B Results of Eucalypt Study in More Detail

In Section 4.1 we examined the assumption of a proportional sampling bias, and discussed how to check this assumption based on the data. We checked whether distance-to-coast, an important bias covariate, had a *species-specific* effect on the sampling bias for the individual species *E. punctata*. The data suggested there was an effect. Figure 10 shows the analogous fitted curve for the 35 other species in the data set. As we see, many of these species exhibit an effect that is either very nearly or not quite distinguishable from no effect.

We also computed cross-validated predictive performance on log-likelihood and AUC in Figures 11– 12. Some of the rarest species only appeared on one or two of the geographic blocks, so we exclude them from the cross-validation results.

For the cross-validated models, the R formula used to model the species distribution is

```
~predict(bc04.basis, newx = bc04)
+ predict(rsea.basis, newx = rsea)
+ predict(bc33.basis, newx = bc33)
+ predict(bc12.basis, newx = bc12)
+ predict(rjja.basis, newx = rjja)
+ bc02 + bc05 + bc14 + bc21
+ bc32 + mvbf + rugg + factor(subs)
+ twmd + twmx
```

Descriptions of the environmental variables can be found in Appendix C. Terms like `predict(var.basis, newx=var)` appear when we have used a customized natural spline basis for the variable `var`.

The R formula used to model the bias is

```
~predict(survey.bg.basis,newx=survey.bg)
+ predict(ld2coast.basis,newx=ld2coast)
+ predict(alongCoast.basis,newx=alongCoast)
+ predict(ld2r.basis,newx=ld2r)
+ predict(d2t.basis,newx=d2t)
+ predict(rugg.basis,newx=rugg)
+ xveg
```

The variable `survey.bg` is a geographic variable corresponding to the logarithm of how many presence-absence survey sites are located in a grid cell. It is meant to proxy for the log-frequency of ecologist visits to locations near a site $s \in \mathcal{D}$. Note that the locations of presence-absence survey sites are not modeled as random in our model, so we are not using the response variable twice by doing this.

The variable `ld2coast` is the logarithm of 1 km plus the distance to coast, and `ld2r` is the logarithm of 1m plus the distance to the nearest road. `d2t` is the distance to the nearest town. `alongCoast` is a projection of geographic location in a direction running parallel to the coast; it is largest in the northeastern part of the study region and smallest in the southwestern part. `rugg` is ruggedness of terrain, and `xveg` is a binary indicator of whether a location has extant vegetation (e.g., it would be 1 in a forest and 0 in a wheat field).

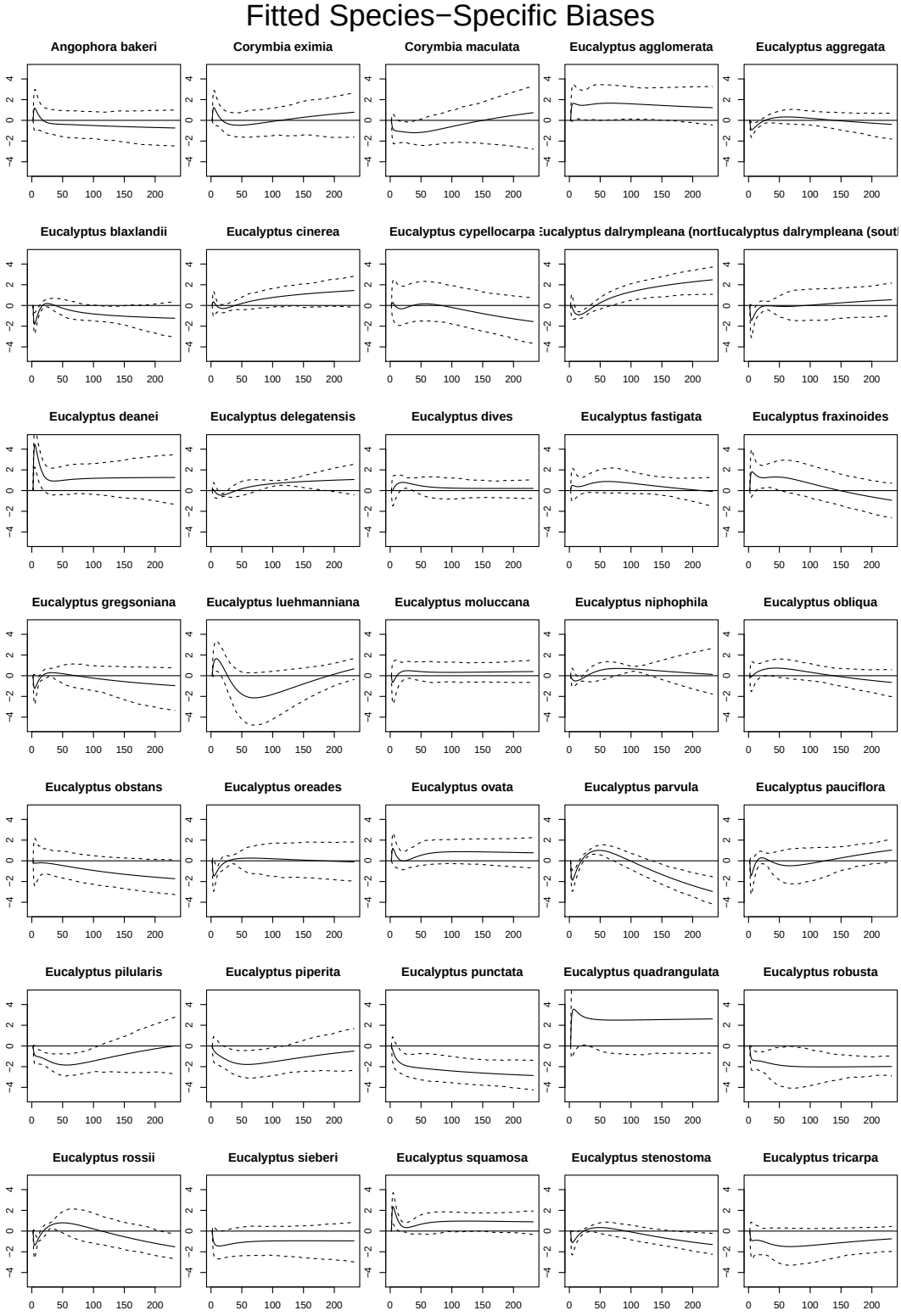

Figure 10: Bootstrap confidence intervals for the species-specific effect of distance-to-coast on log-sampling bias, for each of the 35 species other than *Eucalyptus parramattensis*, whose data set is too small for the block bootstrap. Some of the confidence intervals exclude zero for a significant effect.

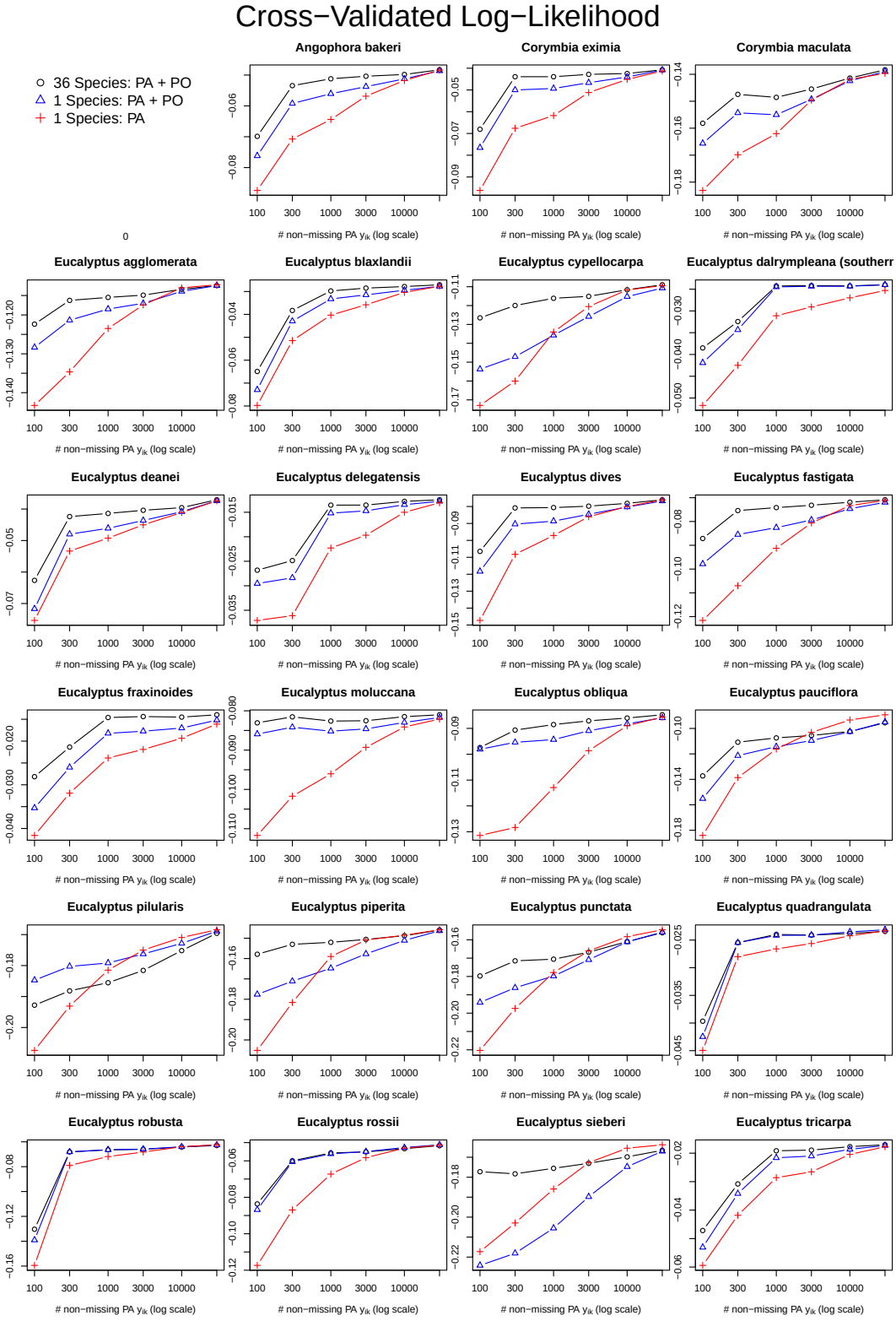

Figure 11: Cross-validation results for all species that were observed in at least 110 different presence-absence sites. Results vary for different species and different methods, but the method that pooled data across all 36 species had consistently superior performance.

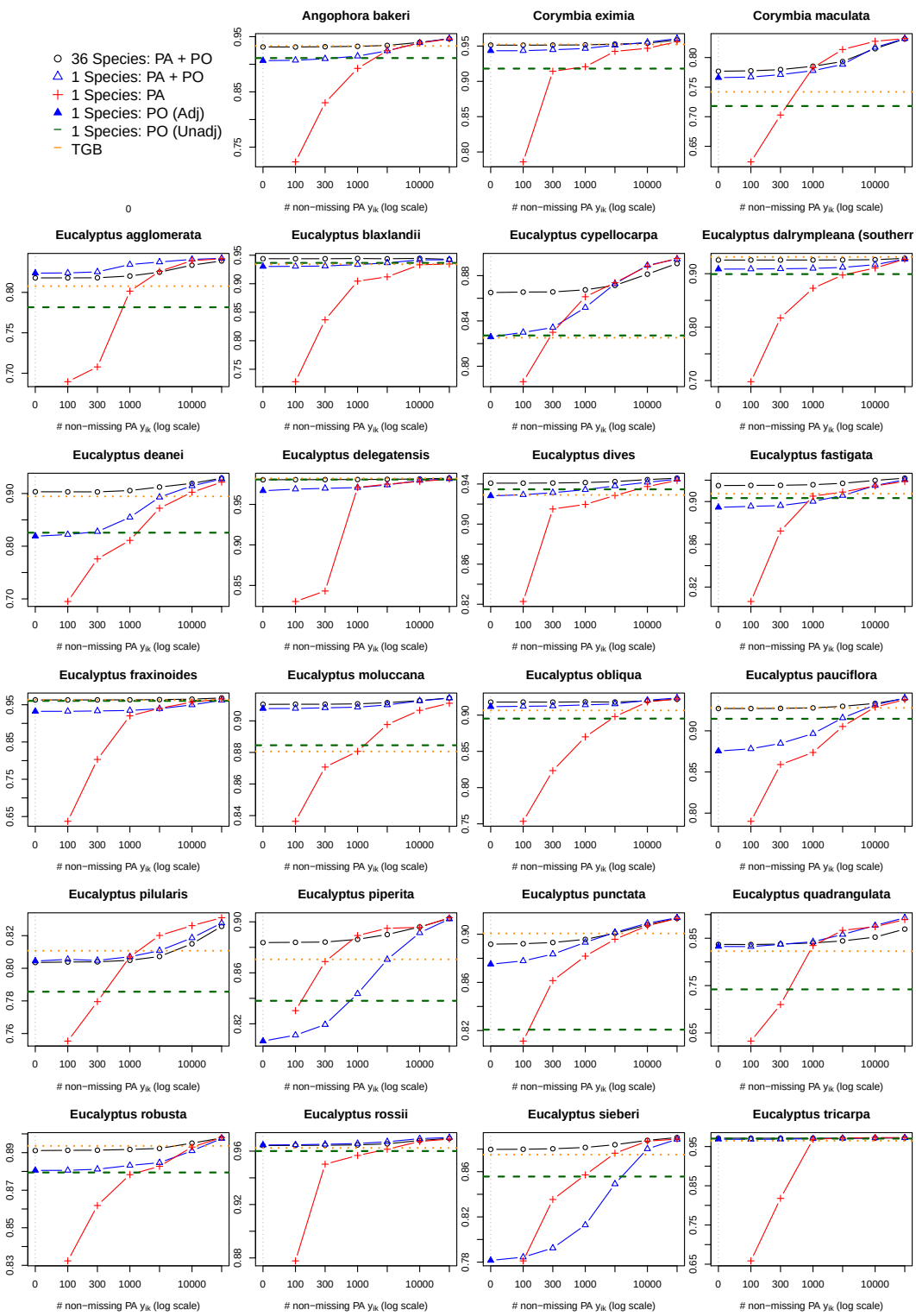

Figure 12: Cross-validation results for all species that were observed in at least 110 different presence-absence sites. Results vary for different species and different methods, but the method that pooled data across all 36 species had consistently superior performance.