that were tested one year after being buried). (C and D) Bayesian p-values for observations for age 2 seeds (seeds that were tested two years after being buried). In all panels, each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

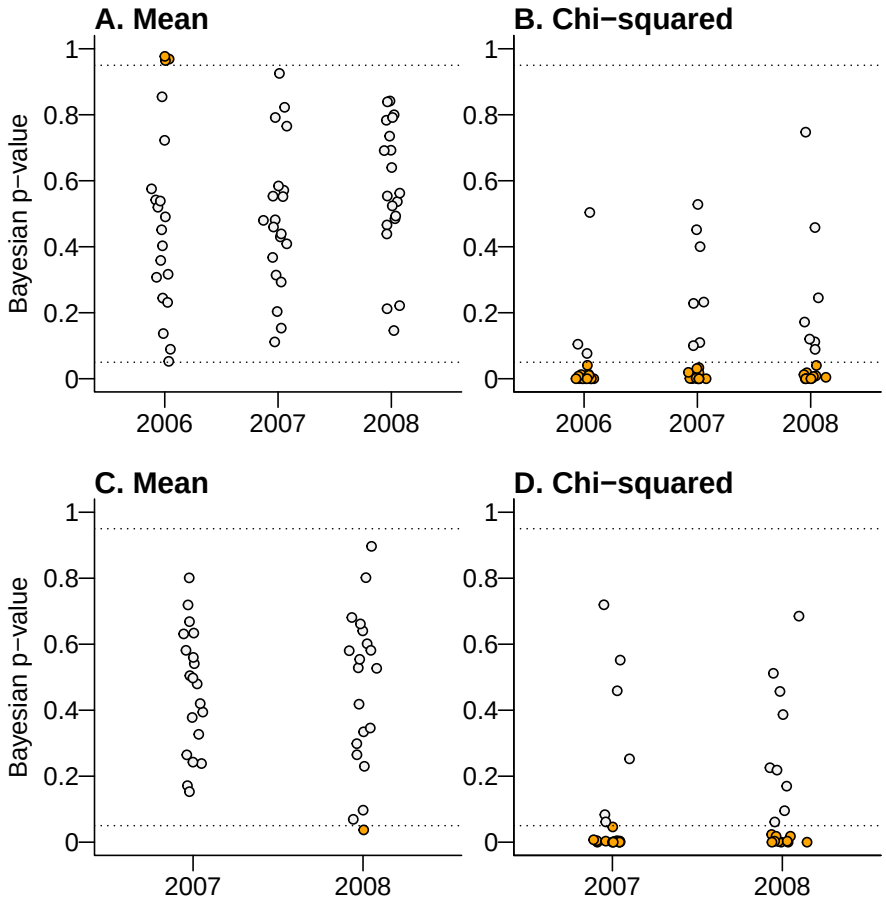


Figure 27: Bayesian p-values for the mean and $\chi^2$ value for observations of seeds staining viable in the viability trials. (A and B) Bayesian p-values for observations for age 1 seeds (seeds that were tested one year after being buried). (C and D) Bayesian p-values for observations for age 2 seeds (seeds that were tested two years after being buried). In all panels, each point corresponds to the p-value in one year for one of the 20 populations. Horizontal dotted lines indicate the 0.05 and 0.95 thresholds for model fit based on the test statistic. Points in light gray indicate a test statistic consistent with good model fit; points in orange indicate a test statistic consistent with poor model fit.

## Literature Cited

Conn, P. B., D. S. Johnson, P. J. Williams, S. R. Melin, and M. B. Hooten. 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs, **88**:526–542.

Hobbs, N. T. 2015. Bayesian models: a statistical primer for ecologists. Princeton University Press, Princeton, New Jersey.