A Plain Language Guide to Model Validation in REMA

Laurinne J Balstad[1,2], Jane Sullivan[3], and Cole Monnahan[3]

September 2024
Joint Groundfish Plan Team

[1]*Department of Environmental Science and Policy, University of California, Davis*
[2]*Center for Population Biology, University of California, Davis*
[3]*Alaska Fisheries Science Center, National Marine Fisheries Service*

Background

Fisheries stock assessments are moving towards state-space estimation, which boasts a range of benefits, including separation and estimation of observation and process error, a more elegant framework for handling missing data, a high degree of flexibility with respect to model architecture and inclusion of different data types, and the potential for improved projections and predictive skill (Aeberhard et al., 2018). The flexibility and relative ease of fitting state-space models means they can increase in complexity and dimensionality rapidly. While easy to fit, state-space models are often challenging to validate, and even simple models can suffer from estimation issues that result in biased inference (Auger-Méthé et al., 2016).

At the North Pacific Fishery Management Council (NPFMC), the "random effects model" (REMA) is by far the most common state-space model used for fishery management (Sullivan et al., 2022). REMA is a state-space random walk model that can be customized to estimate multiple process errors, fit to an additional abundance index, or estimate additional observation error. REMA is used to estimate biomass within Tier 5 groundfish and Tier 4 crab stock assessments and to apportion Acceptable Biological Catches (ABCs) by management area for many stocks. Despite the high impact of this model within the North Pacific fishery management process, we have no standard model validation practices for REMA.

Our goals:

1. Apply established state-space model validation techniques to real life REMA examples.
2. Create a REMA model validation guide with clear descriptions of each method and motivation for its use, an explanation of how to interpret the model validation results, and reproducible code that will run for all operational REMA models at the NPFMC.
3. Provide preliminary recommendations for REMA users and reviewers on which model validation methods are most relevant and informative for REMA.

A living version of this document is available on the REMA website: https://afsc-assessments.github.io/rema/articles/ex4_model_validation.html. This website is code-enhanced with reproducible code to support model validation for all operational REMA models at the NPFMC.

Interested but otherwise busy readers are encouraged to skip to the "When should we care about failed model diagnostics?" and "Preliminary recommendations" sections. Table 1 and Figures 6-8 show examples of a REMA model failing model validation criteria.

A testing framework for REMA model validation

In this paper, we will be testing that (1) REMA is working as expected (i.e., code has been implemented correctly and parameters are estimated without bias), and (2) that the assumptions made when parameterizing and estimating REMA are valid, including assumptions related to random effects and error structure. We use two example stocks:

- Aleutian Islands Pacific cod (AI Pcod; Spies et al., 2023): a simple, univariate case, which fits to a single time series of the AI bottom trawl survey and estimates one process error
- Gulf of Alaska Thornyhead rockfish (GOA Thornyhead; Echave et al., 2022): a complex, multi-strata and multi-survey case, which estimates multiple process errors, a scaling parameter for the longline survey, and two additional observation errors for the GOA bottom trawl and longline surveys

Both models appear to converge per standard REMA and TMB diagnostics (small maximum gradient, invertible Hessian), and fit the data reasonably, with model fits shown in Figure 1. These stocks span the levels of complexity of NPFMC REMA models; testing REMA across this range of complexity helps ensure that model and model assumptions are valid in a variety of realistic, management-relevant cases.

It is important to note that model validation does not mean the model is "more right" or "more correct." Model validation does not help an analyst with model selection (except to identify models that might not function properly), nor does it ensure that the model makes "better" predictions. Rather, model validation is a way to ensure that the model is operating as expected, without introducing bias or violating statistical assumptions upon which the model is based (Auger-Méthé et al., 2016; Auger-Méthé et al., 2021), and that the data reasonably could have come from the model (Thygesen et al., 2017).

The key questions we aim to answer with model validation include the following:

1. Does the model perform as expected, or does it introduce bias? Using a simulation self-check, we will test if the model is coded correctly and is able to recover known parameters.
2. Is it plausible that our data could have been generated by the model? Using one-step ahead (OSA) residuals, the appropriate residual type for state-space models, we test the model assumptions and look for trends in residuals that reveal characteristics or dynamics of the data that aren't adequately captured by the model.
3. Are the normality assumptions made when estimating random effects via the Laplace approximation accurate? By comparing the posterior distribution of fixed effects with and without the Laplace approximation, we test whether the distribution of the random effects are normal and thus the accuracy of the Laplace approximation. This allows us to test the accuracy of the fixed effects estimates with uncertainties.

4. Are the parameters unique and non-redundant? Checking the correlation between parameters helps us identify if parameters are redundant with each other.
5. Is the model converged?

1. Simulation self-test: Can we recover parameters without bias?

Simulation testing ensures the model has been properly coded and performs consistently without introducing bias (Auger-Méthé et al., 2021; Gimenez et al., 2004). First, we use the REMA model to estimate parameters (e.g., process error variance) from the real data. Next, we use the estimated parameters as "true values" to simulate new latent states (random effects) and data conditioned on the states using the REMA equations. We then use the REMA model to re-estimate the model parameters ("recovered parameters") and calculate the relative error (RE; i.e., (true-estimated values)/true value*100)) for the parameters in each simulation replicate (N=500). The AI Pcod model is fitted to bottom trawl data and each simulated data set is a biomass trajectory; the GOA Thornyhead model is fitted to bottom trawl and longline survey data, and each simulation data set includes biomass trajectories in nine strata and longline survey relative population weights (RPWs) in three strata.

Models fail this simulation testing when the recovered parameters are biased (RE $\neq$ 0), or deviate consistently from the true values used to simulate data (span of the recovered parameters is large): this can indicate that the model has a coding error, is non-identifiable, has redundant parameters, or is biased (i.e., there is some other kind of model misspecification).

Results of the simulation self-test

Only 1 out of 500 simulation replicates did not converge for the AI Pcod model, which suggests a high degree of model stability. There were 11 out 500 simulation replicates for the GOA Thornyhead model that did not converge. In Figure 2 the top row within each panel shows the distribution of resulting recovered parameter estimates in each of the models based on the simulations, where the horizontal line is the median estimated value and the black dot is the true value (i.e., the parameters estimates from the real data sets).

As demonstrated in the boxplots of the bottom rows of Figure 2, the parameters in both models are recovered well in simulation testing, with low median relative error. In both models there are instances of long negative tails in the log standard deviation of the process error (log_PE; Figure 2 top row within each panel), suggesting that PE was small or tended towards zero in some simulations (recall that parameters are estimated in log space, and a very negative number in log space is close to zero in natural space). This is also the case for the extra observation error for the biomass survey (log_tau_biomass) in the GOA Thornyhead model (Figure 2, top row within lower panel).

These outlier replicates, which occur when the optimizer used by rema (nlminb) returns a "NA/NaN function evaluation" error, help shed light on why some replicates failed to converge. Under the hood, the optimizer is pushing the PE towards zero (PE=0 is the same as taking a global mean of the biomass time series), violating the central assumption of the REMA model that the biomass has an underlying trend (PE > 0, and thus PE can be log-transformed). The negative log-scale values are most extreme for the GOA

Thornyhead model, which should raise some concern of potential model misspecification or over-parameterization. For example, given the trade-off between process and observation error in REMA (where including an additional observation dampens the biomass trend estimated, pushing PE towards zero), it is possible that the estimation of additional observation error on the biomass survey (log_tau_biomass) might not be justified.

2. Residual analysis

Residuals are used to test the underlying assumptions about the structure and error distributions in the model. For example, in a linear regression, residuals should be independent, normal, and have constant variance. Traditional residuals (e.g., Pearson's residuals) are inappropriate for state-space models like REMA, because the random effects induce correlations in the predicted data such that the residuals are no longer independent. Additionally, process error variance may be overestimated in cases where the model is mis-specified, thus leading to artificially small residuals (see Section 3 of Thygesen et al. 2017 for an example). The appropriate residual type for validation of state-space models are one-step ahead (OSA) residuals. Instead of comparing the observed and expected values at the same time step, OSA residuals use forecasted values based on all previous observations (i.e., they exclude the observed value at the current time step from prediction). In this way, OSA residuals account for non-normality and correlation in the residuals among years.

Under a correctly specified model, resultant OSA residuals should be independent and identically distributed (i.i.d.) with a standard normal distribution N(0,1). This can be tested with a normal QQ plot, where the theoretical quantile values of the standard normal distribution are plotted on the *x*-axis, and the corresponding empirical quantile values of the OSA residuals are plotted on the *y*-axis. If the OSA residuals are standard normally distributed, the points will fall on the 0/1 reference line, and the standard deviation of normalized residuals (SDNR) statistic will be close to 1. If the residuals are not standard normally distributed (SDNR far from 1), the points will deviate from the reference line. In the case where the model includes multiple strata and surveys, OSA residuals should be normally distributed within and across strata and surveys, however small sample sizes (e.g., within a single stratum) may not have sufficient statistical power to identify model misfit.

In addition to the normality of OSA residuals, residuals should be random and independent (i.e., they should not be correlated by year). Standard approaches such as autocorrelation function (ACF) plots or residual runs tests (e.g., Wald-Wolfowitz test) are inappropriate for many REMA applications because of missing years of data in the survey time series, and instead, we directly plot the residuals against the years of the survey time series. Visual patterns or extreme outliers (greater than 3 or less than -3 for N(0,1) distribution) in the residuals can indicate structure in the data (i.e., temporal correlation) that is not adequately captured by the model.

Methods for calculating OSA residuals in REMA have been implemented in the rema::get_osa_residuals() using the TMB::oneStepPredict() function and the fullGaussian method when using lognormal error distributions. The rema::get_osa_residuals() returns tidied dataframes of

observations, REMA model predictions, and OSA residuals for the biomass and CPUE survey data when appropriate, along with QQ and residual-time series diagnostic plots.

Results of the residual analysis

The normal QQ plot for AI Pcod (Figure 3, top panel) suggests slight negative skewness in the residuals (i.e., the majority of the points fall below the 0/1 line), though the small sample size makes interpretation challenging. The SDNR is 0.99 (recall a perfect model would have an SDNR=1.00), indicating assumptions are likely met for the residuals. The plot of OSA residuals by year (Figure 3, bottom panel) shows no obvious patterns in the residuals, suggesting they are independent; there is no evidence of outliers in the residuals that warrant further inspection.

The combined normal QQ plot for GOA Thornyhead OSA residuals (Figure 4, top panel) suggests they follow a normal distribution (the points fall along the 0/1 line), though there is evidence of slight positive, or right skewness (the majority of the points fall above the 0/1 line). The SDNR is approximately 1, indicating the variance assumptions are likely met for the residuals. The QQ plots for the biomass by strata (Figure 4, middle panels) highlight where some of the positive skewness may be coming from (e.g., EGOA 701-1000m, WGOA 0-500 m); however, small sample sizes by stratum make interpretation of these QQ plots difficult. The QQ plots for the CPUE by strata (Figure 4, bottom panel) are mostly normal, though there is some evidence of light tails in the WGOA stratum.

The plot of OSA residuals by year for the GOA Thornyhead OSA residuals by biomass strata (Figure 5, top panels) show no patterns in the residuals, suggesting they are independent. However, there are runs in the residuals for the CPUE strata (Figure 5, bottom panels), especially in the CGOA and WGOA, which might indicate model misspecification or misfit. There is no evidence of outliers.

3. Laplace approximation: Are the model assumptions related to random effects estimation reasonable?

The *rema* library was developed in Template Model Builder (TMB), which uses maximum marginal likelihood estimation with the Laplace approximation to efficiently estimate high dimensional, non-linear mixed effects models in a frequentist framework (Skaug and Fournier, 2006; Kristensen et al., 2016). The primary assumption in models using the Laplace approximation is that the random effects follow a normal distribution. This assumption simplifies the complex integrals that make up the likelihood function. The Laplace approximation is fast and accurate when the normality assumption is met; however, if the true distribution of the random effects is not normal, the Laplace approximation may introduce bias into the fixed effect estimates.

We can test the validity of the Laplace approximation using Markov chain Monte Carlo (MCMC) sampling in the *tmbstan* library, comparing the distributions of the fixed effects parameters from MCMC-sampled models with and without the Laplace approximation (Monnahan and Kristensen, 2018). To do so, we compare the distributions using QQ plots between the two model cases. The Laplace approximation is reasonably accurate if the sampling quantiles for the two models (with and without the

Laplace approximation) are similar, i.e., they fall on the 1:1 line. This can also help us identify bias introduced by the Laplace approximation, for example, if the fixed effects estimates and their associated estimates of uncertainty differ between the base model (run with marginal maximum likelihood estimation using the Laplace approximation) and the MCMC versions with and without the Laplace approximation. MCMC models were run with 1000 iterations, assuming a warmup of 500. As discussed in the Results section, iterations were increased to 5000 with a warmup of 2500 for GOA Thornyhead in an effort to improve diagnostics and model convergence.

Results for testing the validity of the Laplace approximation

Figure 6 gives the compared quantiles from full MCMC testing (*x*-axis) and the Laplace approximation (*y*-axis) for AI Pcod and GOA Thornyhead models.

For the simpler AI Pcod model (Figure 6, left panel), the Laplace approximation and full MCMC testing show a similar distribution, indicating that the Laplace approximation is reasonable. Additionally the estimates of fixed effects are nearly identical between the base model and MCMC models with and without the Laplace approximation (Table 1).

For the more complex GOA Thornyhead model, the preliminary model results based on 1000 iterations exhibited poor diagnostics for all process error parameters and the additional observation error for the biomass survey (log_tau_biomass). We increased the number of MCMC iterations to 5000 in an effort to improve convergence diagnostics; however, the log_tau_biomass parameter continued to show significant deviations between the models with and without the Laplace approximation (Figure 6, right panels). A comparison of the fixed effects estimates between the base models show large differences in the parameter estimates and their associated estimates of uncertainty (Table 1). This indicates that the Laplace approximation is likely inaccurate, and that further investigation into model structure might be necessary.

4. Parameter estimation reliability: Are the model parameters identifiable and non-redundant?

Parameter redundancy refers to the idea that multiple parameters contribute to the model in the same way. An intuitive case is $y \sim \beta1+\beta2+\alpha x$, since the model could estimate many combinations of $\beta1$ and $\beta2$ which minimize the log-likelihood, i.e., the sampled parameters will be correlated. To reduce redundancy, $\beta1+\beta2$ can be redefined as $\beta0$. In more complex, hierarchical models, parameter redundancy is not always intuitive and can be solved by increasing the number of parameters or reparameterizing the model entirely (Gimenez et al., 2004; Cole, 2019).

Using the MCMC sampling framework, we can check for parameter correlations to help us identify possible redundancy in parameters using the *bayesplot* library (Gabry et al., 2019; Gabry and Mahr, 2024). If the sampled parameters are identifiable and non-redundant, we would see no correlation between parameters. For simplicity here, we only use the model case with the Laplace approximation.

Note this is unnecessary for the AI Pcod model, since there is only one fixed effect parameter estimated in that model.

Results of parameter correlation analysis

Figure 7 shows a pairwise correlation matrix of the posterior draws for the GOA Thornyhead model (the MCMC model with the Laplace approximation), with histograms of the univariate marginal distributions on the diagonal and a scatterplot of the bivariate distributions off the diagonal. There are significant unresolved sampling problems of the log_tau_biomass parameter, shown by the non-normal and heavily skewed distribution of log_tau_biomass. The pairwise scatterplots on the off-diagonals show that the heavy left tails in log_tau_biomass are causing interactions (i.e., correlation) with almost all other parameters in the model. This diagnostic is a clear indication that there is model mis-specification that warrants further investigation.

5. Is the model really converged?

Despite the poor diagnostics shown in this vignette for the GOA Thornyhead model (particularly regarding the Laplace approximation, section 4), the standard output in TMB and *rema* suggests that this model is converged (i.e., it has a low maximum gradient component and the standard error estimates appear reasonable). Within the MCMC framework, we can look at the mixing of MCMC across multiple chains using a diagnostic called "Rhat" and a traceplot to assess convergence using the *rstan* library (Stan Development Team, 2024). Properly mixed (i.e., converged) models will have an Rhat close to 1 and exhibit traceplots that are centered around a single value and bounce between a finite span of values, looking like fuzzy caterpillars.

Results of MCMC convergence

From the traceplots for both models in Figure 8, we see that the AI Pcod model (top panel) MCMC chains are fully mixed and the model is converged. However, for the GOA Thornyhead model (bottom panel), we see that the MCMC sampling of additional observation error (log_tau_biomass) is not centered around a single value and diverges towards negative infinity. Additionally, the Rhats associated with this parameter are greater than 1, indicating a lack of convergence (Table 1). This indicates that the estimation of an additional observation error (log_tau_biomass) is causing convergence issues and the MCMC chains are not well mixed.

When should we care about failed model diagnostics?

The REMA model is used within the NPFMC to obtain exploitable biomass estimates for Tier 4 crab and Tier 5 groundfish and as a method to apportion Acceptable Biological Catches among management regions (Sullivan et al., 2022). So then, if its primary purpose is to smooth noisy survey biomass estimates (i.e., if we are using it as a fancy average), do we need to care about potential red flags raised during model validation?

- **Yes:** In the case of the GOA Thornyhead model, for example, it appears there is an issue with the estimation of the additional observation error, pointing to potential over-parameterization of the model and/or parameter redundancy. Because terminal year predictions from the REMA model are the quantities directly used for management, and the estimation of additional observation error by definition increases the "smoothness" of model predictions, the answer is likely yes in this particular case.
- **Maybe not:** In a hypothetical case where the estimation of year predictions shows no diagnostic concerns, but uncertainty estimations show possible diagnostic problems, model validation red flags might be less important. This is because uncertainty results from REMA are not generally used for fisheries management at the NPFMC. In other words, failed model validation criteria are primarily a concern when they impact the quantities of interest in the model (in our case, biomass predictions). Model validation, and subsequent interpretation, should be considered in light of the goal of the model, rather than as a binary pass/fail testing procedure.

Preliminary recommendations

The model validation methods presented here**,** along with reproducible code available on the REMA website (https://afsc-assessments.github.io/rema/articles/ex4_model_validation.html), are tools for stock assessment scientists to evaluate existing models and guide development of future models. From this analysis, we found that the REMA model can recover parameters across the range of complexity relevant to NPFMC Tier 4 and Tier 5 stocks. Our analysis suggests that the MCMC diagnostics and OSA residuals to be the most useful diagnostics for the REMA model. When models fail these diagnostics, it may indicate that the model is too complex and simplifications should be considered. We have highlighted several places in this document how diagnostic features can help make choices about the most appropriate model simplifications (pool process errors, remove additional observation errors, etc.) in light of REMA model assumptions and trade-offs between model parameters. Our results indicate that these diagnostics are most important to explore for existing models using multivariate configurations with multiple process errors, multi-survey versions using CPUE indices like the longline survey, and models estimating additional observation error. While model validation may not be needed for all existing REMA models, we recommend they be used when recommending new models for management.

References

Aeberhard, W.H., Flemming, J.M., Nielsen, A. 2018. Review of State-Space Models for Fisheries Science. Annual Review of Statistics and Its Application, 5(1), 215-235. https://doi.org/10.1146/annurev-statistics-031017-100427

Auger-Méthé, M., Field, C., Albertsen, C.M., Derocher, A.E., Lewis, M.A., Jonsen, I.D., Flemming, J.M. 2016. State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Scientific reports, 6(1), 26677.

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. Leos-Barajas, V., Flemming, J.M., Nielsen, A, Petris, G., Thomas, L. 2021. A guide to state–space modeling of ecological time series. Ecological Monographs, 91(4), e01470.

Campbell, D., & Lele, S. 2014. An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems. Computational Statistics & Data Analysis, 70, 257-267.

Cole, D. J. 2019. Parameter redundancy and identifiability in hidden Markov models. Metron, 77(2), 105-118.

Echave, K.B., Siwicke, K.A., Sullivan, J., Ferriss, B., and Hulson, P.J.F. 2022. Assessment of the Thornyhead stock complex in the Gulf of Alaska. In Stock assessment and fishery evaluation report for the groundfish fisheries of the Gulf of Alaska. North Pacific Fishery Management Council, 605 W. 4th. Avenue, Suite 306, Anchorage, AK 99501. https://apps-afsc.fisheries.noaa.gov/Plan_Team/2022/GOAthorny.pdf

Gabry, J. and Mahr, T. 2024. bayesplot: Plotting for Bayesian Models. R package version 1.11.1, https://mc-stan.org/bayesplot/.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A. 2019. Visualization in Bayesian workflow. J. R. Stat. Soc. A. (182) 389-402. doi:10.1111/rssa.12378 https://doi.org/10.1111/rssa.12378.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M. 2016. TMB: Automatic differentiation and Laplace approximation. J Stat Softw. 2016; 70(5):21. Epub 2016-04-04. https://doi.org/10.18637/jss.v070.i05

Monnahan, C.C., & Kristensen, K. 2018. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. PloS one, 13(5), e0197954.

Skaug, H.J., and Fournier, D.A. 2006. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Computational Statistics & Data Analysis, 51(2), 699-709.

Spies, I., Barbeaux, S., Hulson, H., Ortiz, I. 2023. Assessment of the Pacifc cod stock in the Aleutian Islands. In Stock assessment and fishery evaluation report for the groundfish fisheries of the Bering Sea and Aleutian Islands. North Pacific Fishery Management Council, 605 W. 4th. Avenue, Suite 306, Anchorage, AK 99501. https://apps-afsc.fisheries.noaa.gov/Plan_Team/2023/AIpcod.pdf

Stan Development Team. 2024. RStan: the R interface to Stan. R package version 2.32.6, https://mc-stan.org/.

Sullivan, J., Monnahan, C. Hulson, P. Ianelli, J. Thorson, J. Havron, A. 2022. REMA: A consensus version of the random effects model for ABC apportionment and Tier 4/5 assessments. Plan Team Report, Joint Groundfish Plan Teams, North Pacific Fishery Management Council. 605 W 4th Ave, Suite 306 Anchorage, AK 99501. Available through the Oct 2022 Joint GPT e-Agenda.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., & Nielsen, A. 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics, 24(2), 317-339. https://doi.org/10.1007/s10651-017-0372-4

Tables

Table 1. AI Pcod and GOA Thornyhead fixed effect estimates with 95% confidence or credible intervals for maximum likelihood or MCMC models, respectively (CI). Fixed effects estimates are compared between the base model run with marginal maximum likelihood estimation assuming the Laplace approximation ("Base_MMLE_Laplace") and MCMC models with and without the Laplace approximation ("MCMC_Laplace" and "MCMC_withoutLaplace", respectively). The Rhat statistic is provided for MCMC models and is a convergence metric (convergence = Rhat = 1). Nonconverged parameters are highlighted in bold. The parameter estimate for the MCMC models is the median of the posterior samples.

Stock	FixedEffect	Model	Estimate (95% CI)	Rhat
AI Pcod	log_PE	Base_MMLE_Laplace	-1.87 (-2.53, -1.21)	-
AI Pcod	log_PE	MCMC_Laplace	-1.88 (-2.58, -1.19)	1.003
AI Pcod	log_PE	MCMC_withoutLaplace	-1.89 (-2.68, -1.19)	1.018
GOA Thornyhead	log_PE[1]	Base_MMLE_Laplace	-2.16 (-2.76, -1.55)	-
GOA Thornyhead	log_PE[1]	MCMC_Laplace	-1.84 (-2.28, -1.42)	1.000
GOA Thornyhead	log_PE[1]	MCMC_withoutLaplace	-1.85 (-2.32, -1.41)	1.011
GOA Thornyhead	log_PE[2]	Base_MMLE_Laplace	-2.26 (-2.85, -1.66)	-
GOA Thornyhead	log_PE[2]	MCMC_Laplace	-2.08 (-2.52, -1.66)	1.000
GOA Thornyhead	log_PE[2]	MCMC_withoutLaplace	-2.08 (-2.53, -1.67)	1.004
GOA Thornyhead	log_PE[3]	Base_MMLE_Laplace	-1.49 (-2.09, -0.90)	-
GOA Thornyhead	log_PE[3]	MCMC_Laplace	-1.21 (-1.58, -0.86)	1.000
GOA Thornyhead	log_PE[3]	MCMC_withoutLaplace	-1.19 (-1.56, -0.86)	1.002
GOA Thornyhead	log_q	Base_MMLE_Laplace	-0.51 (-0.58, -0.44)	-
GOA Thornyhead	log_q	MCMC_Laplace	-0.51 (-0.57, -0.46)	1.000
GOA Thornyhead	log_q	MCMC_withoutLaplace	-0.52 (-0.57, -0.46)	1.000
GOA Thornyhead	log_tau_biomass	Base_MMLE_Laplace	-1.72 (-2.29, -1.15)	-
GOA Thornyhead	**log_tau_biomass**	**MCMC_Laplace**	**-3.16e+13 (-3.7e+16, -8.71e+11)**	**3.335**
GOA Thornyhead	**log_tau_biomass**	**MCMC_withoutLaplace**	**-1.21e+10 (-1e+11, -2.57e+05)**	**2.793**
GOA Thornyhead	log_tau_cpue	Base_MMLE_Laplace	-1.93 (-2.26, -1.61)	-
GOA Thornyhead	log_tau_cpue	MCMC_Laplace	-1.87 (-2.23, -1.58)	1.000
GOA Thornyhead	log_tau_cpue	MCMC_withoutLaplace	-1.88 (-2.23, -1.58)	1.000

Figures

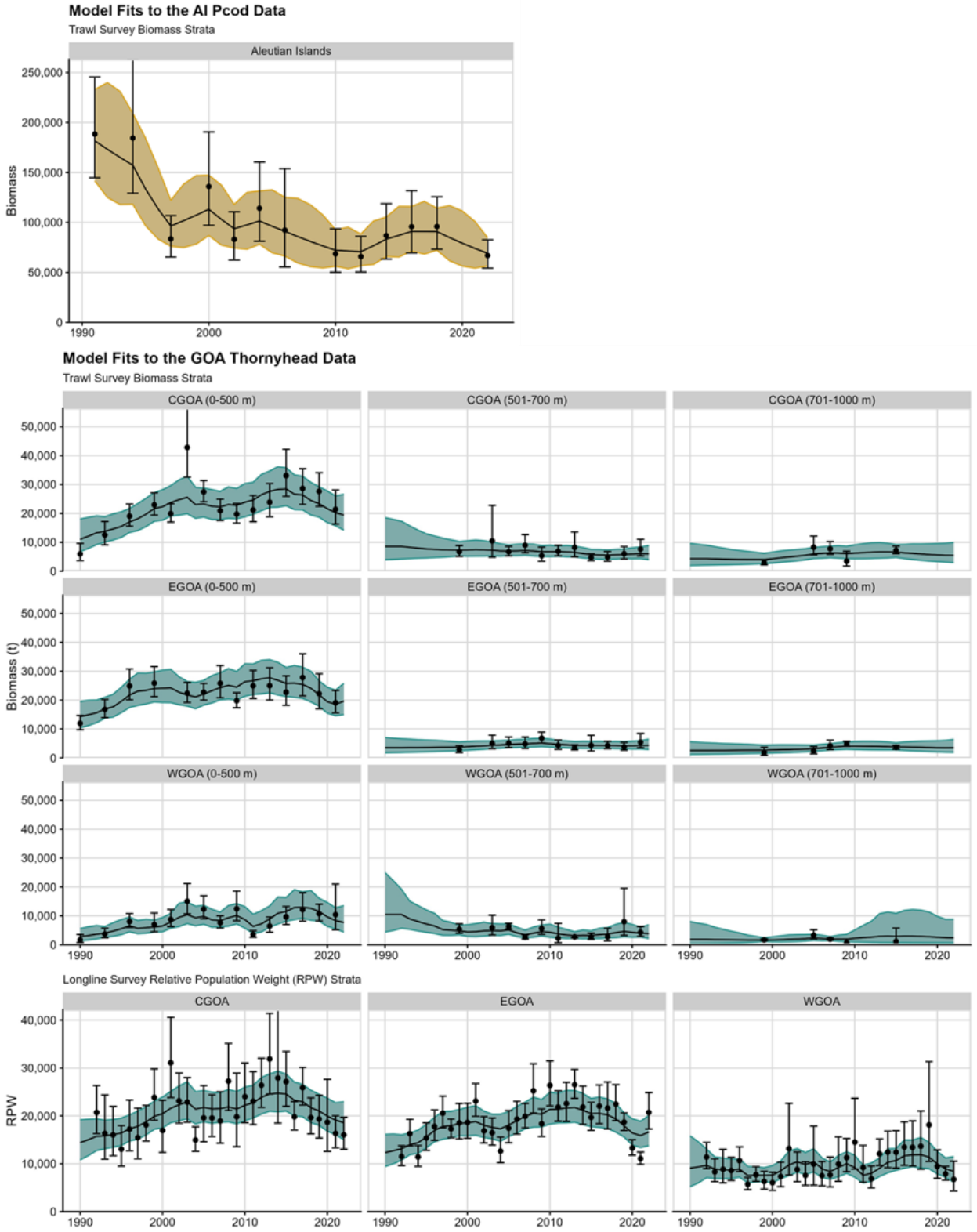

Figure 1. Fits to the AI Pcod (top panel; gold) and GOA Thornyhead (bottom panels; teal) models.

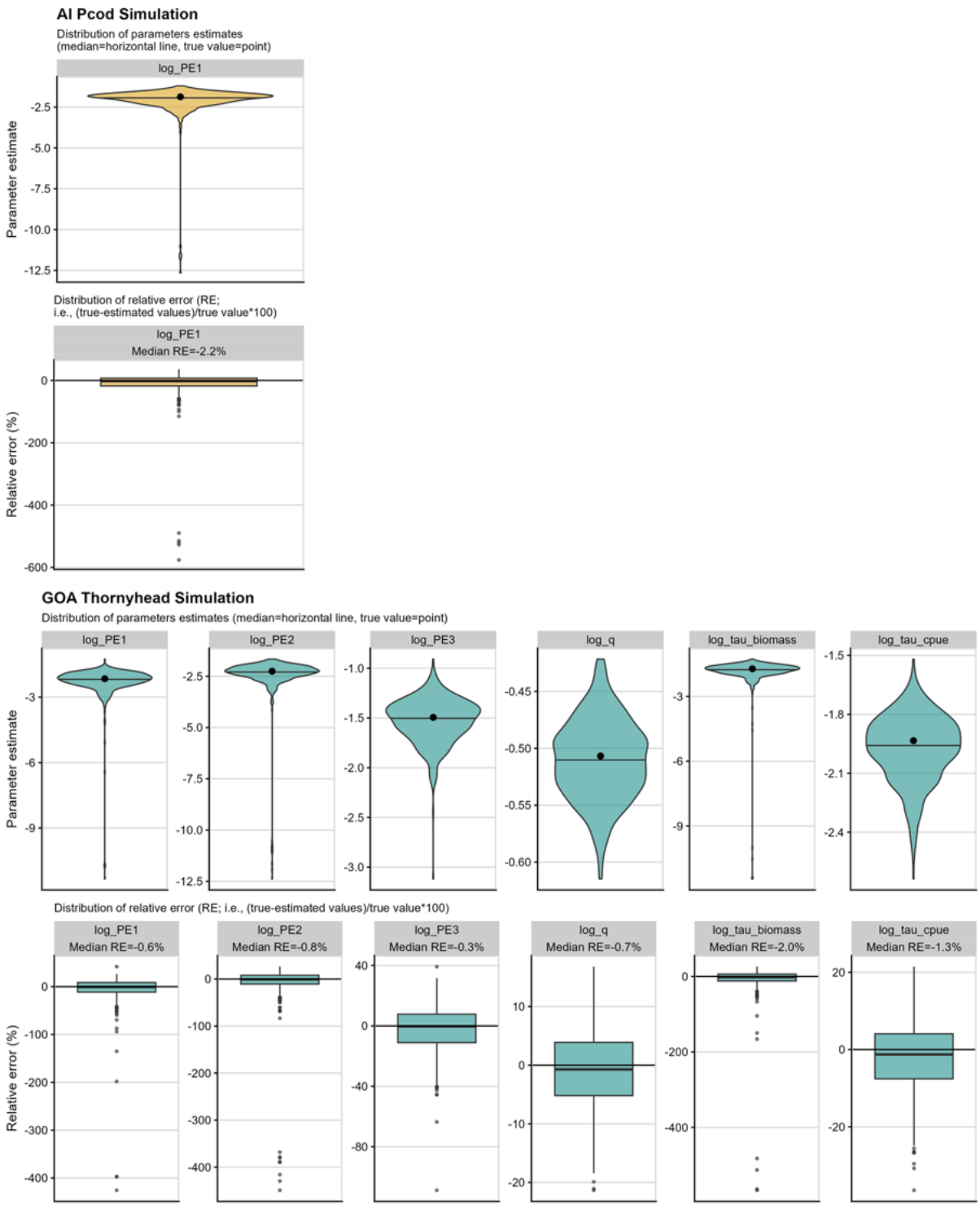

Figure 2. Simulation self-test results for AI Pcod (top panels; gold) and GOA Thornyhead (bottom panels; teal). The upper row of each panel gives the distribution of the recovered parameters, with the dot giving the "true value" used to simulate the data. The lower row of each panel gives the relative error of the recovered parameters, with the zero line indicating the simulation returns the "true value," the box plot line giving the median of the recovered parameter's relative errors, and the whiskers and dots of the boxplots giving the spread of the recovered parameter's relative errors.

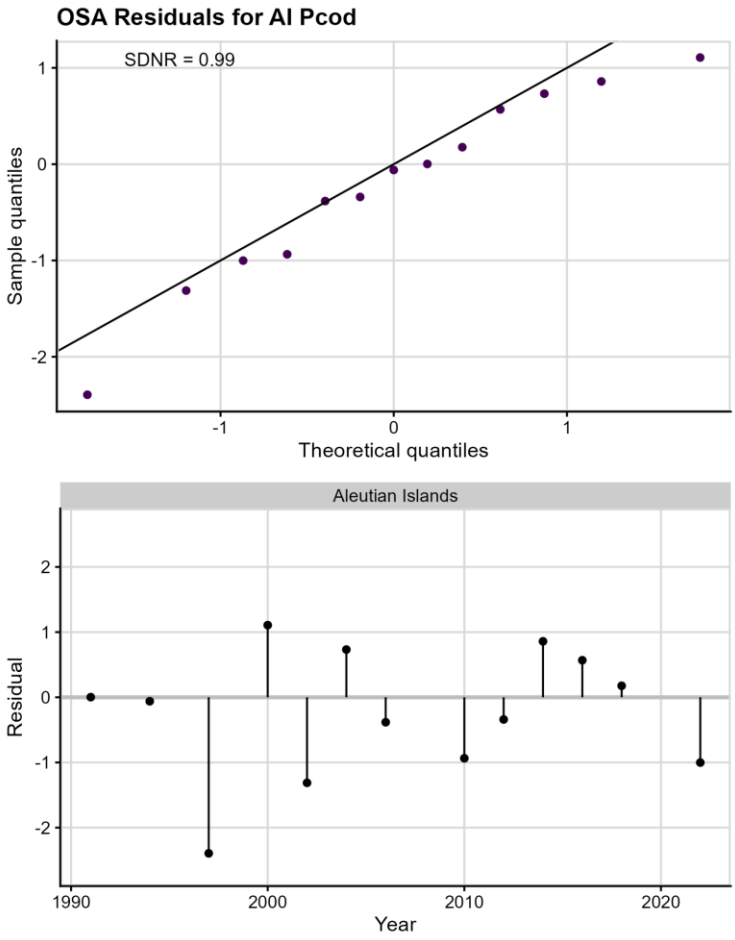

Figure 3. One-step ahead (OSA) residual plots for AI Pcod. The top panel gives the QQ plot, comparing the normal distribution quantiles (*x*-axis) to the OSA residual distribution quantiles (*y*-axis). The diagonal line is the 0-1 line, where the two quantiles are equal, and the SDNR statistic is given in the upper left of the plot. The bottom panel gives the residuals (*y*-axis) plotted by year (*x*-axis) which is used to check for independence.

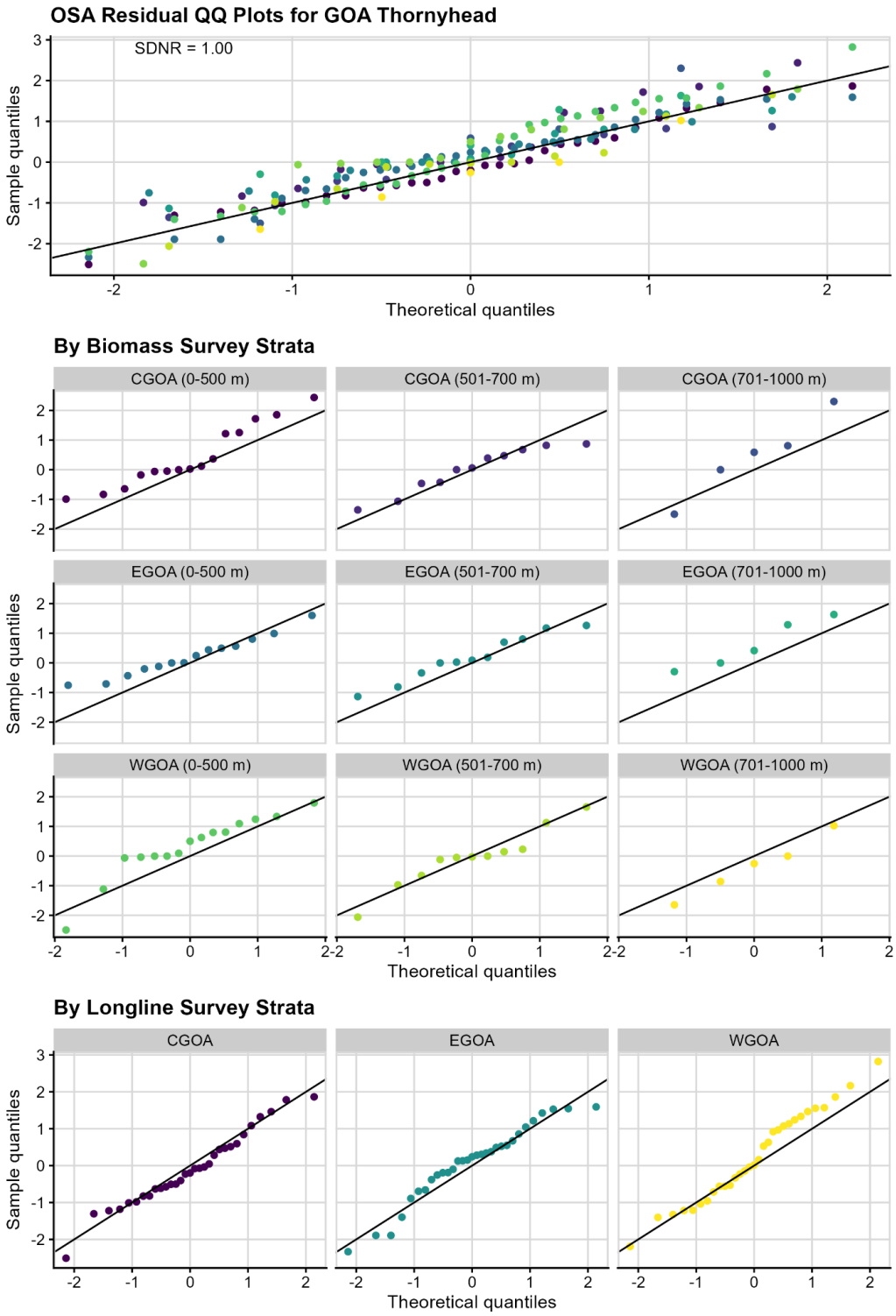

Figure 4. One-step ahead (OSA) QQ plots across (top panel) and within strata (middle panel, bottom trawl survey strata; and bottom panel, longline survey strata) for GOA Thornyhead. In both panels, the colors correspond to bottom trawl survey strata. The SDNR statistic is given for the across strata QQ plot; the within strata plots are useful to check for extreme skew but lack statistical power.

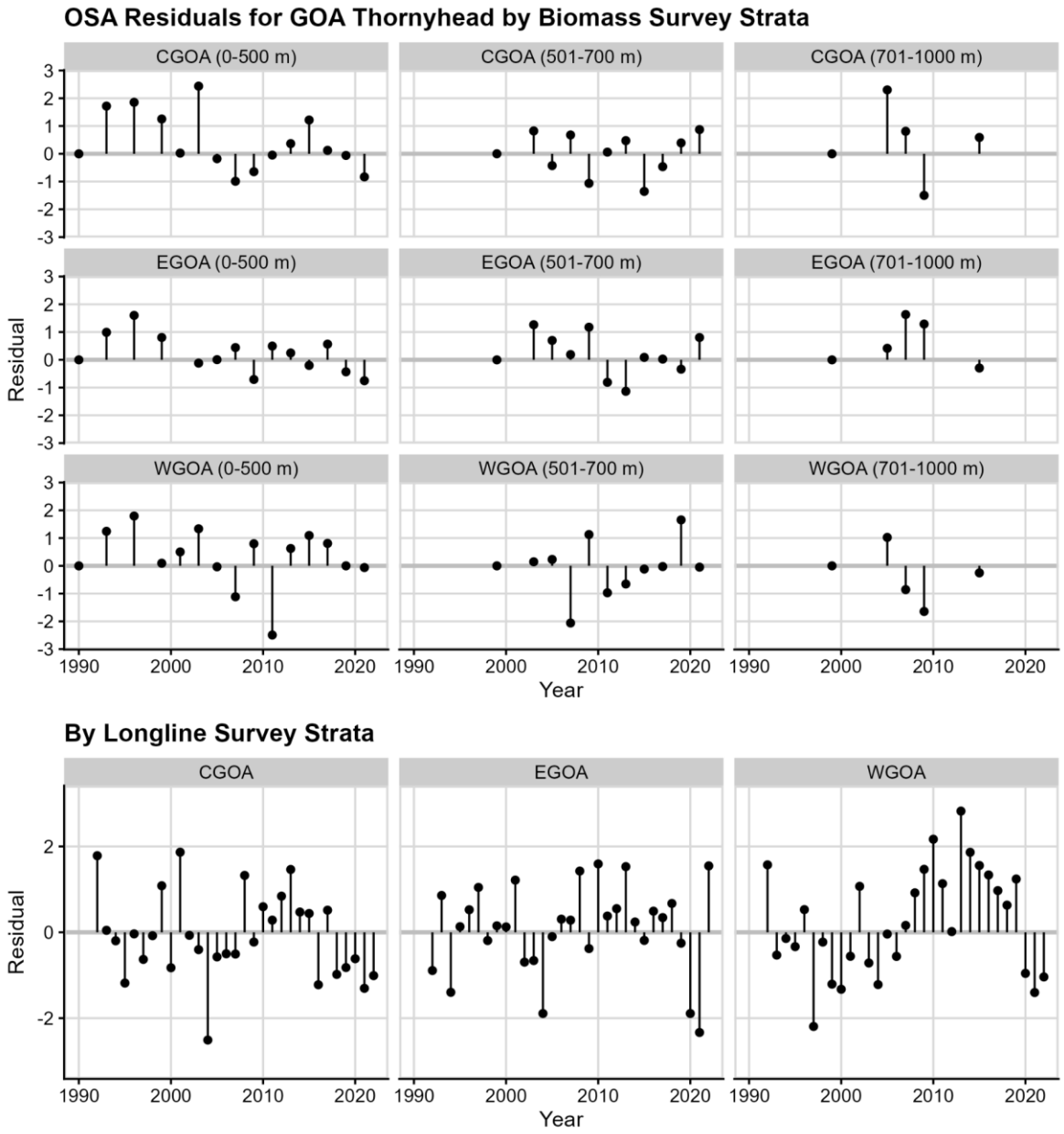

Figure 5. One-step ahead (OSA) residual plots for GOA Thornyhead. The residuals (*y*-axis) for each annual time step (random effect; *x*-axis) for each survey stratum.

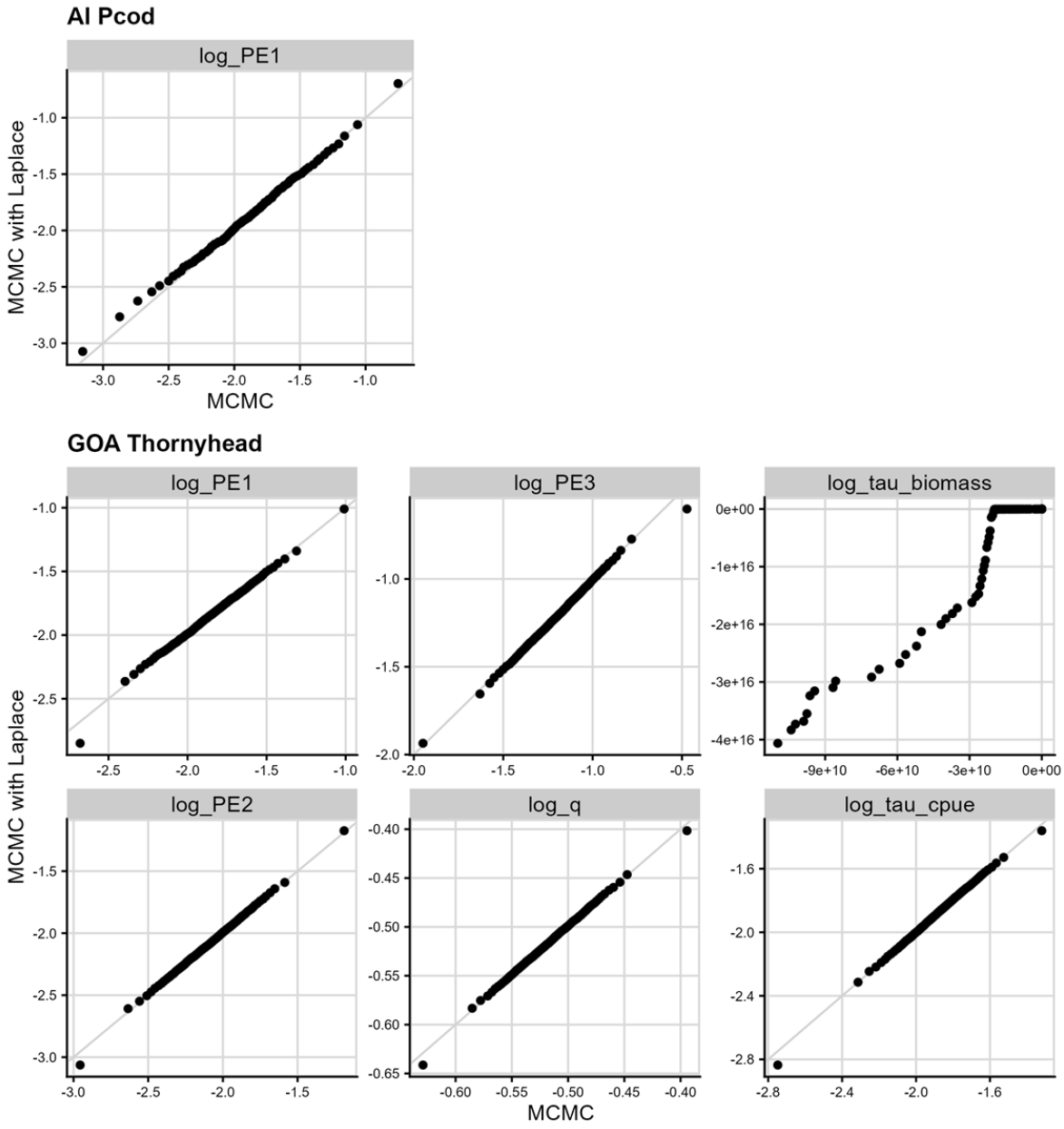

Figure 6. Results from Laplace approximation testing. We compare the results of running a model without assuming normality of random effects (MCMC, *x*-axis) and running a model assuming normality of random effects (Laplace approximation, *y*-axis), where the quantiles of each are plotted against each other for each fixed effect in the model. The gray line is the 1:1 line; points that fall on the gray line indicate that the quantile value is the same between the two cases. The top panel gives the plot for the AI Pcod model (one fixed effect), and the bottom panels give the plots for the GOA Thornyhead model (six fixed effects).

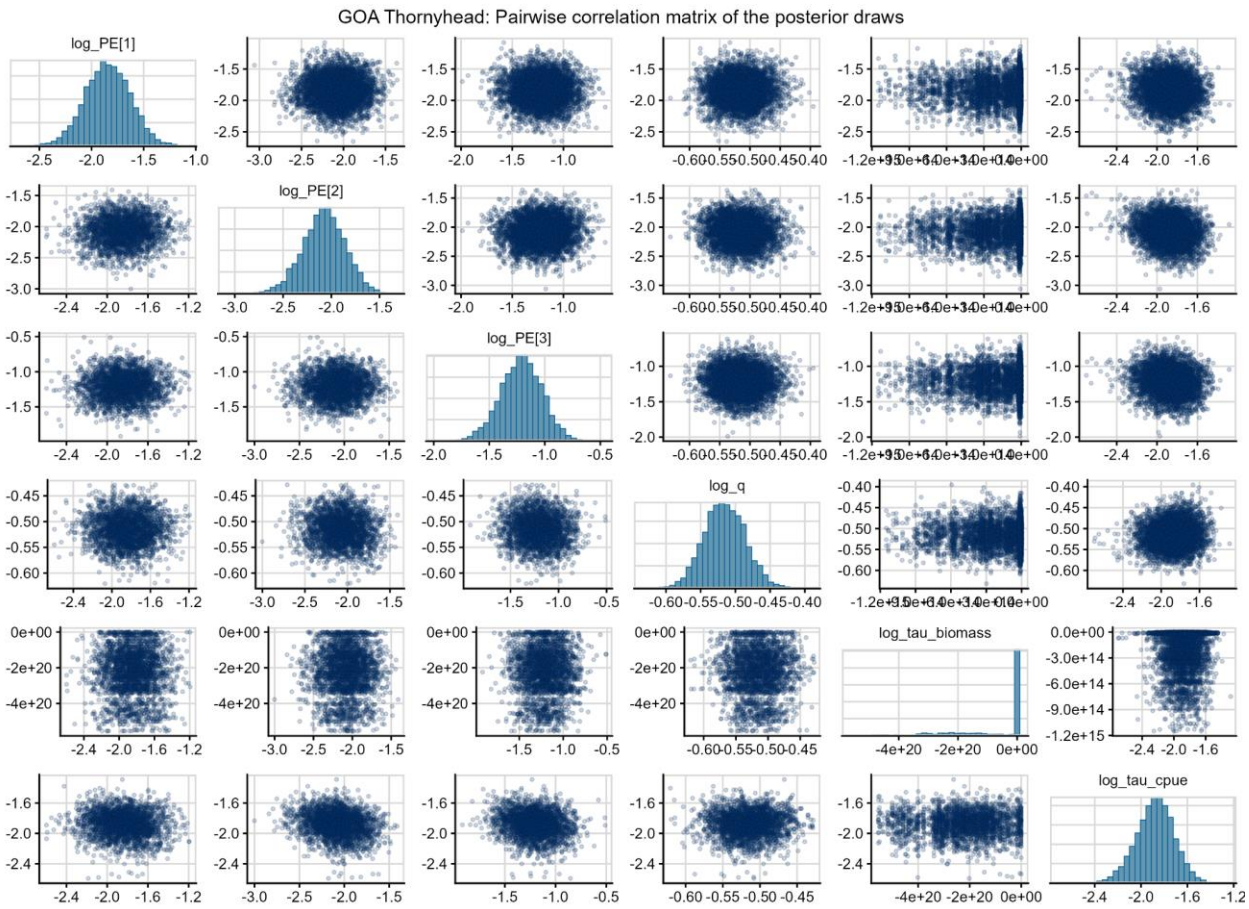

Figure 7. Pairs plot for the GOA Thornyhead MCMC model assuming the Laplace approximation. The diagonal plots give the distribution of each fixed effect for 2500 MCMC draws. The off-diagonal elements give the parameter-by-parameter correlation, where each point gives the parameter values estimated from the same MCMC simulation.

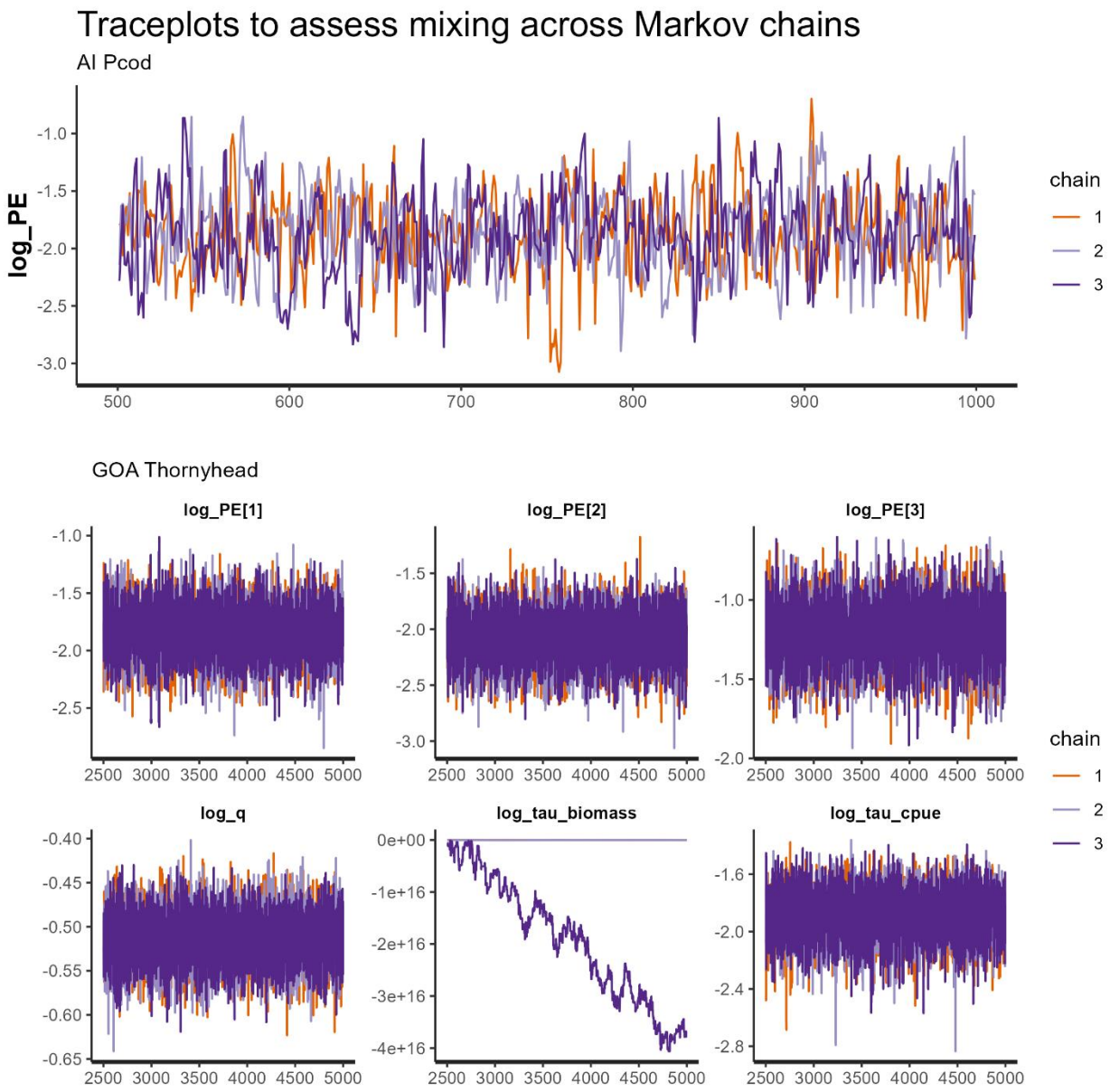

Figure 8. MCMC traceplots for AI Pcod (top panel) and GOA Thornyhead (bottom panels). Each subpanel is an individual parameter, and the plot gives the value of the parameter (*y*-axis) for each simulation draw (*x*-axis) for each model chain (color).