1 Model adequacy

Our first model comparison is based on in-sample AUC estiamtes. These estimates describe the relative and absolute adequacy of our models to describe the data to which they were fit. Comparison of the in-sample AUC estimates between the four models reveals similar results as the selection criteria (Fig. 1). While the parameter rich "past and vary" model has the greatest mean AUC when compared to the other three models, there is substantial overlap in their posterior distributions. Additionally, all of our models are concentrated around an AUC of 0.77 which is interpreted as "fine but not good" performance. It is hard, then, to conclude that there is one "best" model when all of them are approximately equivalent. However, our mode complex models have relatively higher performance than our simplest models – in particular, allowing covariate effects to vary over time appears to improve in-sample model performance as much as including the historical covariates.

Additional comparison of in-sample model performance at each of the time intervals reveals just how similar in adequacy our models performance are to each other (Fig. 2). There are few if any obvious differences between the models in their performances.

We can also compare the by-taxon in-sample performances of the four models.

Finally, we can also examine how the by-taxon in-sample performances of the four models vary over time.

2 Cross-validation

The approximate out-of-sample predictive performance for each of our four models, as measured by AUC, are approximately identical to each other and to our in-sample performance (Fig. 5). The quality of performance, however, is not great, with average out-of-sample AUC for each of our models estimated to be about 0.77 which is far from perfect. This result means that while we expect our model to yield consistent results when provided with new data, we do not expect that our predictions will be very accurate.

When our cross-validation estimates are presented for each time interval, the

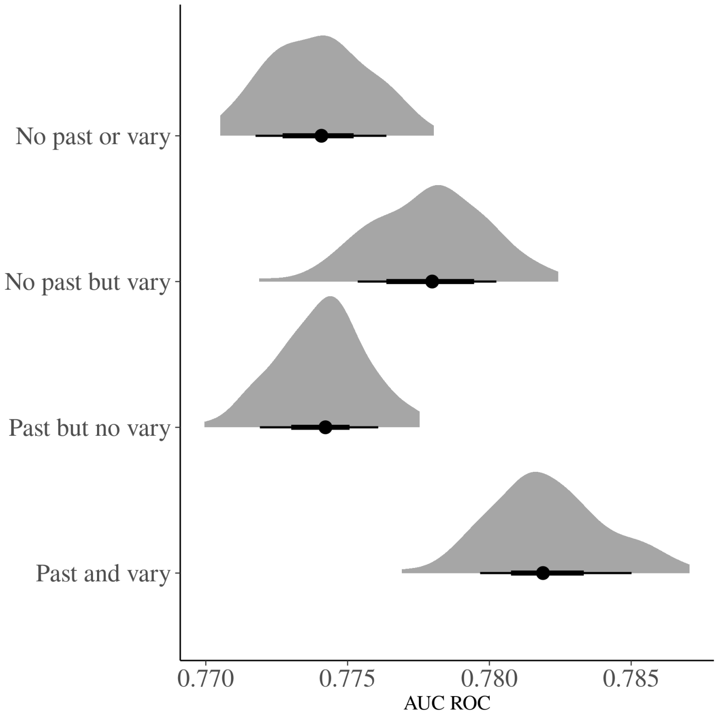

Figure 1: Posterior predictive AUC estimates for each of the four models being compared. These estimates are calculated from each of the models posterior predictive distribution compared to the empirical values. Models with a higher AUC values indicate better performance over models with lower AUC values.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

similarity in out-of-sample performance between the models becomes even more obvious – while there are minor differences in the posterior estimates for individual time points, the overall structures are nearly identical.

We can also compare expected out-of-sample performance based on taxonomic group

Finally, expected out-of-sample performance is compared between taxonomic groups over time.

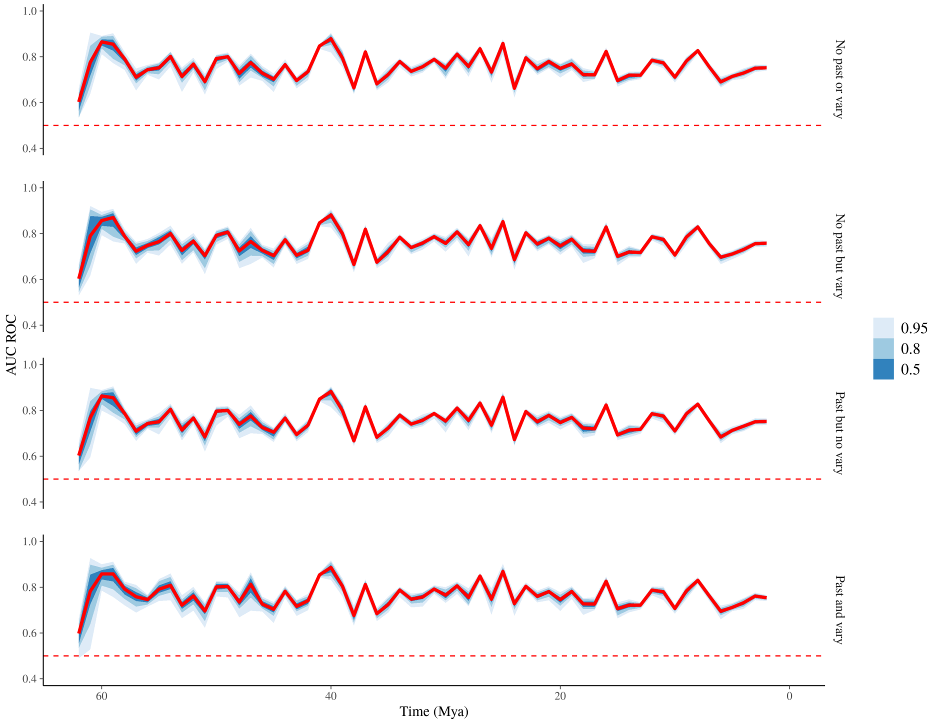

Figure 2: Comparison between the posterior predictive AUC estimates for each of the time intervals for each of the four models. These estimates are reflections of each model's fit to the various time intervals. The red line corresponds to the median AUC value, while the envelopes correspond to multiple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

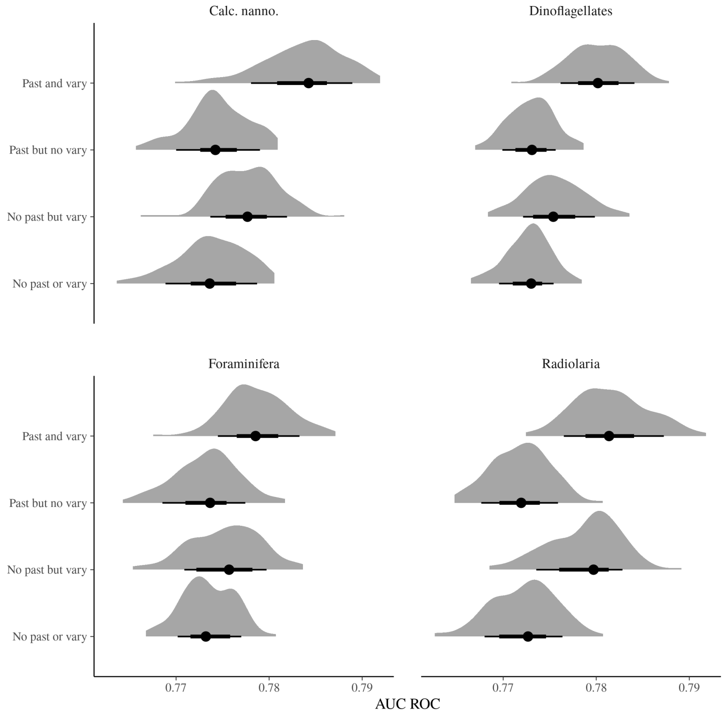

Figure 3: Comparison of posterior predictive AUC estimates for each of the four models, arranged by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups present in this analysis. The densities reflect the posterior distribution of the estimates, and below each density 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.

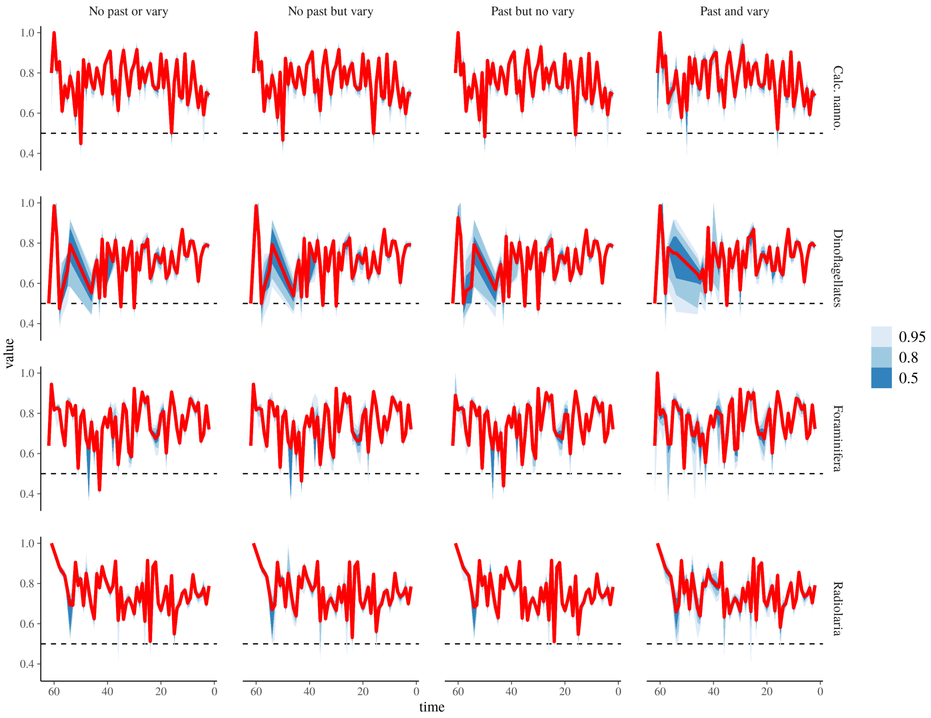

Figure 4: Comparison of posterior predictive AUC estimates for each of the four models, arranged over time and by taxonomic group. These estimates reflect each model's fit to the various taxonomic groups over time. The red line corresponds to the median AUC value, while the envelopes correspond to multiple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

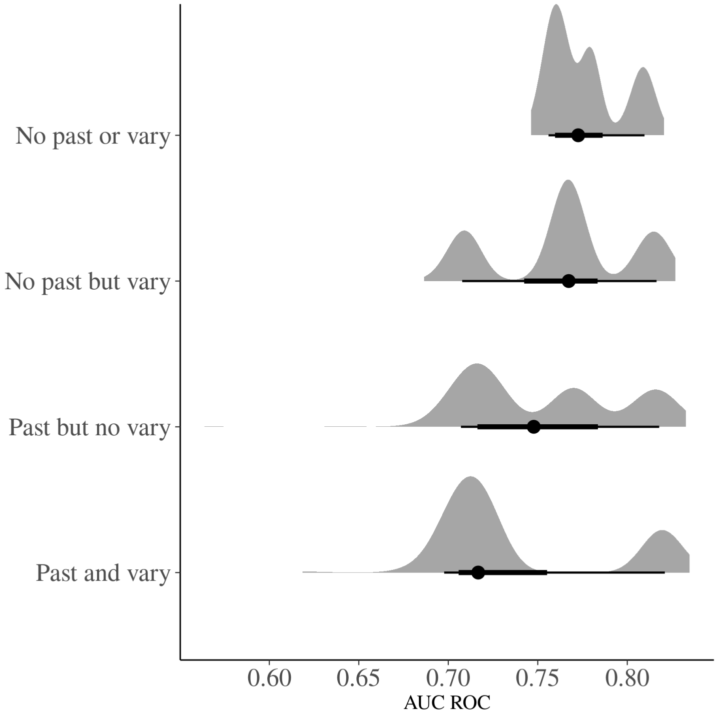

Figure 5: Results from our five-fold cross-validation of the time-series. Each labeled distribution of AUC values correspond to expected out-of-sample performance as estimated from that fold. Each fold represents a section of data being predicted from a model fit to all data before the start of that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

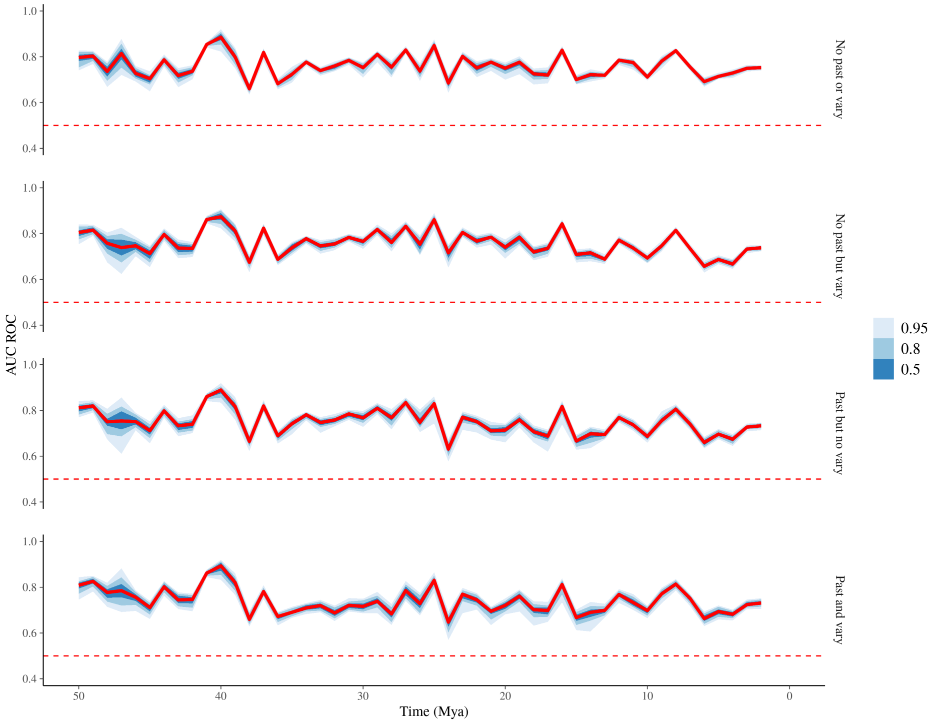

Figure 6: Comparison of out-of-sample AUC values calculated for each of the My intervals for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 12-million 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.

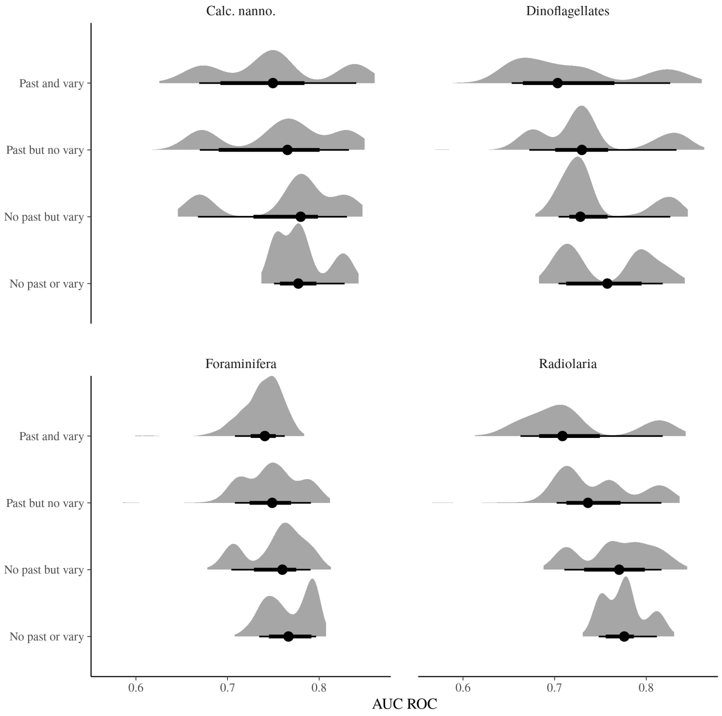

Figure 7: Comparison of out-of-sample AUC values aggregated by taxonomic group for each of the four models. Depicted for each taxon-model combination is an aggregate density of all posterior predictive estimates for each of four folds – cross-validation estimates are commonly multi-modal as each fold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

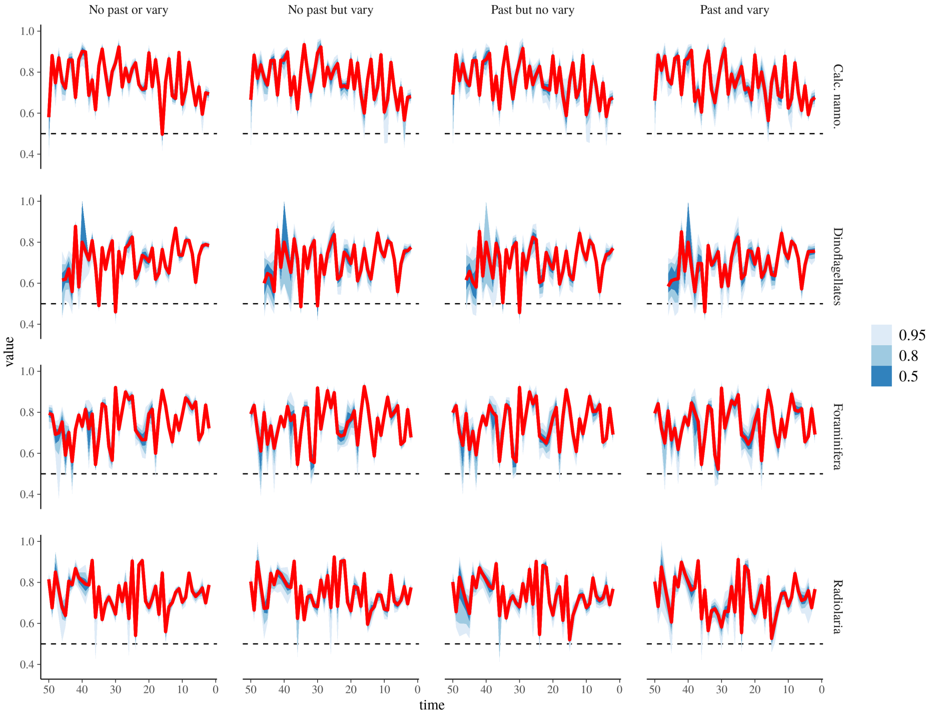

Figure 8: Comparison of out-of-sample AUC values over time as aggregated by taxonomic group for each of the four models. The AUC of the individual My intervals within each fold is plotted to highlight the heterogentity in performance within and between folds. This presentation decomposes each of the 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.