1 Model adequacy

The in-sample model comparisons are for determining their relative adequacy, or a model's ability to represent the data it was fit to. Comparison between the posterior predictive estimates of in-sample AUC for each of the four models demonstrates that, overall, all of the models have approximately equal in-sample performance (Fig. 1). The parameter rich "past and vary" model has the greatest median in-sample AUC when compared to the other three models, but there is substantial overlap in their posterior distributions. Additionally, while our parameter rich "past and vary" model is possibly the most adequately performing model, the difference or improvement to performance is minimal at best – all four models have approximately equal in-sample AUC posterior distributions. All of the in-sample AUC estimates from our models are concentrated around an AUC of 0.77 which is interpreted as "fine but not good" performance. It is then hard to conclude that there is one "best" model which we can rely upon.

When the posterior predictive distributions of the in-sample AUC estimates are presented over time, the similarity in adequacy between the models becomes more apparent (Fig. 2). There are few major or obvious differences in model adequacy between the four models.

When the posterior predictive distributions of the in-sample AUC estimates are presented by taxonomic group, some heterogeneity in model adequacy is revealed (Fig. 3). While in all cases the model with the highest average in-sample AUC is the parameter-rich "past and vary" model, the amount of difference between the models varies by taxonomic group in ways not observable from the pooled estimates (Fig. 1). For example, the difference between the "past and vary" model and the others is more pronouced for Calcareous nannoplankton and Dinoflagellates, and smaller for the Foraminifera and Radiolaria.

Finally, when the posterior predictive distributions of the in-sample AUC estimates are presented over time and by taxonomic group, the reasons for the differences in model adequacy between the taxonomic groups becomes more apparent (Fig. 4), in particular the marginally better performance of the "past and vary" model over the other three.

For many taxon/model combinations there are one or more time periods

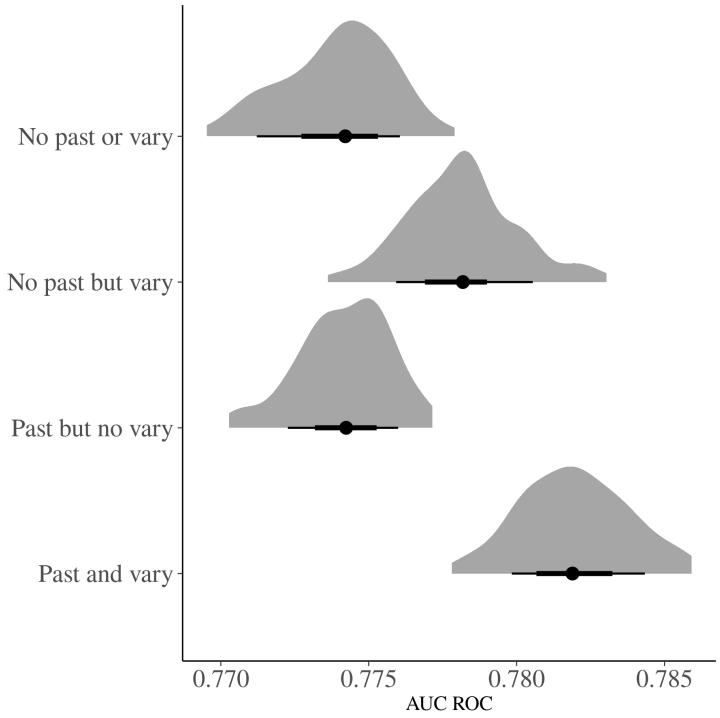

Figure 1: Posterior predictive AUC estimates for each of the four models being compared. These estimates are calculated from each of the models posterior predictive distribution compared to the empirical values. Models with a higher AUC values indicate better performance over models with lower AUC values. AUC is bounded between 0.5 and 1.

where posterior predictive in-sample AUC has a median value less than or equal to 0.5 – AUC value of 0.5 indicates that the model's predictions are no better than random (Fig. 4). However, this pattern is absent for the posterior predictive distribution of Foraminfera and Radiolaria for the "past and vary" model. Additionally, these periods of low model performance are rarer for the posterior predictive distribution of the "past and vary" model for calcareous nannoplankton and Dinoflagellates when compared to the other three models.

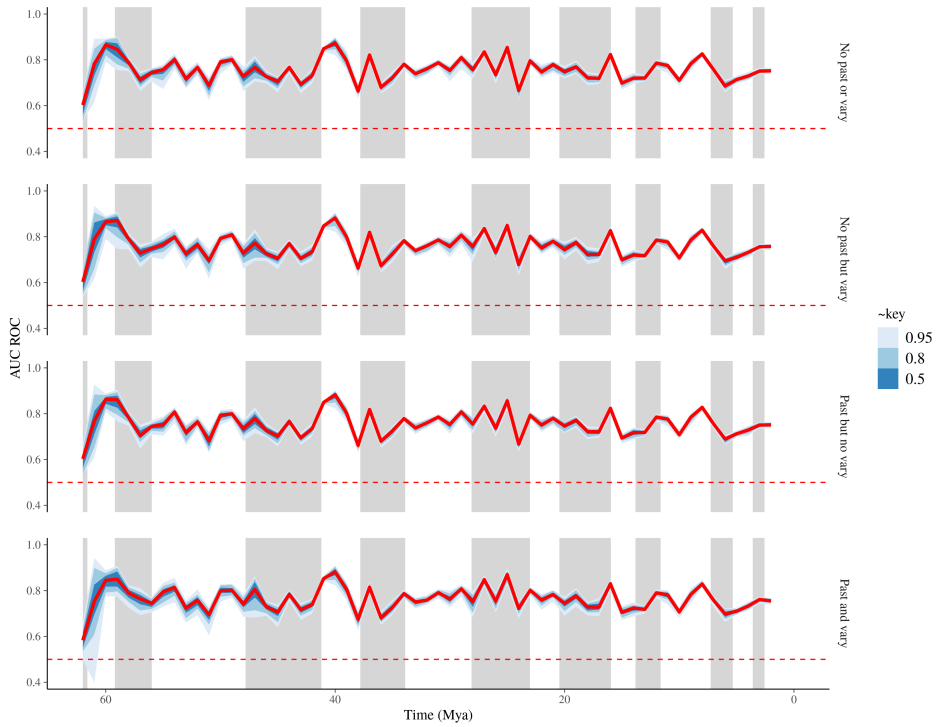

Figure 2: Comparison between the posterior predictive AUC estimates for each of the time intervals for each of the four models. These estimates are reflections of each model's fit to the various time intervals. The red line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

2 Cross-validation

Expected out-of-sample predictive performance was estimated using five-fold cross-validation, modified for time series data CITATION. This procedure yields four posterior (predictive) distributions, each corresponding to AUC values calculated from model-based predictions compared to the extinction state of the hold-out data. These four posterior predictive distributions are pooled to yield a posterior predictive distribution of expected out-of-sample performance – the resulting distributions tend to be very multimodal due to their very nature being fit to and estimated from different data sets and

amounts of data CITATION. Additionally, multimodality increases with model complexity (Fig. 5) – this makes sense as the more complex models allow for predictor effects to vary with time, allowing for a greater range in possible parameter values which in turn yield a greater range of posterior predictions.

Comparison between the posterior predictive distributions of expected out-of-sample AUC (Fig. 5) reveals a similar range in plausible values for all models as the in-sample AUC posterior predictive distributions (Fig. 1). Interestingly, the differences between the posterior predictive distributions for the models have decreased. For example, the "past and vary" model not clearly better than either of the "no past or vary" and "past and no vary" models (Fig. 5), which were shown earlier to be obviously worse-performing models based on in-sample performance (Fig. 1). These differences means that the rank order of median out-of-sample AUC is different from the rank order of median in-sample AUC. However, the shapes of the posterior distributions means interepting from the median values is incorrect – the models are effectively indistinguishable in their expected out-of-sample AUC values.

Additionally, the quality of expected out-of-sample performance is not great, with average out-of-sample AUC for each of our models estimated to be between 0.7 and 0.8 which is far from perfect. This result means that we would expect to correctly rank two species in order of most to least likely to go extinct 70-80% of the time. However, this expected out-of-sample performance is approximately the same as the in-sample performance results (Fig. 1), indicating that our models would yield consistent results when generalized to future extinctions.

When the posterior predictive distribution of expected out-of-sample AUC is presented as a time series, the similarity between the models is even more apparent (Fig. 6). While the width of the credible intervals at various time points varies between the models, the overall picture of expected out-of-sample AUC is almost identical when you compare the models – periods of relatively better or worse performance map identically between the time series.

We can also compare expected out-of-sample AUC by taxonomic group for each of the models (Fig. 7). These comparisons reveal a lot about the differences in predictive potential of the taxonomic groups. For example, the posterior predictive distributions of out-of-sample AUC for Foraminifera from all four models are approximately identical. In contrast, expected out-of-

sample AUC for Radiolaria exhibits the same or similar pattern in relative model performance to the pooled comparisons (Fig. 5). Additionally, we can state that our out-of-sample predictions for calcareous nannoplankton and dinoflagellates are not necessarily as precise as our estimates for Foraminifera or even Radiolaria. These results indicate that out-of-sample predictions may be easier for some taxonomic groups than others (e.g. Foraminifera versus Dinoflagellates.

Finally, we can present the posterior predictive distribution of expected out-of-sample AUC over time and taxonomic group for each of the four models (Fig. 8). For each taxonomic group, the time-series of posterior predictive values for each model are broadly congruent.

In the analysis of the posterior predictive distributions of the in-sample AUC values for the four models, we noted that there were time intervals where the models' predictions were no better than random (Fig. 4). This occurrence is generally much rarer for the posterior predictive distribution of out-of-sample AUC values – the major exception to this is Dinoflagellates, which for all four models has at least one time interval where the median the AUC of out-of-sample data were no better random. In contrast, the only other group for which median posterior predictive estimate of out-of-sample AUC reaches 0.5 is calcareous nannoplankton, and then only with the "no past or vary" model.

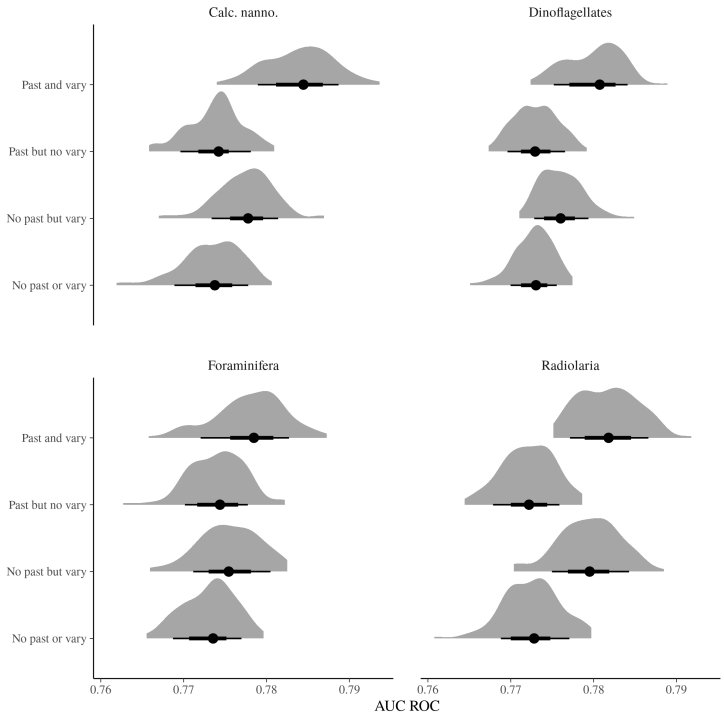

Figure 3: Comparison of posterior predictive AUC estimates for each of the four models, arranged by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups present in this analysis. The densities reflect the posterior distribution of the estimates, and below each density is marked the median AUC value along with the 50% and 80% credible intervals. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

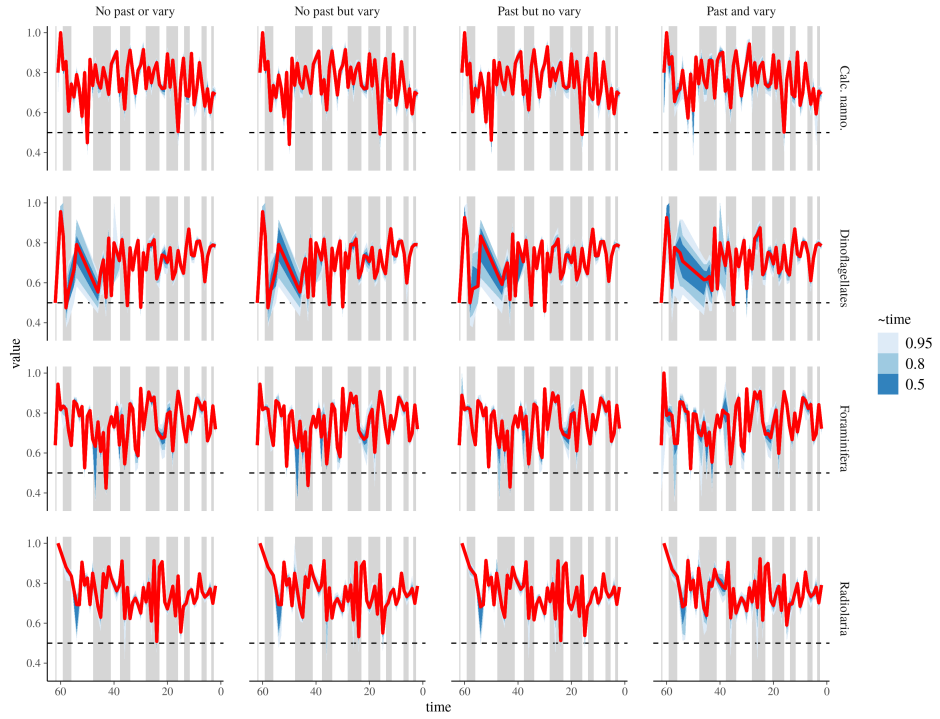

Figure 4: Comparison of posterior predictive AUC estimates for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups over time. The red line corresponds to the median AUC value, while the envelopes correspond to multiple credible intervals as indicated in the legend. In all cases, higher AUC values indicate greater predictive performance versus lower AUC values.

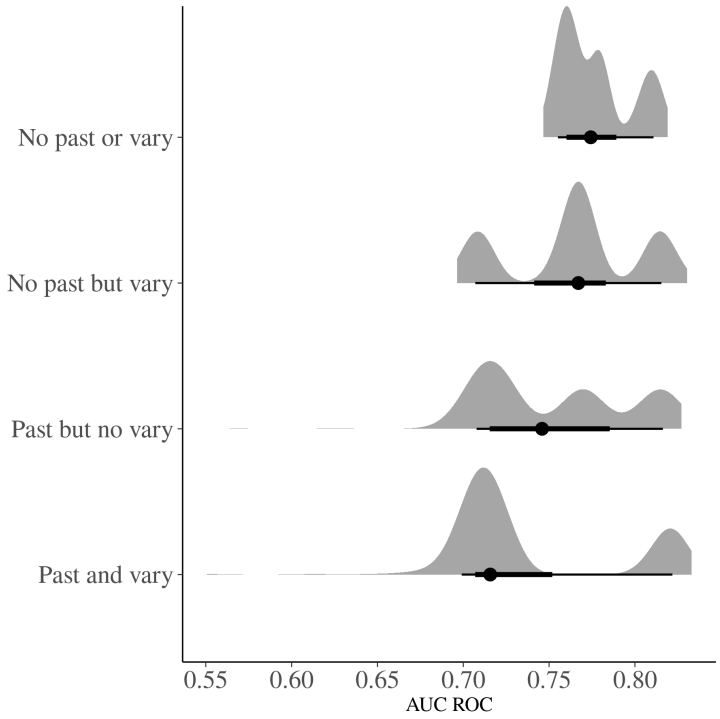

Figure 5: Results from our five-fold cross-validation of the time-series. Each labeled distribution of AUC values correspond to expected out-of-sample performance as estimated from that fold. Each fold represents a section of data being predicted from a model fit to all data before the start of that fold. Given that there are only five folds, performance is measured from predictions for four of the folds.

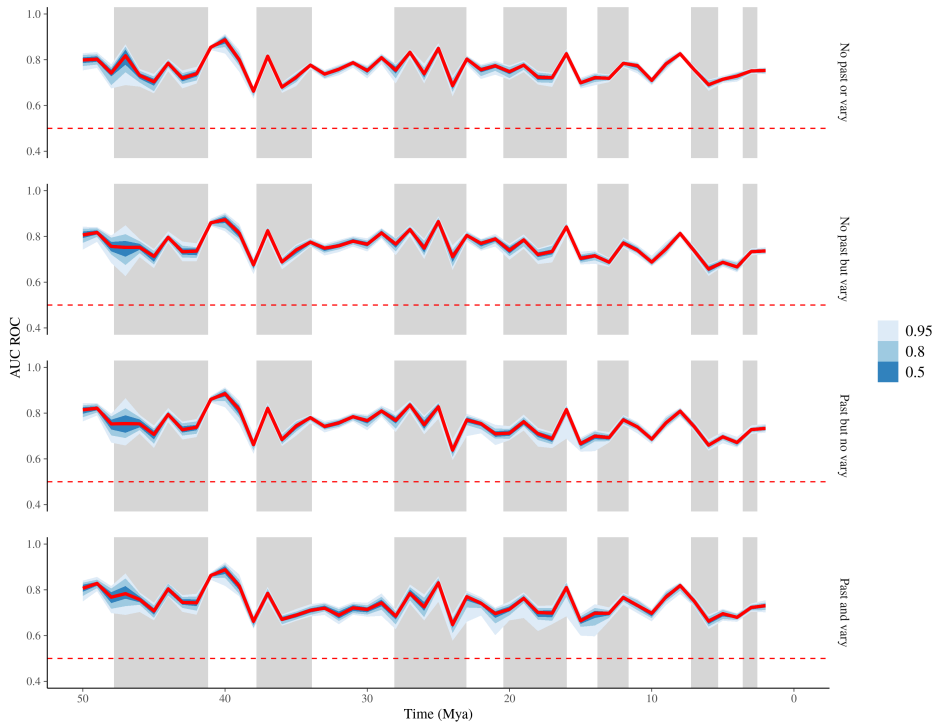

Figure 6: Comparison of out-of-sample AUC values calculated for each of the My intervals for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds (Fig. 5) into the predictions made for each of the million-year intervals. The red line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend.

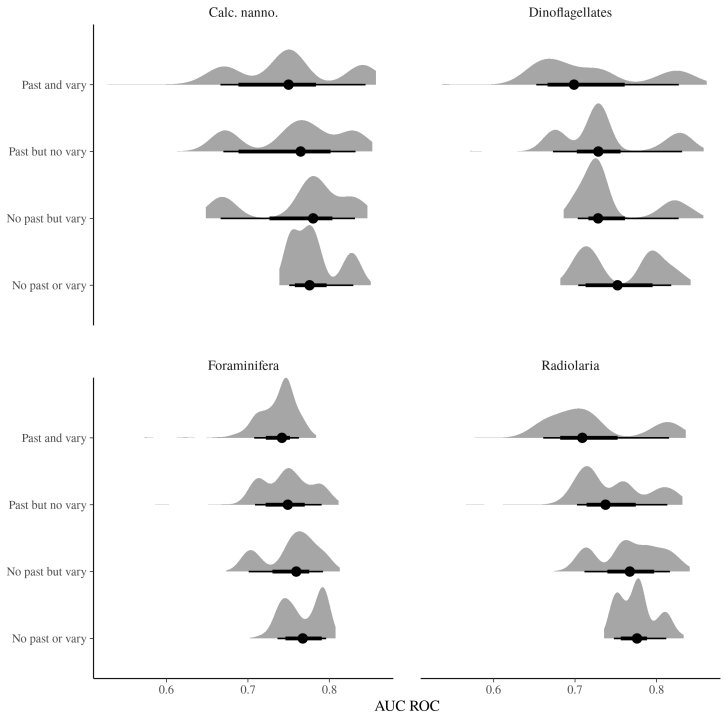

Figure 7: Comparison of out-of-sample AUC values aggregated by taxonomic group for each of the four models. Depicted for each taxon-model combination is an aggregate density of all posterior predictive estimates for each of four folds – cross-validation estimates are commonly multi-modal as each fold presents its own challenges for prediction. Beneath these densities is marked the median estimate along with 50% and 80% credible intervals.

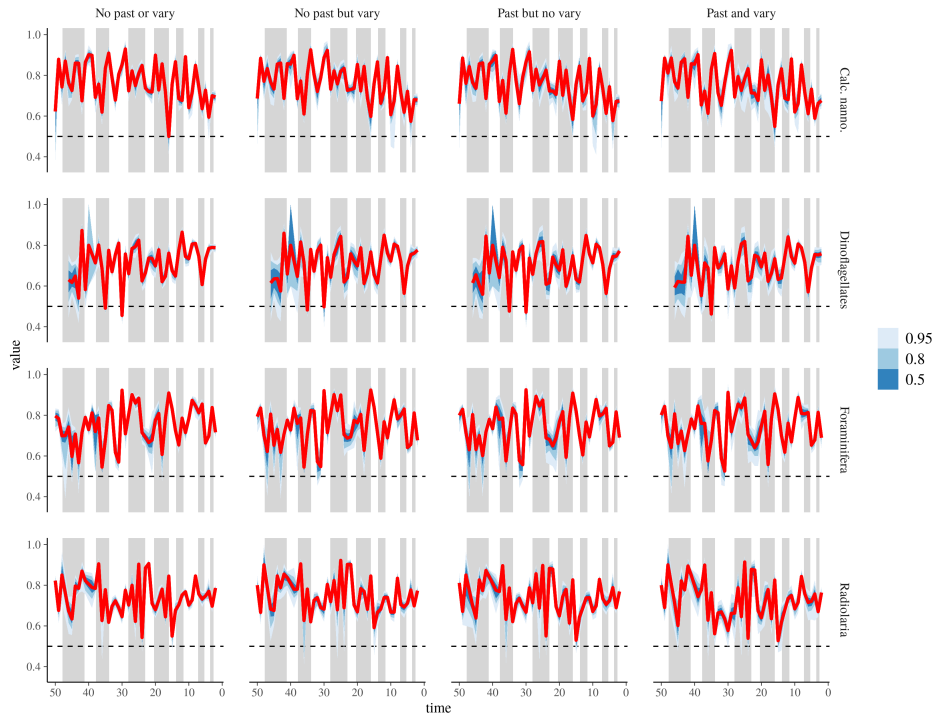

Figure 8: Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million year folds by each of the taxonomic groups (Fig. 7) into the predictions made for each of the million-year intervals. The red line corresponds to the median AUC estimate, with the envelopes corresponding to multiple credible intervals as indicated in the legend.