Potential CPUE model improvements for the primary index of Southern Bluefin Tuna abundance.

Simon Hoyle

Hoyle Consulting, simon.hoyle@gmail.com.

Working Paper CCSBT-ESC/2108/35 prepared for the CCSBT Extended Scientific Committee for the 26th Meeting of the Scientific Committee, 23-31 August 2021.

Abstract

Generalized additive models were used to compare approaches for fitting aggregated CPUE data for southern bluefin tuna. Data aggregation caused some problems with model fit by introducing an inverse correlation between expected catch rate and residual variance. Delta lognormal models were selected as the preferred approach. The extreme value diagnostic was improved to account for the sizes of the extreme values. Models with different components were compared using maximum likelihood (ML) and final smoothers fitted using restricted ML (REML). Final models used *ti()* terms to specify each model component, and gamma of 2 to reduce the effective sample size. The final models fitted the data well and generated plausible values in strata without observations. Indices were adjusted to account for differences in the ocean areas of spatial strata.

Introduction

The CPUE standardization methods used for SBT need to be updated because of problems with recent estimates, particularly an anomalously high value in 2018 (CCSBT 2020). The main reason for the unreliability of the estimates has been identified as increasing aggregation of fishing effort, which has caused parameter estimation problems for the models used to date. Analyses in 2020 (Hoyle 2020) developed an alternative approach using generalized additive models (GAMs) implemented with the R package *mgcv* (Wood 2011). Data were fitted with multi-dimensional smoothers which share information among adjacent values of continuous variables. This work was considered preliminary (CCSBT 2020), and additional work needs were identified in several areas.

The preliminary analyses followed previous work (Nishida and Tsuji 1998; Itoh and Takahashi 2019) in assuming that the uncertainty distribution was lognormal after adding a constant to the CPUE values. This constant was defined as 10% of the mean CPUE across the whole dataset (Campbell et al. 1996; Campbell 2004). This historically common 'lognormal constant' approach has been superseded (Maunder and Punt 2004; Bellego and Pape 2019) by methods that either permit zero observations, or use a hurdle process to separately model the probability of nonzero catch and catch rates in nonzero catches (Lo et al. 1992). Work was therefore required to consider alternative uncertainty distributions, including both models that permit zero observations and hurdle methods.

Preliminary models were fitted using the GCV criterion and fitting procedure. However, the generalized cross-validation criterion GCV has a tendency to under-smooth (Wood 2011). In contrast, REML and ML smoothness selection more strongly penalizes overfit than GCV, which gives them more clearly defined optima and greatly reduces the risk of undersmoothing. However, their use can considerably increase the computation time, is unstable for some basis functions, and comparison of fixed effect models using REML is not valid. Model fitting using both ML and REML was explored.

The *mgcv* package determines the appropriate smoothness of model components. Increasing the gamma parameter from the default value of 1 produces smoother models by multiplying the model's effective degrees of freedom, equivalent to reducing the effective sample size. Kim,Gu (2004) recommend setting gamma to 1.4 to reduce the chance of overfitting when using GCV, given its tendency to slightly overfit, but this problem is worse with CPUE data which are generally over-dispersed which leads to overfitting. Overfitting is a particular problem with operational set by set data because sets are not independent, but also occurs in aggregated data. An important cause is parameters that affect CPUE but are not available for inclusion in the model, such as environmental factors and information sharing among vessels. To avoid overfitting, the gamma parameter was set to 2, but there was a need to explore the effects of alternative values of the gamma parameter.

Smooth terms in GAMs are specified in a formula using one or more *s*, *te*, *ti*, and/or *t2* terms (Wood 2011). Each term identifies one (*s*) or more (*te, ti, t2*) variables, along with the dimension and class of the basis function used to represent the smooth term. The preliminary modeling considered alternative combinations of interactions among variables but did not explore different approaches for specifying the interactions, or alternative basis functions. Further exploration of these issues was recommended. Wood states that "gam allows nesting (or 'overlap') of te and s smooths, and automatically generates side conditions to make such models identifiable, but the resulting models are much less stable and interpretable than those constructed using ti terms."

Problems with recent estimates were associated with extreme values being estimated in strata with no observations. Lack of observations meant that there was no constraint on predictions in these areas. A diagnostic developed (Hoyle 2020) to identify extreme values among these predictions

counted the number of strata in which the predicted value was greater than the largest observed stratum value for the year. This approach did not account for how much the predicted CPUE was larger than the highest observed value, and further work was requested to explore improvements to the diagnostic.

Statistical area weighting has been recommended for CPUE analyses, based on work done for eastern Pacific tunas and southern bluefin tuna (Punsly 1987; Campbell 2004). This approach is designed for situations in which time-area interactions are not accounted for in the model, so is probably unnecessary for models that do take spatiotemporal trends into account. The effect of using this approach was explored.

The preliminary models developed indices using the mean annual CPUE across all strata, which assumed that the density implied by CPUE was directly related to relative abundance. However, abundance is the product of density and area, and different spatial cells include different areas of ocean. Some spatial cells include land as well as ocean, and cell areas diminish further from the equator. The presence of time-area interactions in the models indicated the potential for ocean area to affect abundance trends. There was therefore a need to change the analysis approach so that indices were based on trends in abundance rather than cell density.

Estimating trends through time requires predictions of CPUE for the same cells in each year. To avoid inclusion of spatial cells with too much uncertainty due to insufficient sampling effort, the preliminary approach included only spatial cells with at least 15 observations in the aggregated dataset (at the year-month-cell level). The criterion of 15 observations was chosen without considering alternatives, and there was a need to explore the effects of other values.

Methods

Input data

These analyses were based on a slightly different dataset from that used for the primary analyses (Itoh & Takahashi 2019), which is only available to Japanese scientists. The available dataset was sufficiently similar to the primary dataset to provide useful insights. The main differences between the two datasets are listed below.

- The primary dataset uses a set of core vessels that have high SBT catches for at least 3 years, whereas the available dataset includes data from all vessels.
- The primary dataset includes catches of bigeye and yellowfin tuna, but the available dataset does not.
- The primary dataset is available as operational (set by set) data (but is aggregated for the main analysis) whereas the available dataset is aggregated.

The data file ‘CPUEInputs_2021_June.txt’, available from the private area of the CCSBT website, was used for the analysis. These data are aggregated by year, month, and 5° latitude and longitude, with catches reported by age class based on spatially and temporally stratified size sampling.

The following processes were then applied to the dataset:

- Filter to include effort from 1986 to 2018, with DATA_CODE ‘COMBINED’, in statistical areas 4 to 9, and months 4 to 9. Include strata with more than 10 000 hooks. Include latitudes north of 50° S.
- Create numeric catch variable, the sum of catches of all SBT 4+ and older.

- Create categorical llf variable, indicating 5° square that combines latitude and longitude.
- Create categorical areaf variable, which merges statistical area 4 with 5 and statistical area 6 with 7.
- Create categorical variables yf, latf, and mf, for year, latitude, and month.
- Adjust numeric longitude variable (lon) by adding 360 to all longitudes between -180 and -100, to provide continuity across the spatial domain of the fishery. Longitudes are recorded as -180 to 180 and so the range of the adjusted longitude variable was from -95 to 260.
- Create numeric cpue variable = catch per 1000 hooks.
- Remove a single outlier with cpue > 120.

Error distribution assumption

Using the variables and interactions previously selected (Hoyle 2020), a series of models were run with alternative assumptions about the uncertainty distribution.

Previous models used the lognormal constant approach, i.e. a normal distribution with response variable *log(cpue + k)*, where k is 10% of the mean of all cpue values. Alternative distribution functions included the following: binomial distribution with two link function options, lognormal and gamma models for nonzero catch rates, and Tweedie and negative binomial for discrete distributions.

All models started with the same sets of variables on the right-hand side of the equation (RHS). The *mgcv* package uses the offered terms and initial basis dimension (*k*) as a starting point for a search. The *k* parameter sets the upper limit on the degrees of freedom associated with a smooth s, while for a *te* or *ti* tensor product smooth the upper limit is the product of the *k* values for each marginal smooth. The basic form of the model was selected by AIC after fitting with maximum likelihood or GCV, and the optimal smooths were obtained by fitting with REML.

RHS = yf + te(lon2, LAT, k = c(40,4)) + te(MONTH, LAT, k = c(6,4)) + te(lon2, MONTH, k = c(10, 5)) + te(YEAR, LAT, k = c(20, 4)) + te(YEAR, MONTH, k = c(20, 5)) + te(LAT, lon2, MONTH, k = c(4,15, 6)) + te(LAT, lon2, YEAR, k = c(4,10, 9))

The lognormal model used *log(cpue)* as the response, with identity link and gaussian error distribution, while the Gamma model used *cpue* as the response, with log link function and Gamma error distribution (Table 1).

Binomial models added effort to the RHS formula above, to account for the effect of effort on the probability of non-zero catch in a stratum. Effort was included as a spline rather than a straight line or offset, to allow for potential nonlinearity in the relationship.

catch > 0 ~ RHS + s(log(N_HOOKS))

Models with discrete response terms (Tweedie, negative binomial) used *catch* as the response variable and included effort on the right-hand side of the equation.

catch ~ RHS + s(log(N_HOOKS))

The Tweedie parameter *p* and the negative binomial scale parameter *theta* were difficult to estimate with full models, which were unstable when estimating these parameters. Instead, these parameters were estimated with a simpler gam (see below) and provided to the model for more complex analyses. For Tweedie distributions *p* was estimated using the tw() family and models with fixed *p* fitted using the Tweedie() family. For the negative binomial, *theta* was estimated using the nb() family and full models fitted with the negbin() family.

sbt ~ yf + te(lon2, LAT, k = c(20,3)) + te(MONTH, LAT, k = c(6,4)) + te(lon2, MONTH, k = c(9, 4)) + te(YEAR, MONTH, k = c(10, 4)) + s(log(N_HOOKS))

Table 1: Settings used in mgcv to compare models with different distributions.

Distribution	Family	Dataset	response	Link function	Likelihood
lognormal + constant	Gaussian	all	log(cpue+c)	identity	ML
Binomial (DLN)	Binomial	all	cpue > 0	logit	REML
Lognormal (DLN)	Gaussian	nonzero	log(cpue)	identity	ML
Binomial (delta Poisson)	Binomial	all	cpue > 0	cloglog	REML
Gamma (delta Poisson)	Gamma	nonzero	cpue	log	ML
Tweedie	Tweedie	all	sbt	log, p = 1.5	REML
Negative binomial	negbin	all	sbt	log, *theta*=0.9	REML

delta11 <- gam(cpue > 0 ~ formula, data = a, gamma = 2, method = 'REML', family = binomial)

delpo11 <- gam(cpue > 0 ~ formula, data = a, gamma = 2, method = 'REML', family = binomial(link=cloglog))

pos11 <- bam(log(cpue) ~formula, data = apos, gamma = 2)

gamma11 <- gam(cpue ~ formula, data = apos, gamma = 2, family = Gamma(link = "log"))

modtw11 <- gam(sbt ~ formula, data = a, gamma = 2, method = 'REML', family = Tweedie(*p*=1.5))

modnb11 <- gam(sbt ~ formula, data = a, gamma = 2, method = 'REML', family = negbin(0.9))

Model diagnostics

To explore how data aggregation may affect residuals and consistency with model assumptions, patterns in the residuals from the lognormal constant model were explored by fitting generalized additive models. First, the absolute value of residuals was modelled as a function of effort: abs(resid) ~ s(n_HOOKS), and the relationship was plotted. Next, CPUE predicted from the fitted model in strata with effort was modelled as a function of effort after adjusting for year: pred_CPUE ~ s(N_HOOKS) + yf. Finally, residual variance was modelled as a function of predicted CPUE after adjusting for year: abs(residuals) ~ s(predicted CPUE) + yf.

Models fits were initially explored using the standard R diagnostic plots as implemented in the mgcViz package (Fasiolo et al. 2020). These include a Q-Q plot with theoretical quantiles based on simulated residuals, a histogram of residuals, a scatter plot of predicted values versus residuals, and a scatter plot of predicted versus observed values.

Further diagnostic analyses were carried out using the DHARMa package (Hartig 2020). These diagnostics involve simulating residuals from the fitted model and comparing the distributions of simulated and observed residuals, with a number of plots and statistical tests available.

Q-Q plots were generated suing simulated residuals, and each Q-Q plot also ran a uniformity test (Kolmogorov-Smirnov test for overall uniformity of the residuals) and a dispersion test (a simulation-based test for over/under-dispersion). For further details see Fasiolo et al. (2020).

Boxplots of residuals were plotted for each categorical variable, to compare the relationship between the quantile distribution and the predicted response. For each boxplot tests were run to check whether residuals within each group were distributed according to model assumptions (multiple Kolmogorov-Smirnov tests, with adjustment of p-values for multiple testing), and whether the variance between groups was heterogeneous (using a Levene test). Residuals were also plotted against the smoothed year effect and model predictions, and tests carried out to detect whether quantiles deviated from the expected values. For further details of these tests see Fasiolo et al. (2020). Finally, the median residuals were plotted against rank-transformed model predictions.

Extreme value diagnostic

The previous version of extreme value diagnostic was based on counting the number of predictions per year that exceed the maximum value observed in that year.

The diagnostic was updated by adding two components

1. Differences by stratum between the predicted stratum value and the maximum observed value in a year across all strata, for those predictions that exceeded the maximum observed value. Only strata without observations were included in the diagnostic since strata with observations are constrained by their contribution to the likelihood.
2. The above value squared.

These components were provided summed by year and summed across all years.

Changes to modeling approach

Statistical area weighting

The recommended approach for fitting CPUE models that do not include time-area interactions (Punsly 1987; Campbell 2004) gives the same statistical weight to each unit of area, by counting the number of rows of data n(a, year) associated with each area and applying statistical weight of 1/n(a,yr). The effect of this approach was explored through simulation and by applying statistical reweighting to the indices.

Alternative gamma adjustments to effective sample size

Both the delta and lognormal positive models were rerun with different levels of gamma: 1, 2, 4, 10.

Alternative smoothers

It is recommended (Wood 2011) that models with multiple levels of interactions should, rather than using *te()* smoothers for all terms, specify main effects using either *s()* or *ti()* and interaction terms with *ti()*. Trials in 2020 with the lognormal constant model found excess extreme values when using the *ti()* approach, but these models included a mixture of thin plate regression splines (*s()* terms) and cubic splines (*ti()* terms), so were different from the *te()* models which used cubic splines only. These trials were revisited for the delta lognormal approach, by comparing models using *te()* with models using *ti()* for all terms.

Model 11 RHS = yf + ti(lon2, k=40) + ti(LAT, k=4) + ti(MONTH, k = 6) + ti(lon2, LAT, k = c(40,4)) + ti(MONTH, LAT, k = c(6,4)) + ti(lon2, MONTH, k = c(10, 5)) + ti(YEAR, LAT, k = c(20, 4)) + ti(YEAR, MONTH, k = c(20, 5)) + ti(lon2, YEAR, k =c(10, 9)) + ti(LAT, lon2, MONTH, k = c(4,15, 6)) + ti(LAT, lon2, YEAR, k = c(4,10, 9))

Alternative optimisation criteria

Models were fitted using either the generalized cross-validation (GCV), maximum likelihood (ML), or restricted maximum likelihood (REML). Delta models were fitted using ML to compare models with different number of terms, and REML in final models to estimate the smooth terms. Positive models

were fitted using GCV to compare models with different number of terms, and REML in final models to estimate the smooth terms. Positive models were unstable when fitted with ML.

Alternative basis functions

Mgcv provides a variety of basis functions. The default basis function for the s() function is 'tp', the thin plate regression spline, while the default for tensor products *te()* and *ti()* is 'cr', the cubic regression spline. Wood (2019) states that thin plate regression splines tend to give the best MSE performance within s() terms, but are slower to set up than the other bases, and within tensor products seem to offer no advantage over cubic splines. Nevertheless, to complement the default approach with the *ti()* model above of using 'cr' for *ti()* terms, models were set up applying 'tp' basis functions to all terms.

RHS = yf + ti(lon2, k=40, bs = 'tp') + ti(LAT, k=4, bs = 'tp') + ti(MONTH, k = 6, bs = 'tp') + ti(lon2, LAT, k = c(40,4), bs = 'tp') + ti(MONTH, LAT, k = c(6,4), bs = 'tp') + ti(lon2, MONTH, k = c(10, 5), bs = 'tp') + ti(YEAR, LAT, k = c(20, 4), bs = 'tp') + ti(YEAR, MONTH, k = c(20, 5), bs = 'tp') + ti(lon2, YEAR, k =c(10, 20), bs = 'tp') + ti(LAT, lon2, MONTH, k = c(4,15, 6), bs = 'tp') + ti(LAT, lon2, YEAR, k = c(4,10, 9), bs = 'tp')

However, models set up in this way were unstable and caused R to crash, so this approach was not taken further.

As a final step the *ti* model was specified using a 'shrinkage' version of the cubic spline smooth (bs = "cs"), which can penalise a curve to zero and effectively eliminate it from the model.

Changes to preparation of indices

Abundance adjustment for ocean area

Ocean areas of all cells were calculated (Hoyle and Langley 2020) and used to adjust CPUE estimates before calculating indices.

Criteria for including spatial strata

The effects were explored of using alternative minimum numbers of observations as the criterion for including spatial cells in the dataset (which is aggregated at the year-month-cell level). Alternatives trialled were 1, 5, 10, 15, 20, and 25 observations.

Effect of extreme values on indices

Extreme values in strata without observations are problematic if they are sufficiently large to affect the index. To explore the potential effects of extreme values on the index, indices were recalculated after reducing all extreme values in each year to be equal to the largest observed value in that year.

Model selection

A series of binomial models with logit link and lognormal positive models was run with a range of smoother configurations with a wider range using the *te()* approach (Table 3), followed by focused analyses using the *ti()* approach (Table 4). Models were compared using AIC after ML fitting, and extreme value prediction performance for versions of the models fitted using REML.

The *ti()* approach involved separately fitting all main effects and interaction terms, and specifying all terms involved in higher order terms in the model. This differed from the *te()* approach, in which all one-way and some two-way terms were implicitly specified more than once, by being included in two or more higher order terms.

R code

All R code is available at the github repository https://github.com/hoyles/R_ccsbt_cpue. Please email simon.hoyle@gmail.com to request permission to access the repository.

Results

Data aggregation

Residuals showed a strong pattern of declining variance with increasing effort (Figure 1, top left), as expected since higher effort means a stratum CPUE is the average of more sets with the consequent reduced variability. More interesting but also unsurprising was that higher effort was associated with higher CPUE (Figure 1, top right). The SBT fleet wants to catch more SBT, so tends to focus on areas with higher catch rates. The consequence of these two effects is a negative relationship between expected CPUE and residual variance (Figure 1, lower left), which differs from the assumptions of the available error models.

Distributions

Diagnostics run for all distributions included standard arrays of four diagnostic residual plots (Figures 2 - 5), Q-Q plots with tests of uniformity (Kolmogorov-Smirnov) and dispersion (Figure 6), plots of residual distributions by group with tests of within-group uniformity and Levene tests of homogeneity of variance (Figures 7 - 13), and scaled residual plots (Figure 14).

Most of the models showed some lack of fit, as expected given the use of aggregated data. With aggregated data the residual variance is related to the effort associated with each row of data. Effort varies spatially and temporally and tends to be greater in areas with higher catch rates, so the mean and variance are likely to be negatively correlated. This violates the model assumption that variance is homoscedastic.

All error distributions except the two delta models showed a pattern of residuals that were more peaked than expected in the middle, indicated by residual histograms and the S-shape of the Q-Q plots (Figures 2 - 5). The lognormal constant, lognormal positive and Gamma models (Figures 2, 3 and 4) also showed overdispersion with excess variability of outliers at both ends, while the Tweedie (Figure 2) and negative binomial (Figure 5) models fitted better at the lower end but with excess outliers at the upper end.

All models failed the dispersion tests, and all but the delta models failed the tests of uniformity (Figure 6).

More detailed explorations of residuals by input variable showed substantial variation in residual variation for the lognormal constant model by latitude and to a lesser extent by month (Figure 7). There was no apparent pattern by year in median quantile deviation by year, but there was strong dependency of residuals on model predictions.

Both delta models fitted the data well, with the logit link (Figure 8) fitting slightly better than the complementary log-log (cloglog) link (Figure 9). The cloglog link function is asymmetric and preferred for extreme events where the probability is close to either 0 or 1, whereas in this dataset the average proportion of nonzero catches is more moderate at 84.8%, varying between 63% and 93% (Figure 15).

Neither the lognormal positive (Figure 10) nor the Gamma positive model (Figure 11) fitted the data well. Residuals were heteroscedastic with much larger residuals at lower latitudes. Quantile tests showed significant lack of fit. Fits appeared slightly worse and potentially more biased for the

gamma distribution compared to the lognormal, with more bias in the residuals particularly at lower predicted values (Figure 14).

The Tweedie model also had substantial heteroscedasticity, and some potentially concerning bias in median residuals by year (Figure 12). However, it fitted considerably better than the negative binomial model (Figure 13), which showed large deviations from the expected distribution.

Changes to modeling approach

Statistical area weighting

Giving consistent statistical weights to each spatial stratum noticeably changed the trend of the model. However, simulations showed that this adjustment is unnecessary for models that include time-area interactions, and that the adjustment of statistical weighting increases uncertainty and can cause bias.

Alternative gamma adjustments to effective sample size

Changing the gamma assumption about effective sample size changed the trend in the model since 2013 but had relatively little effect on the index before this (Figure 16).

Alternative smoothers

Changing the smoother from *te()* to *ti()* slightly changed the trend after 2013 but had little effect on the index before this (Figure 17).

Changes to preparation of indices

Abundance adjustment for ocean area

Including the ocean area had negligible impact on the index (Figure 18).

Criteria for including spatial strata

Changing the criterion for the minimum number of strata per area slightly reduced the number of strata available to estimate trends (Table 2). Changes had only a small effect on the indices (Figure 19), suggesting that strata with few samples do not have a large effect on the model results.

Table 2: Alternative minimum strata criteria for including spatial cells, and the resulting number of strata in the models.

Minimum strata per area	Total number of strata	Prediction strata per year
1	4344	672
5	4307	552
10	4183	450
15	4089	396
20	3921	336
25	3793	300

ML/GCV versus REML

The optimization criterion used in the model had relatively little impact on the resulting index (Figure 2), since the indices fitted using different methods largely overlaid one another.

Model selection

Model selection was explored for the different arrangements of smoothers (Table 3 and 4). The more flexible smoothed models fitted the data considerably better than the categorical variable

models in both the binomial (Table 5) and lognormal (Table 6) components, and also in most cases had many fewer parameters. The three best fitting delta models according to AIC fitted using the *te()* approach were models 14, 7, and 11. The three best-fitting lognormal positive models according to AIC were models 11, 12, and 14.

The AIC values from the *ti()* approach were slightly but not substantially higher than those from the *te()* approach with similar models (Tables 7 and 8). The delta shrinkage model 15s effectively removed the terms *ti(mon, lat)* and *ti(lat, lon, mon)*, so an additional non-shrinkage model 16 was run that omitted those terms. The best fitting *ti()* models were 16, 15s and 15 for the delta component, with models 11 and 14 also fitting relatively well. For the positive component models 15s fitted best, followed by models 15, 11 and 14.

Delta and positive models were combined by multiplication at the stratum level to generate predicted catch rates. AIC values were added to provide an overall score for each index (Tables 9 and 10).

Predictions were also assessed using the extreme value prediction diagnostic. Almost all extreme value predictions were in the last 7 years, with the largest effect in 2018. Data are increasingly sparse towards the end of the series (Figure 21). Most smoothed models had relatively few extreme values that were small compared to the sum of the predicted values. Each year has 396 (Table 2) predicted values (Figure 22), with annual averages between 0.78 and 4.88. For example, the sum of all predicted values from the delta14pos11 model for 2018 was 3718, which an excess prediction of 12 in that year would change by 0.3%. When extreme values were removed by limiting the maximum predicted values of all strata by year to the maximum observed CPUE in the year, indices for the model delta14pos11 were largely unchanged (Figure 23).

Tables of parameter estimates and statistical significance for the factors and statistical significance for the smooth terms are provided for the best-fitting delta (Tables 11 and 12) and positive (Tables 13 and 14) models.

Indices were plotted for the 6 combined delta lognormal models with the lowest total AIC values (Figure 24). All models were largely unaffected by extreme values.

Discussion

Comparison of model distribution diagnostics suggests that the delta lognormal model should be preferred to other distributions. The apparently poor fit of the lognormal positive component is largely due to the aggregation of the data, which affects all models of the positive component. This is difficult to avoid when working with aggregated data. Models of the positive component assume either homoscedasticity, i.e. constant variance (lognormal model) or residual variance positively correlated with the mean (negative binomial, Gamma and Tweedie), whereas in this dataset residual variance is negatively correlated with mean CPUE.

The models that use the *ti()* approach have been preferred to those using *te()*, which duplicate some terms and tended to be less stable than the *ti()* models. The *ti()* approach is recommended by the author of the mgcv package. Nevertheless, the *te()* models generally had slightly lower AIC values than the corresponding *ti()* models. It may be rewarding to explore why this occurs.

The best fitting index used model 16 for the delta component and model 15s for the positive component. Almost all the difference in trend comes from the positive component, so this index completely overlaid the next best model which used model 15s for both components. Indices from

models 15 (both components) and delta 16 pos 15 were also very similar to the best fitting index. Small differences in the indices were observed for models 14 and 11, which fitted slightly worse with delta AIC of 66.4 and 69.2 respectively.

Model selection is challenging and unreliable with aggregated CPUE data, given the violation of distribution assumptions, the limited variables available, and spatial and temporal autocorrelation. AIC is used as an indicator but is not a particularly reliable as a criterion for weighting models of aggregated CPUE data. A reasonable approach may be to select a group of models that fit the data comparably well and to give them similar weights as alternative outcomes.

Most of the variability among model results occurs at the end of the time series, when data are sparse, and predictions depend on assumptions about what happens in locations with little information. As such, this variability among models seems to represent real uncertainty. The variability between model structures may better represent true uncertainty than parameter uncertainty.

Most of the approaches used in this analysis can be applied to analyses of operational data. The use of *ti()* terms is recommended, as are the use of the extreme value diagnostic to check index predictions, and DHARMa residuals to check model fits. The delta lognormal model is likely to be effective for operational data, but it would be also useful to explore other distributions such as the Tweedie, Gamma, and negative binomial.

Bibliography

Bellego, C., Pape, L.-D. (2019) Dealing with logs and zeros in regression models. *Série des Documents de Travail*(2019-13).

Campbell, R., Tuck, G., Tsuji, S., Nishida, T. (1996) Indices of abundance for southern bluefin tuna from analysis of fine-scale catch and effort data. *Second CCSBT Scientific Meeting*, Hobart, Australia, August 26 - September 6 1996.

Campbell, R.A. (2004) CPUE standardisation and the construction of indices of stock abundance in a spatially varying fishery using general linear models. *Fisheries Research*, 70(2-3): 209-227.

CCSBT (2020) Report of the Twenty Fifth Meeting of the Scientific Committee.

Fasiolo, M., Nedellec, R., Goude, Y., Capezza, C., Wood, S.N. (2020) mgcViz: Visualisations for generalized additive models. *Computer software]. https://CRAN. R-project. org/package= mgcViz*.

Hartig, F. (2020) DHARMa: residual diagnostics for hierarchical (Multi-Level/Mixed) regression models https://CRAN.R-project.org/package=DHARMa

Hoyle, S.D. (2020) Investigation of potential CPUE model improvements for the primary index of Southern Bluefin Tuna abundance, CCSBT-ESC/2008/29. *CCSBT Extended Scientific Committee 25*: 33.

Hoyle, S.D., Langley, A.D. (2020) Scaling factors for multi-region stock assessments, with an application to Indian Ocean tropical tunas. *Fisheries Research*, 228: 105586.

Itoh, T., Takahashi, N. (2019) Update of the core vessel data and CPUE for southern bluefin tuna in 2019. *CCSBT Extended Scientific Committee*.

Kim, Y.J., Gu, C. (2004) Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2): 337-356.

Lo, N.C.H., Jacobson, L.D., Squire, J.L. (1992) Indices of relative abundance from fish spotter data based on delta-lognormal models. *Canadian Journal of Fisheries and Aquatic Sciences*, 49(12): 2515-2526.

Maunder, M.N., Punt, A.E. (2004) Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, 70(2-3): 141-159. 10.1016/j.fishres.2004.08.002

Nishida, T., Tsuji, S. (1998) Estimation of abundance indices of southern bluefin tuna (*Thunnus maccoyii*) based on the coarse scale Japanese longline fisheries data (1969-97). *Fourth CCSBT Scientific Meeting*, Shimizu, Shizuoka, Japan, July 23-31, 1998.

Punsly, R. (1987) *Estimation of the relative annual abundance of yellowfin tuna, Thunnus albacares, in the eastern Pacific Ocean during 1970-1985*. Inter-American Tropical Tuna Commission, La Jolla, CA.

Wood, S. (2019) MGCV: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version, 1.8-31.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1): 3-36.

Tables

Table 3: Specifications for models run using mgcv. The factors column reports all variables included as categorical variables. Smooth terms include two-way, three-way, and four-way ('All') interactions.

Label	**Factors**	**Smooth terms**								**4-way**
Base										
Base plus	.+mf:latf									
Base_noYrAr	.-yf:areaf									
glmmYrAr										
gam 2	yf+mf	lon,lat								
gam 3	yf+mf							mn,lon,lat		
gam 4	yf									All
gam 5	yf+llf									All
gam 6	yf+llf		mn,lat							All
gam 7	yf	lon,lat	mn,lat							All
gam 8	yf	lon,lat	mn,lat	lon,mn						All
gam 9	yf	lon,lat	mn,lat	lon,mn	yr,lat			mn,lon,lat	lat,lon,yr	
gam 10	yf	lon,lat	mn,lat	lon,mn	yr,lat	yr,lon	yr,mn			
gam 11	yf	lon,lat	mn,lat	lon,mn	yr,lat		yr,mn	lat,lon,mn	lat,lon,yr	
gam 12	yf	lon,lat	mn,lat	lon,mn	yr,lat	yr,lon	yr,mn	lat,lon,mn	lat,lon,yr	
gam 13	yf+mf	lon,lat	mn,lat	lon,mn	yr,lat	yr,lon				
gam 14	yf	lon,lat	mn,lat	lon,mn	yr,lat	yr,lon,mn	yr,mn	lat,lon,mn	lat,lon,yr	

Table 4: Specifications for models fitted with smooth terms specified using ti(). The factors column reports all variables included as categorical variables. Smooth terms include one-way terms and two-way and three-way interactions. All models include categorical variables for year (yf) and one-way smooth terms for longitude (lon), month (mn), and latitude (lat). Within interactions year is specified as a smooth term (yr). Terms marked 0 were included but effectively removed via shrinkage.

Model	Factors	lon, lat	mn, lat	lon, mn	yr, lat	yr, lon	yr, mn	yr, lon, mn	lat, lon, mn	lat, lon, yr	lat, mn, yr
9	yf	x	x	x	x	x			x	x	
10	yf	x	x	x	x	x	x				
11	yf	x	x	x	x	x	x		x	x	
13	yf+mf	x	x	x	x	x					
14	yf	x	x	x	x	x	x	x	x	x	
15	yf	x	x	x	x	x	x	x	x	x	x
15s	yf	x	0	x	x	x	x	x	0	x	x
16	yf	x		x	x	x	x	x		x	x

Table 5: Comparison of delta models fitted using te() after optimisation using maximum likelihood, to compare main effects. The degrees of freedom for smoother models is the estimated degrees of freedom (edf). The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model.

Delta	edf	AIC	ΔAIC
Base	259	1882.8	440.1
BasePlus	274	1897.7	455.0
Base_noYrArea	160	1840.4	397.7
glmm_YrArea	151.0	1822.5	379.8
1	277.8	1687.2	244.5
2	63.6	1569.2	126.5
3	84.8	1569.2	126.4
4	105.5	1534.2	91.5
5	173.5	1579.5	136.7
6	104.1	1512.0	69.3
7	94.5	1446.0	3.3
8	98.3	1473.6	30.9
9	89.3	1491.4	48.6
10	94.8	1479.6	36.9
11	95.1	1457.1	14.4
12	92.6	1469.7	27.0
13	81.8	1488.2	45.4
14	91.1	1442.7	0.0

Table 6: Comparison of lognormal positive models after optimisation using generalized cross-validation (GCV), to compare main effects. The degrees of freedom for smoother models is the estimated degrees of freedom (edf). The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model.

Pos	edf	AIC	ΔAIC
Base	255	10483.5	1486.3
BasePlus	270	10425.3	1428.1
Base_noYrArea	157	10531.9	1534.6
glmm_YrArea	166.5	10507.5	1510.3
1	311.5	9344.9	347.7
2	96.9	10240.5	1243.3
3	149.7	9491.6	494.4
4	225.3	9419.4	422.1
5	250.7	9180.9	183.6
6	244.1	9125.4	128.2
7	191.1	9125.1	127.9
8	200.5	9108.3	111.1
9	197.3	9090.0	92.8
10	171.9	9200.8	203.6
11	217.8	8997.2	0.0
12	217.9	9018.2	21.0
13	167.3	9197.0	199.8
14	230.0	9078.1	80.9

Table 7: Comparison of delta models fitted using ti, optimized with maximum likelihood (ML) or restricted maximum likelihood (REML). The degrees of freedom for smoother models is the estimated degrees of freedom (edf). The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model.

Model		ML			REML		
Delta	Shrinkage	edf	AIC	ΔAIC	edf	AIC	ΔAIC
9	N	80.8	1505.2	59.5	94.5	1478.9	58.9
10	N	86.6	1485.5	39.8	97.7	1463.6	43.5
11	N	89.4	1468.5	22.8	97.8	1441.5	21.4
14	N	89.5	1465.8	20.2	98.4	1444.9	24.8
15	N	93.2	1447.5	1.9	102.8	1420.1	0.0
15s	Y	108.4	1447.4	1.7	122.0	1429.7	9.6
16	N	91.2	1445.7	0	99.6	1431.9	11.9

Table 8: Comparison of lognormal positive models, optimised using optimized with generalized cross-validation (GCV) or restricted maximum likelihood (REML). The degrees of freedom for smoother models is the estimated degrees of freedom (edf). The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model.

		GCV			REML		
Pos	shrinkage	edf	AIC	ΔAIC	edf	AIC	ΔAIC
9	N	202.0	9292.3	65.8	210.8	9309.3	56.9
10	N	178.9	9468.5	242.0	185.9	9482.1	229.7
11	N	225.7	9272.9	46.4	214.2	9302.3	49.9
14	N	215.9	9272.7	46.3	227.8	9296.5	44.0
15	N	224.5	9269.4	43.0	238.4	9286.6	34.2
15s	Y	224.8	9226.5	0	226.8	9252.4	0.0

Table 9: Sums of delta and lognormal AIC values, and extreme value diagnostic results for the te() smoothed models, optimized using maximum likelihood (delta) and generalized cross-validation (positive). The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model. Models with an asterisk have ΔAIC less than 100.

Delta	Pos	Total AIC	ΔAIC	max	diff	diff^2
1	1	11032.1	592.2	16	35.9	168.0
2	2	11809.7	1369.8	57	53.2	91.8
3	3	11060.8	620.9	0	0.0	0.0
4	4	10953.6	513.7	38	420.2	11338.8
5	5	10760.4	320.5	6	8.9	22.0
6	6	10637.4	197.5	20	59.1	266.9
7	7	10571.1	131.2	32	76.2	350.4
8	8	10581.9	142.0	2	0.8	0.4
9	9	10581.4	141.5	0	0.0	0.0
10	10	10680.4	240.5	5	3.6	3.5
11	11	10454.3	14.4	13	16.8	32.8
12	12	10487.9	48.0	14	28.2	80.0
13	13	10685.2	245.3	6	7.6	12.2
14	14	10520.8	80.9	0	0.0	0.0
14	11	10439.9	0.0	13	17.0	33.1

Table 10: Sums of delta and lognormal AIC values, and extreme value diagnostic results for the ti() smoothed models. The delta AIC (ΔAIC) is the difference between the individual model's AIC and that of the best fitting model. Models with an asterisk have ΔAIC less than 100.

Delta	Pos	ML		REML				
		Total AIC	ΔAIC	Total AIC	ΔAIC	max	diff	diff^2
9	9	10797.5	125.3	10788.2	106.1	0	0	0
10	10	10953.9	281.8	10945.7	263.6	13	15.8	28.3
11	11	10741.4	69.2	10743.8	61.7	15	34.9	140.3
14	14	10738.6	66.4	10741.4	59.3	12	5.2	3.5
15	15	10717.0	44.8	10706.6	24.5	5	5.2	6.9
15s	15s	10673.9	1.7	10682.1	0.0	3	1.9	2.2
16	15	10715.1	43.0	10718.5	36.4	5	5.2	6.9
16	15s	10672.2	0.0	10684.4	2.3	3	1.9	2.2

*Table 11: Table of parameter estimates, standard errors and statistical significance for the categorical year factors in delta model 15. The final column includes codes to indicate significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	8.08	1.00	8.11	5.20E-16	***
yf1987	0.03	0.58	0.06	9.54E-01	
yf1988	-1.58	0.64	-2.46	1.39E-02	*
yf1989	-1.86	0.73	-2.54	1.12E-02	*
yf1990	-2.50	0.81	-3.09	2.02E-03	**
yf1991	-2.41	0.81	-2.96	3.04E-03	**
yf1992	-1.54	0.87	-1.77	7.70E-02	.
yf1993	-2.08	0.88	-2.36	1.84E-02	*
yf1994	-4.73	0.89	-5.30	1.16E-07	***
yf1995	-3.98	0.92	-4.33	1.46E-05	***
yf1996	-2.91	0.93	-3.14	1.72E-03	**
yf1997	-2.32	0.96	-2.42	1.54E-02	*
yf1998	-2.49	0.96	-2.58	9.80E-03	**
yf1999	-1.67	0.98	-1.71	8.78E-02	.
yf2000	-2.96	0.99	-2.99	2.79E-03	**
yf2001	-3.06	1.00	-3.06	2.19E-03	**
yf2002	-2.50	1.08	-2.31	2.10E-02	*
yf2003	-0.11	1.12	-0.10	9.19E-01	
yf2004	-1.15	1.06	-1.09	2.77E-01	
yf2005	-2.24	1.01	-2.21	2.70E-02	*
yf2006	-2.21	1.06	-2.08	3.73E-02	*
yf2007	-0.67	1.06	-0.63	5.31E-01	
yf2008	-2.25	1.03	-2.18	2.90E-02	*
yf2009	-1.78	1.06	-1.68	9.24E-02	.
yf2010	-1.53	1.03	-1.49	1.37E-01	
yf2011	-2.08	1.03	-2.01	4.44E-02	*
yf2012	-2.75	1.01	-2.72	6.50E-03	**
yf2013	-0.52	1.12	-0.46	6.45E-01	
yf2014	-1.43	1.08	-1.32	1.86E-01	
yf2015	-0.53	1.09	-0.48	6.28E-01	
yf2016	-1.04	1.13	-0.92	3.55E-01	
yf2017	-1.44	1.16	-1.24	2.13E-01	
yf2018	-1.58	1.22	-1.30	1.94E-01	
yf2019	-1.16	1.26	-0.92	3.56E-01	
yf2020	-0.64	1.32	-0.48	6.28E-01	

*Table 12: Table of effective degrees of freedom, Chi square and p-values for the smooth terms in delta model 15. The final column includes codes to indicate significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

	edf	Chi.sq	p-value	
ti(lon2)	12.47	86.91	3.76E-12	***
ti(LAT)	2.13	86.23	1.61E-18	***
ti(MONTH)	3.27	91.79	1.34E-18	***
ti(lon2,LAT)	5.77	27.03	1.38E-08	***
ti(MONTH,LAT)	1.96	2.29	3.49E-01	
ti(lon2,MONTH)	1.00	8.20	4.20E-03	**
ti(YEAR,LAT)	3.24	9.09	5.83E-02	.
ti(YEAR,MONTH)	6.95	23.69	7.61E-03	**
ti(lon2,YEAR)	3.18	20.48	5.53E-04	***
ti(LAT,lon2,MONTH)	1.00	0.23	6.30E-01	
ti(YEAR,lon2,MONTH)	1.00	14.15	1.69E-04	***
ti(LAT,lon2,YEAR)	4.23	28.57	1.63E-08	***
ti(LAT,MONTH,YEAR)	5.82	28.36	2.05E-04	***
ti(log(N_HOOKS))	1.48	134.75	1.27E-28	***

*Table 13: Table of parameter estimates, standard errors and statistical significance for the categorical year factors in positive model 15s. The final column includes codes to indicate significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.15	0.07	-2.12	3.44E-02	*
yf1987	0.12	0.08	1.36	1.74E-01	
yf1988	-0.02	0.09	-0.22	8.26E-01	
yf1989	-0.11	0.09	-1.22	2.21E-01	
yf1990	0.02	0.09	0.26	7.93E-01	
yf1991	0.07	0.09	0.77	4.40E-01	
yf1992	0.22	0.09	2.36	1.84E-02	*
yf1993	0.51	0.10	5.26	1.53E-07	***
yf1994	0.58	0.11	5.21	2.03E-07	***
yf1995	0.51	0.11	4.88	1.09E-06	***
yf1996	0.23	0.10	2.24	2.53E-02	*
yf1997	0.08	0.10	0.83	4.04E-01	
yf1998	0.12	0.10	1.26	2.06E-01	
yf1999	0.24	0.10	2.43	1.52E-02	*
yf2000	0.14	0.11	1.32	1.85E-01	
yf2001	0.37	0.11	3.43	6.11E-04	***
yf2002	0.61	0.11	5.39	7.45E-08	***
yf2003	0.38	0.11	3.35	8.14E-04	***
yf2004	0.14	0.11	1.34	1.79E-01	
yf2005	0.10	0.11	0.99	3.23E-01	
yf2006	-0.25	0.11	-2.33	1.99E-02	*
yf2007	-0.40	0.11	-3.72	2.05E-04	***
yf2008	0.17	0.11	1.49	1.36E-01	
yf2009	0.75	0.12	6.34	2.53E-10	***
yf2010	0.70	0.12	6.00	2.21E-09	***
yf2011	0.92	0.12	7.74	1.25E-14	***
yf2012	0.77	0.12	6.45	1.25E-10	***
yf2013	1.05	0.12	8.58	1.38E-17	***
yf2014	1.13	0.13	8.76	3.00E-18	***
yf2015	1.33	0.12	10.95	1.78E-27	***
yf2016	1.12	0.13	8.67	6.51E-18	***
yf2017	1.22	0.14	8.84	1.44E-18	***
yf2018	1.36	0.14	9.45	6.23E-21	***
yf2019	1.38	0.14	9.74	3.88E-22	***
yf2020	1.33	0.13	9.99	3.51E-23	***

*Table 14: Table of effective degrees of freedom, Chi square and p-values for the smooth terms in positive model 15s. The final column includes codes to indicate significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

	edf	F	p-value	
ti(lon2)	20.25	17.14	3.00E-69	***
ti(LAT)	2.94	306.59	7.25E-167	***
ti(MONTH)	4.26	54.75	7.61E-51	***
ti(lon2,LAT)	36.03	9.80	1.30E-160	***
ti(MONTH,LAT)	5.70	16.00	1.11E-75	***
ti(lon2,MONTH)	8.75	1.83	5.55E-17	***
ti(YEAR,LAT)	9.77	0.89	1.17E-15	***
ti(YEAR,MONTH)	7.67	0.34	1.30E-07	***
ti(lon2,YEAR)	6.62	0.59	6.14E-11	***
ti(LAT,lon2,MONTH)	10.17	0.46	1.23E-22	***
ti(YEAR,lon2,MONTH)	2.79	1.84	1.66E-05	***
ti(LAT,lon2,YEAR)	34.63	1.41	3.19E-52	***
ti(LAT,MONTH,YEAR)	4.74	0.26	3.61E-08	***

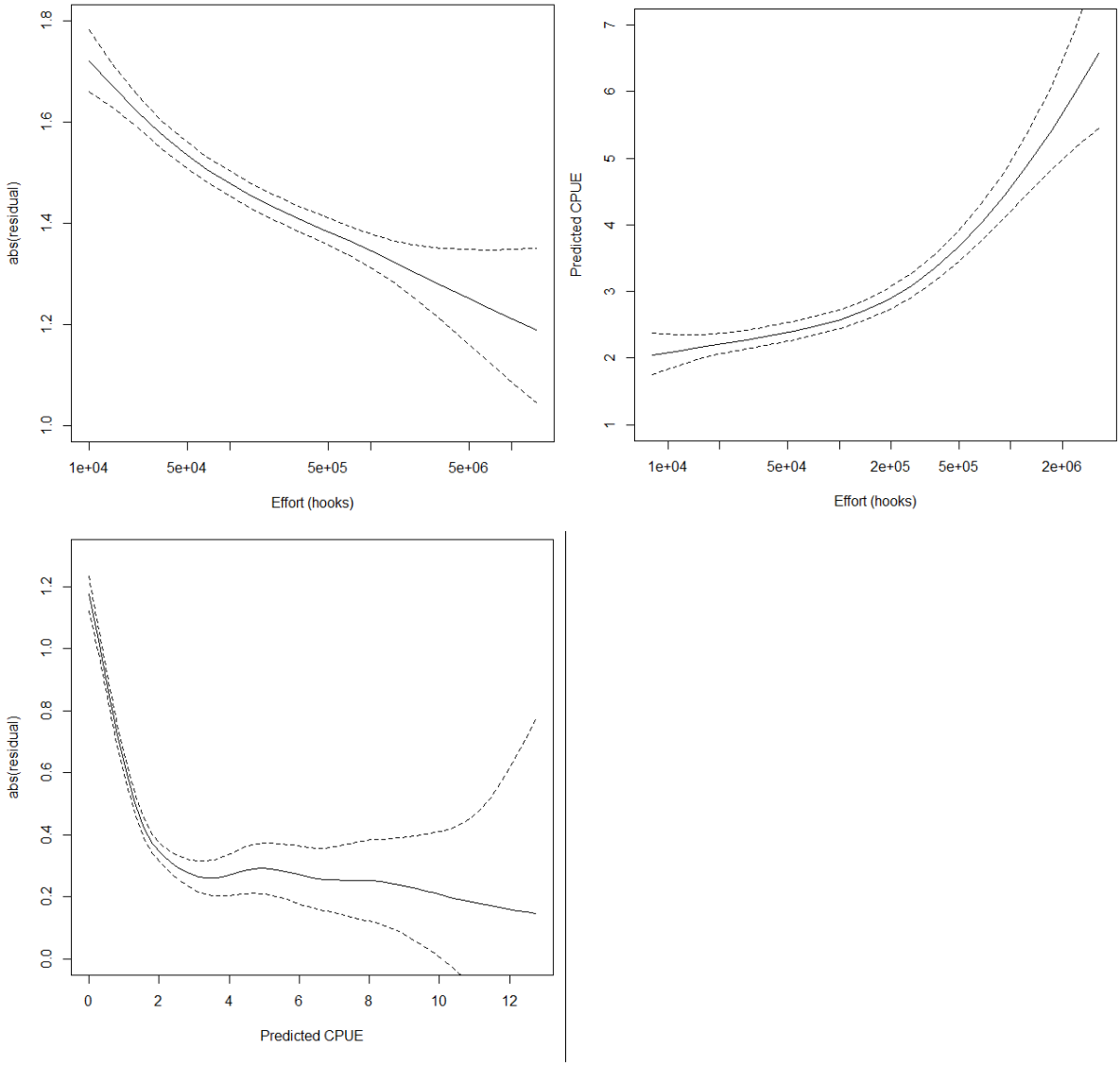

Figure 1: Relationships between effort, residual variance, and CPUE in the aggregated catch and effort data for soutghern bluefin tuna.

Figures

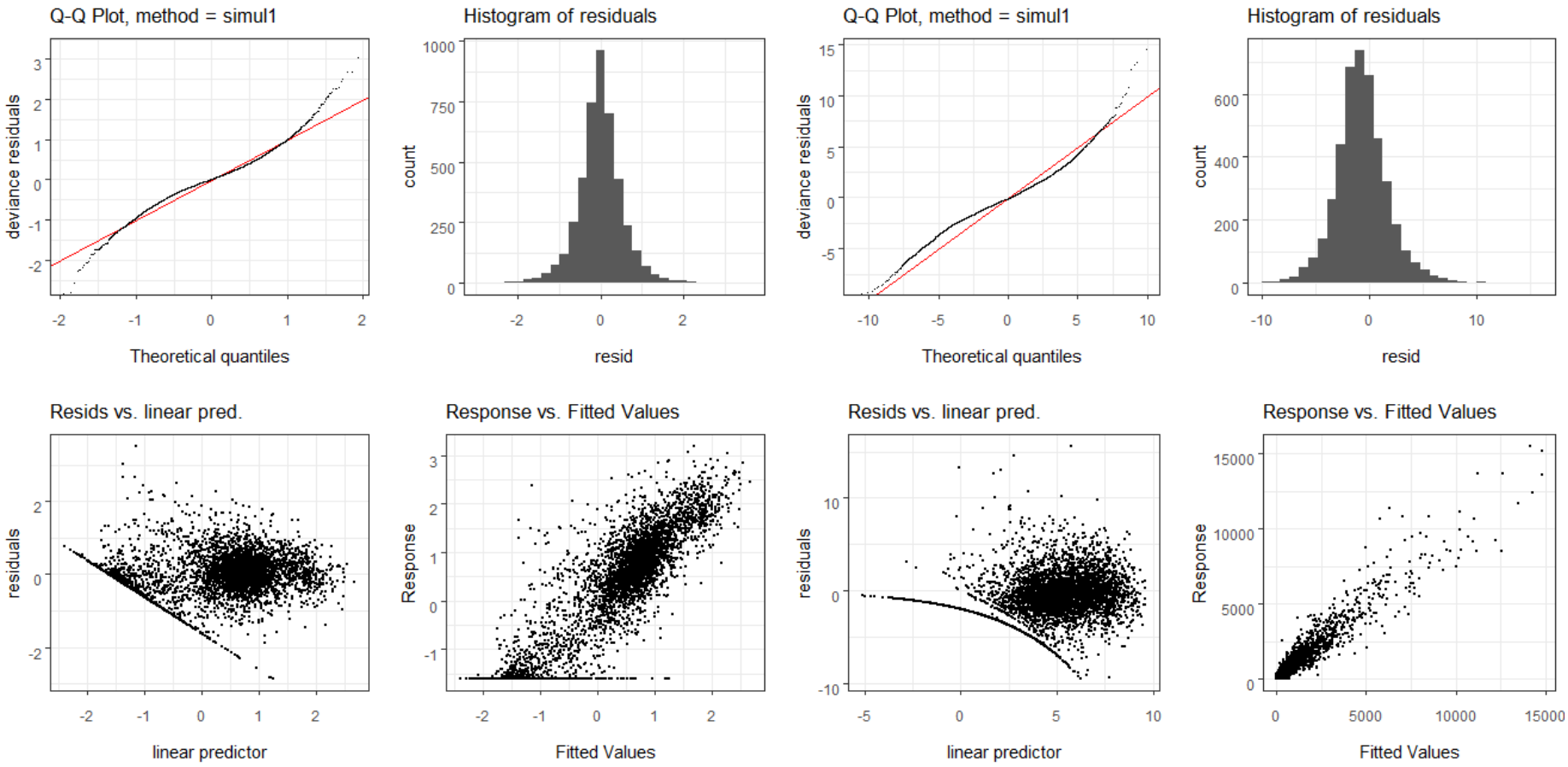

Figure 2: Diagnostic plots for the lognormal constant model and the Tweedie model (right).

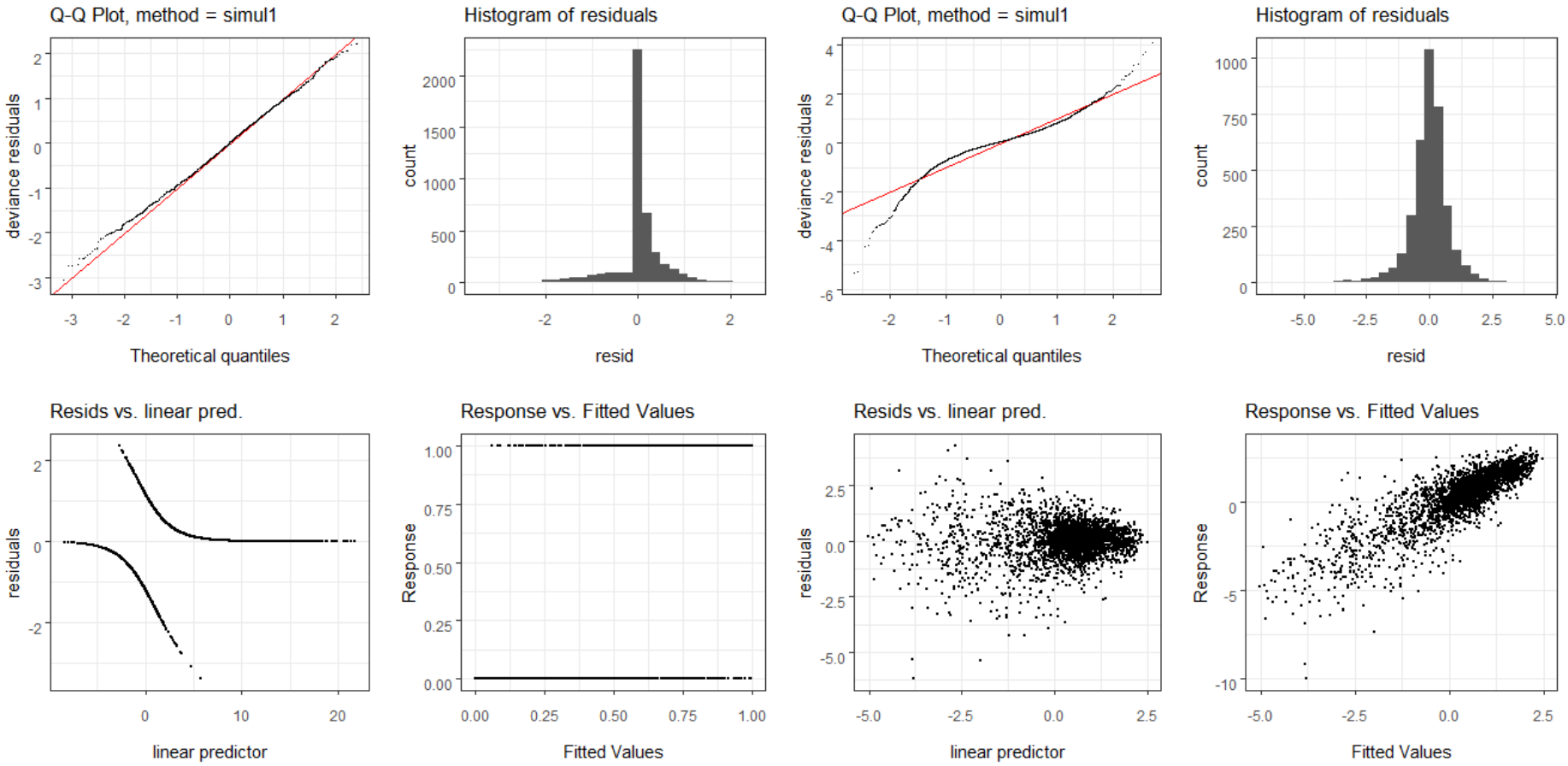

Figure 3: Diagnostic plots for the delta model with logit link (left) and the lognormal model (right).

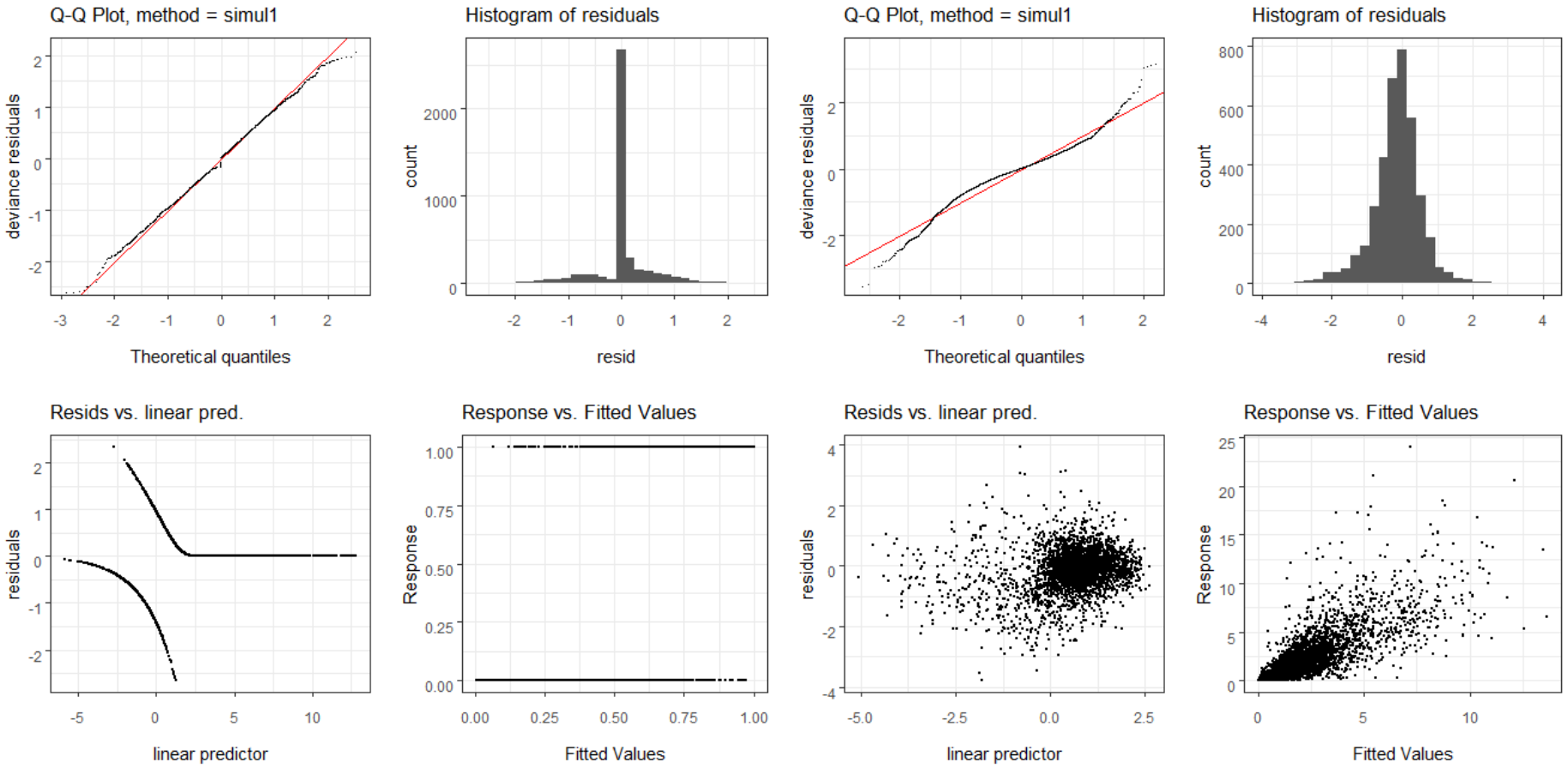

Figure 4: Diagnostic plots for the delta model with cloglog link (left) and the Gamma model (right).

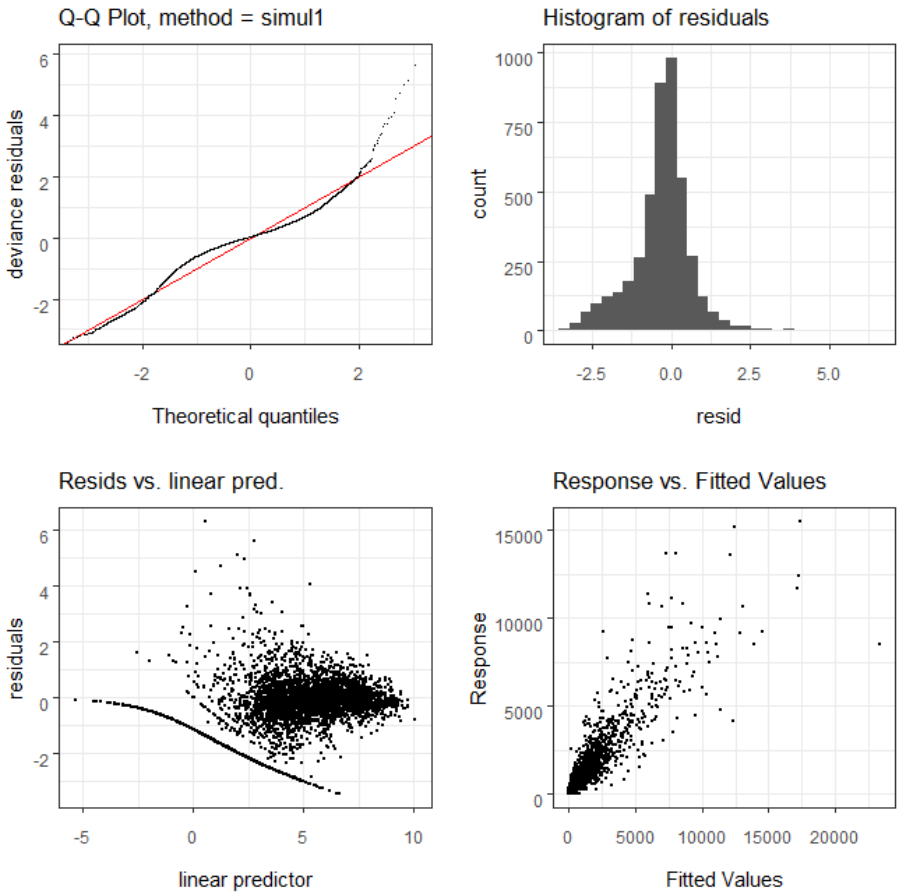

Figure 5: Diagnostic plots for the negative binomial model.

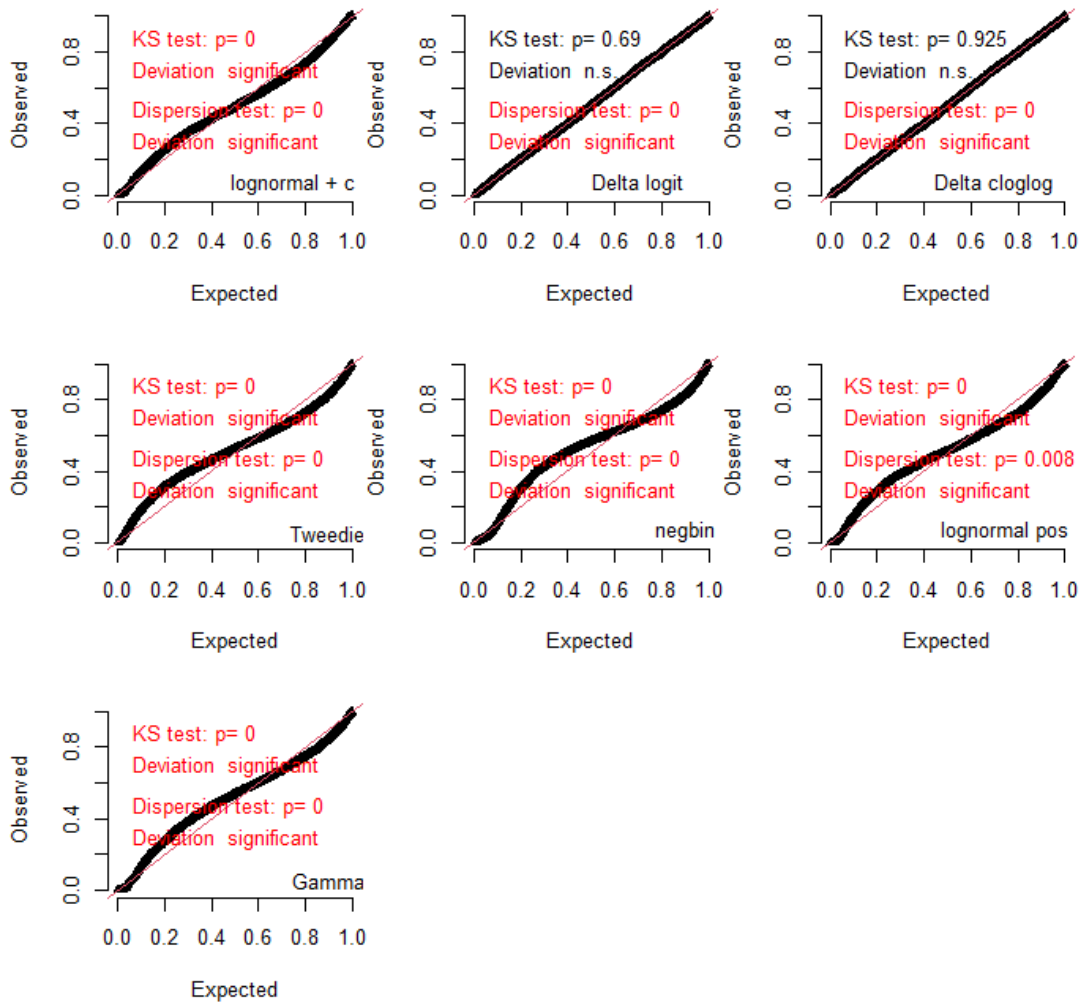

Figure 6: Q-Q plots based on simulated residuals for the models.

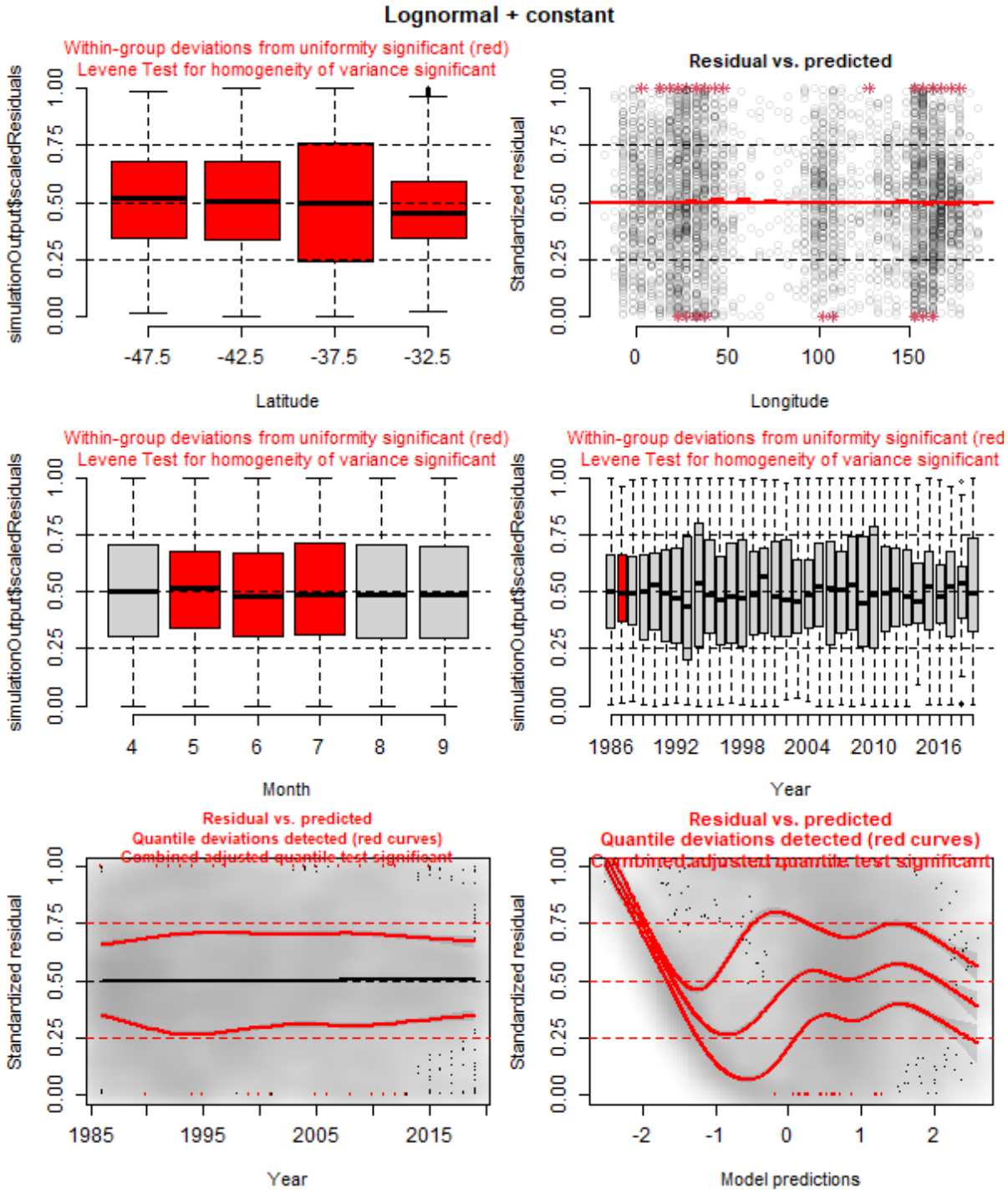

Figure 7: Simulated residuals by covariate for model 11 with the cpue + constant response variable and lognormal distribution.

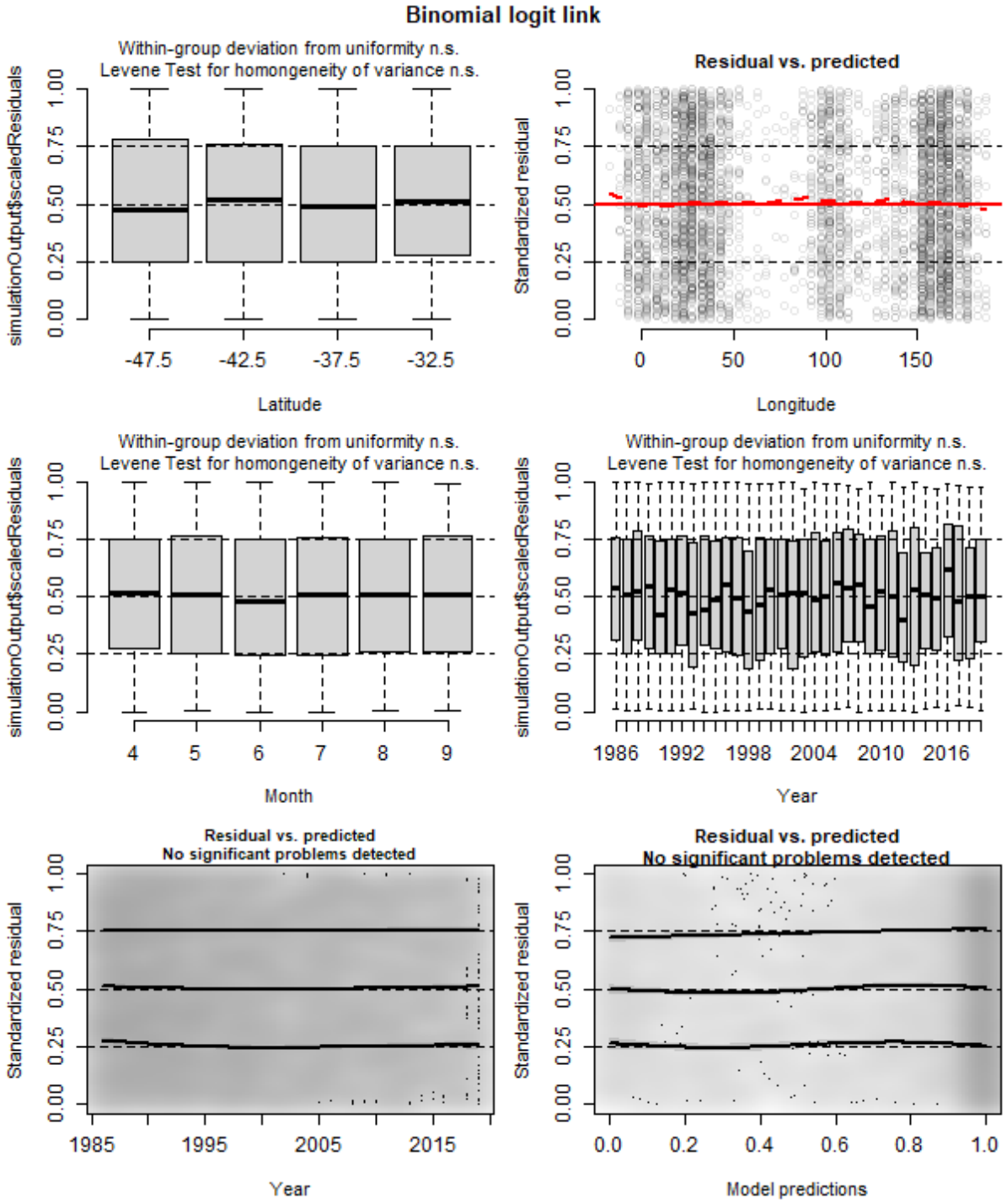

Figure 8: Simulated residuals by covariate for model 11 with the binomial (delta) distribution.

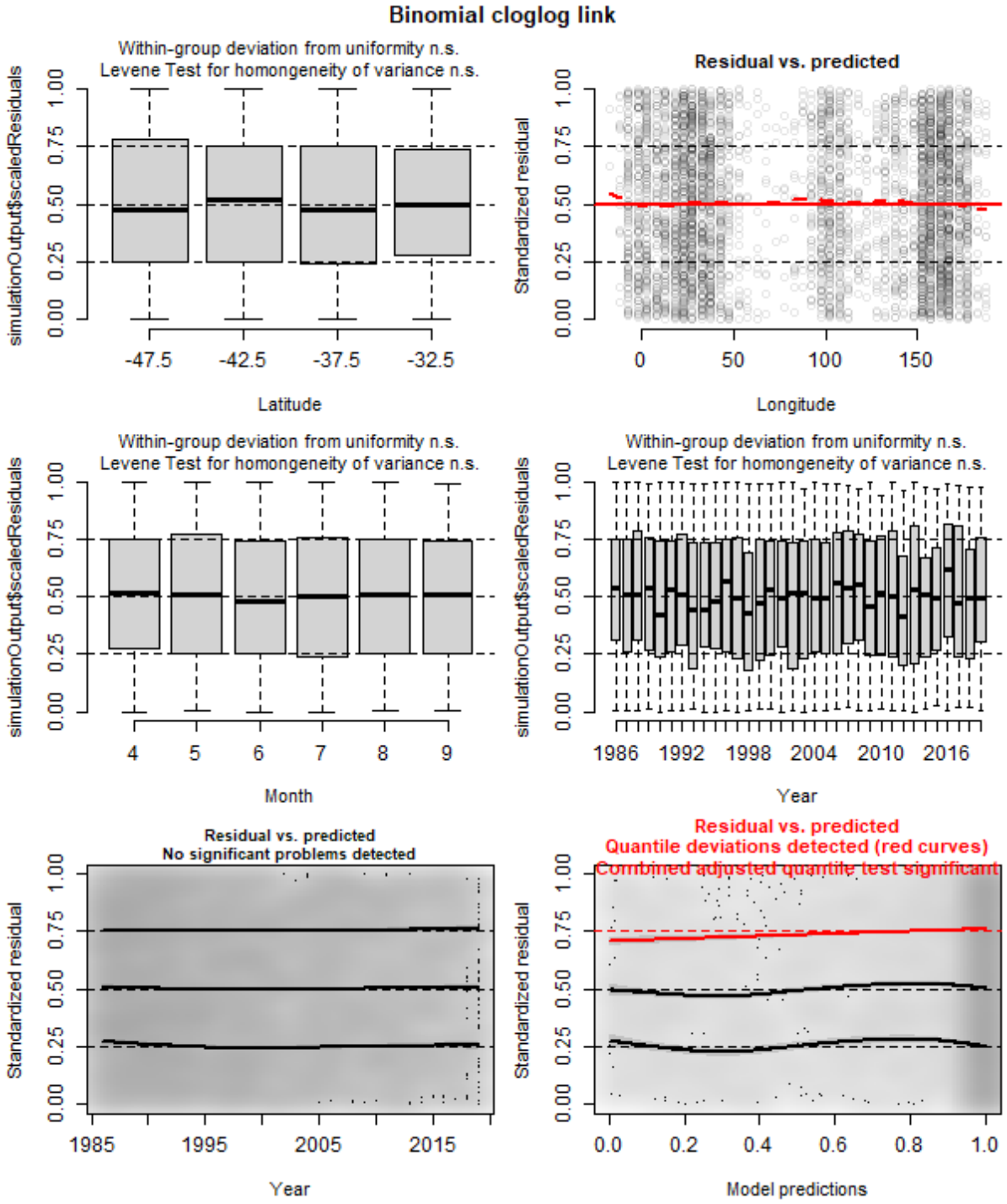

Figure 9: Simulated residuals by covariate for model 11 with the complementary log-log link, as specified for the Delta Poisson distribution.

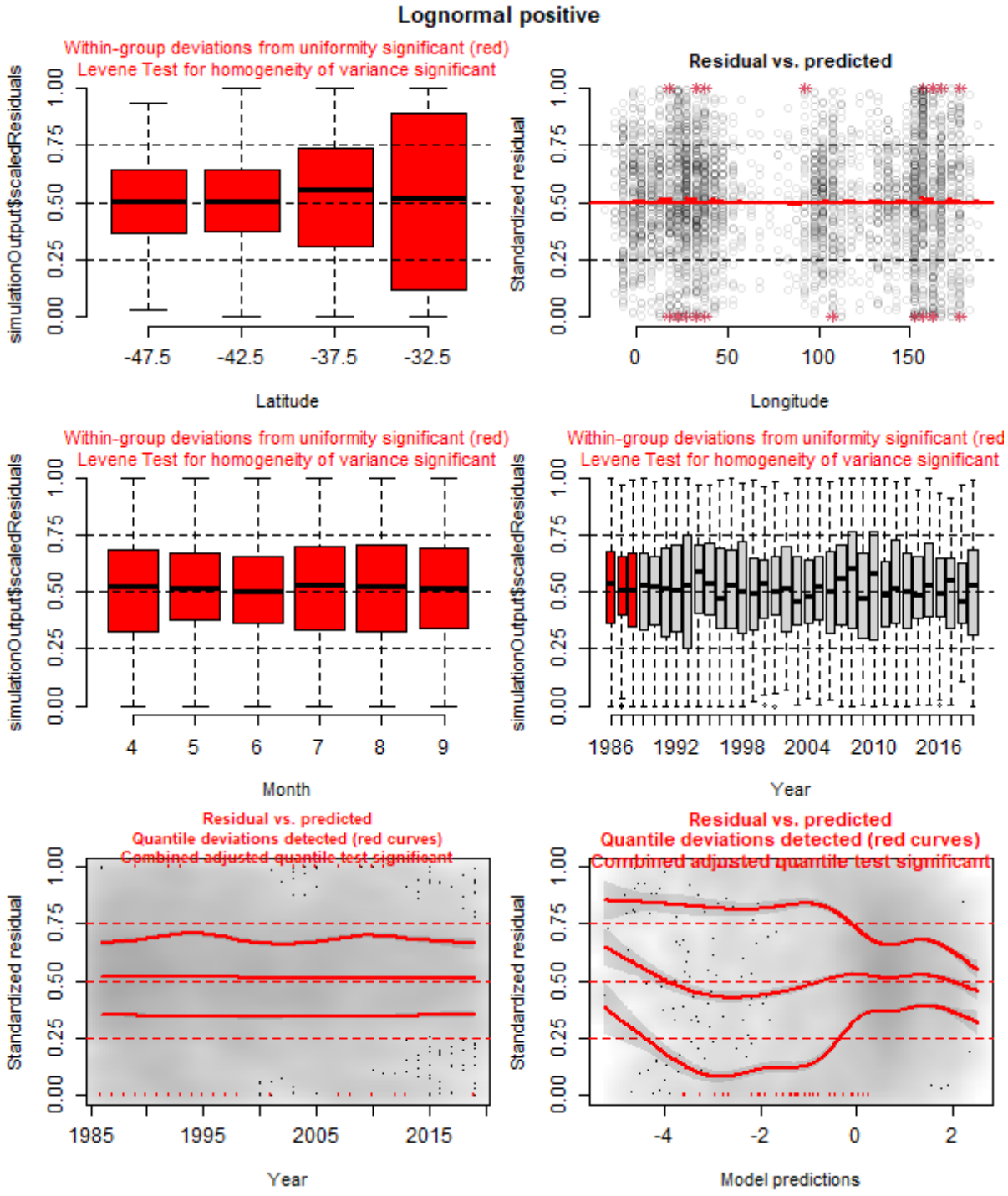

Figure 10: Simulated residuals by covariate for model 11 with the lognormal distribution and nonzero catches.

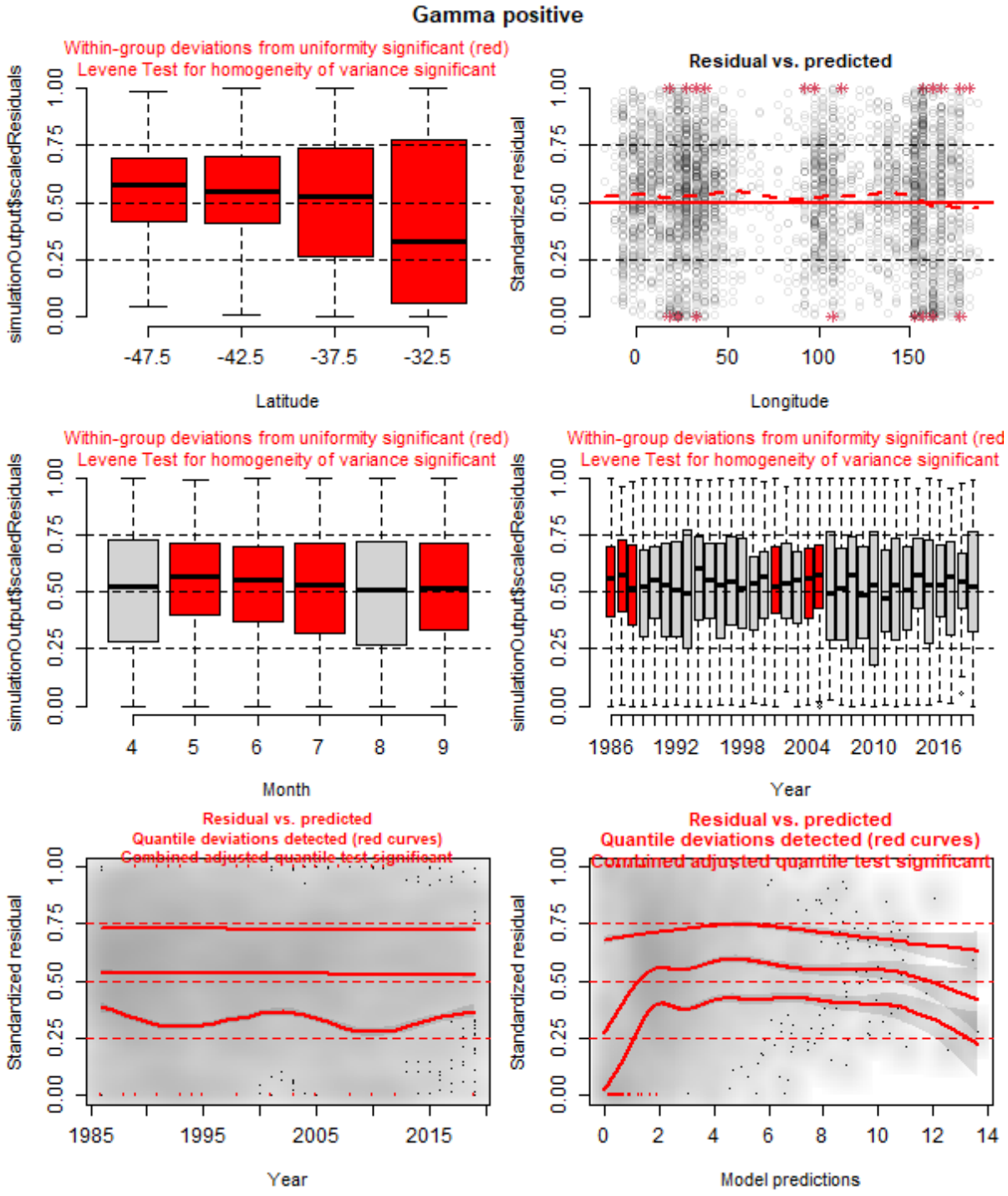

Figure 11: Simulated residuals by covariate for model 11 with the Gamma distribution.

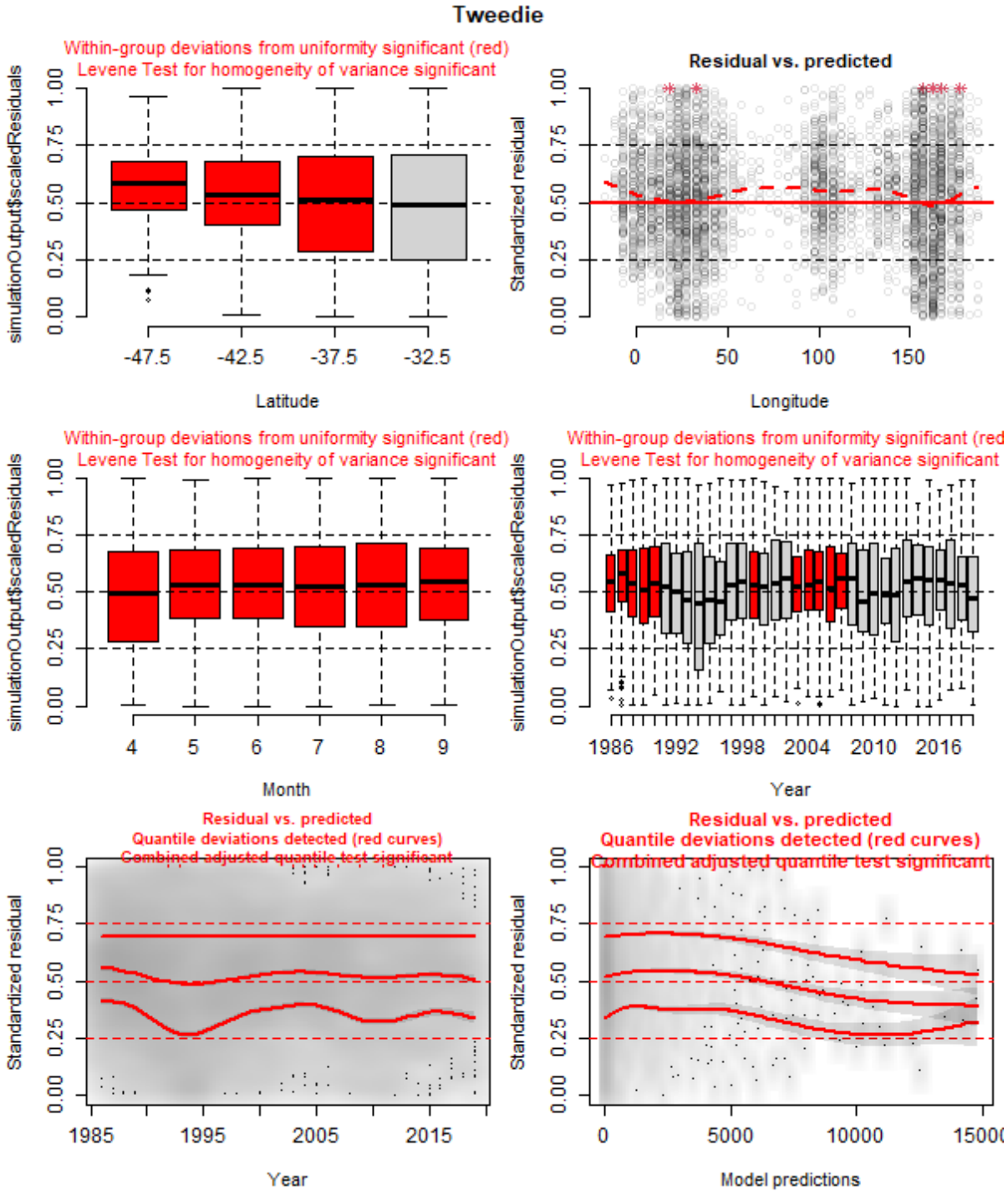

Figure 12: Simulated residuals by covariate for model 11 with the Tweedie distribution.

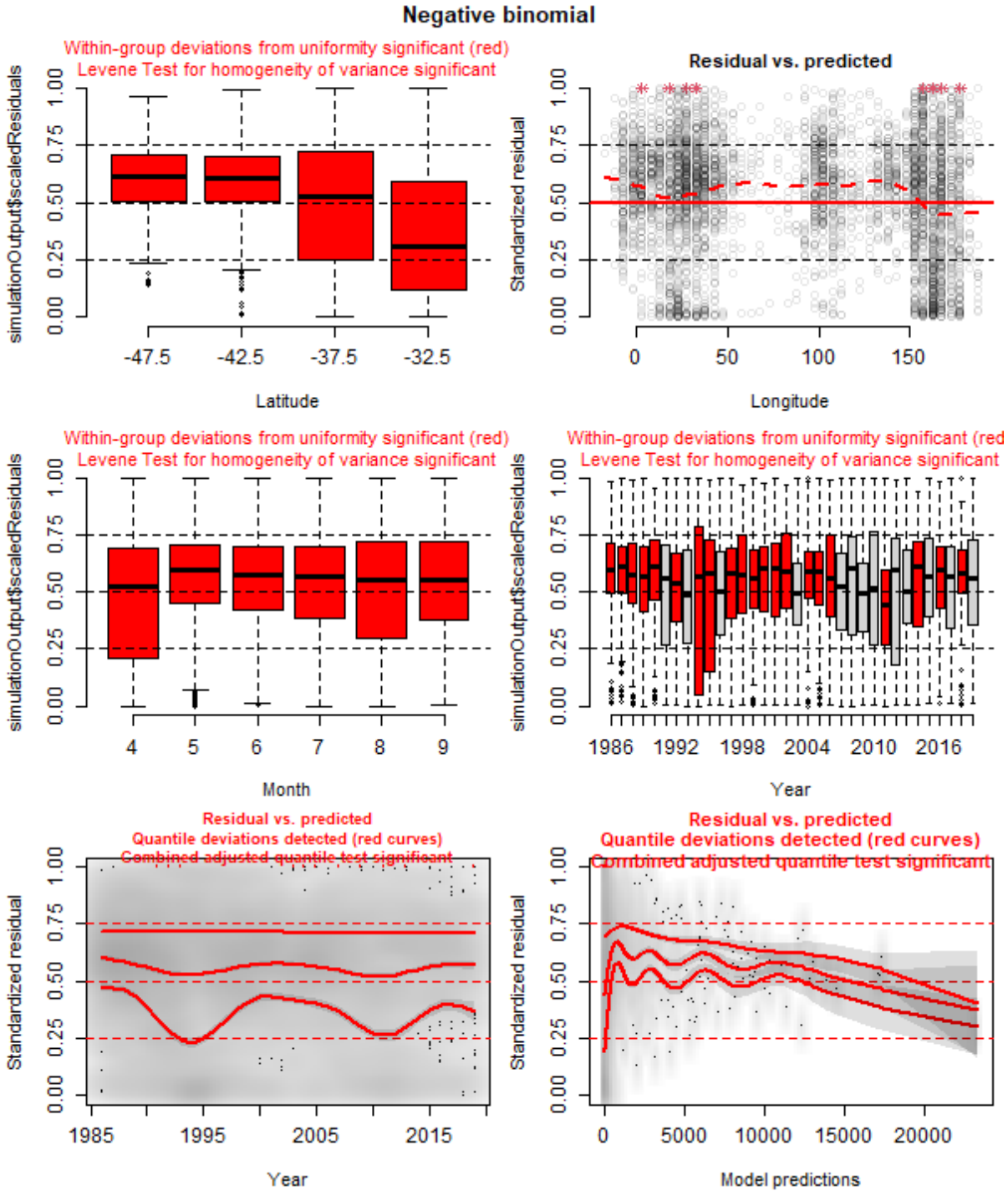

Figure 13: Simulated residuals by covariate for model 11 with the negative binomial distribution.

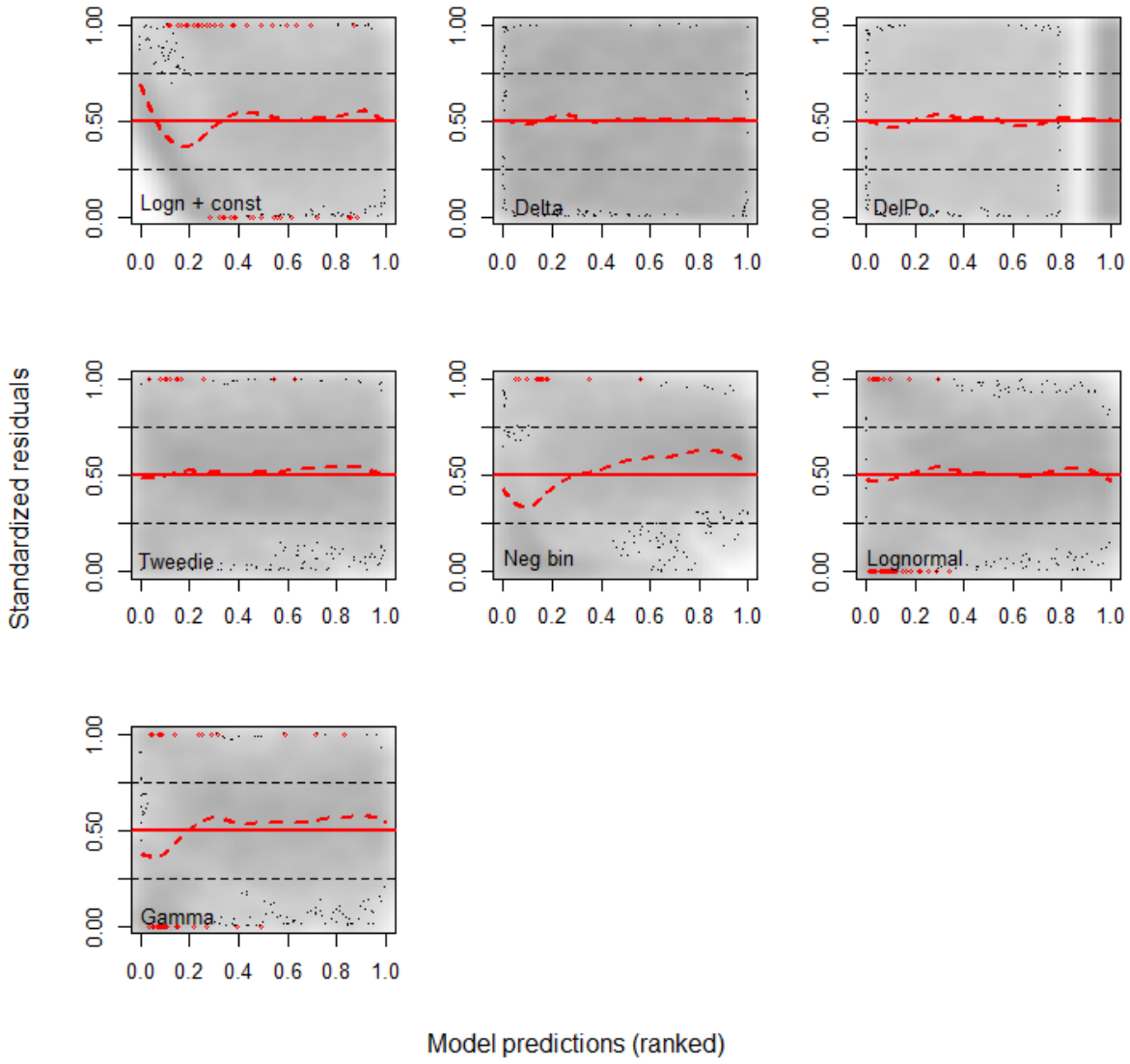

Figure 14: Scaled residual plots for model 11 and all distribution options, with the x-axes representing ranked model predictions, and y-axes representing standardized residuals.

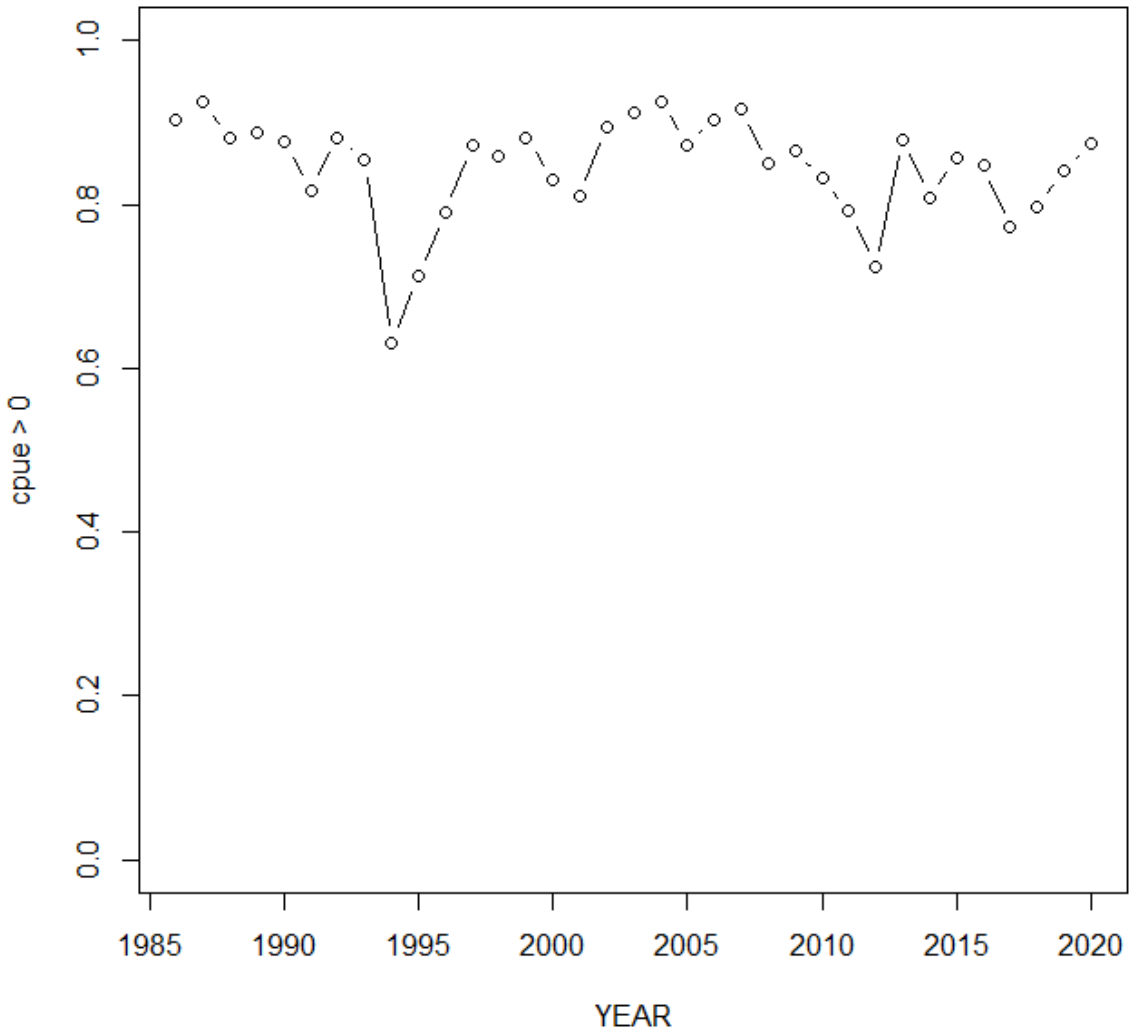

Figure 15: Annual proportion of nonzero catch strata in the reported effort.

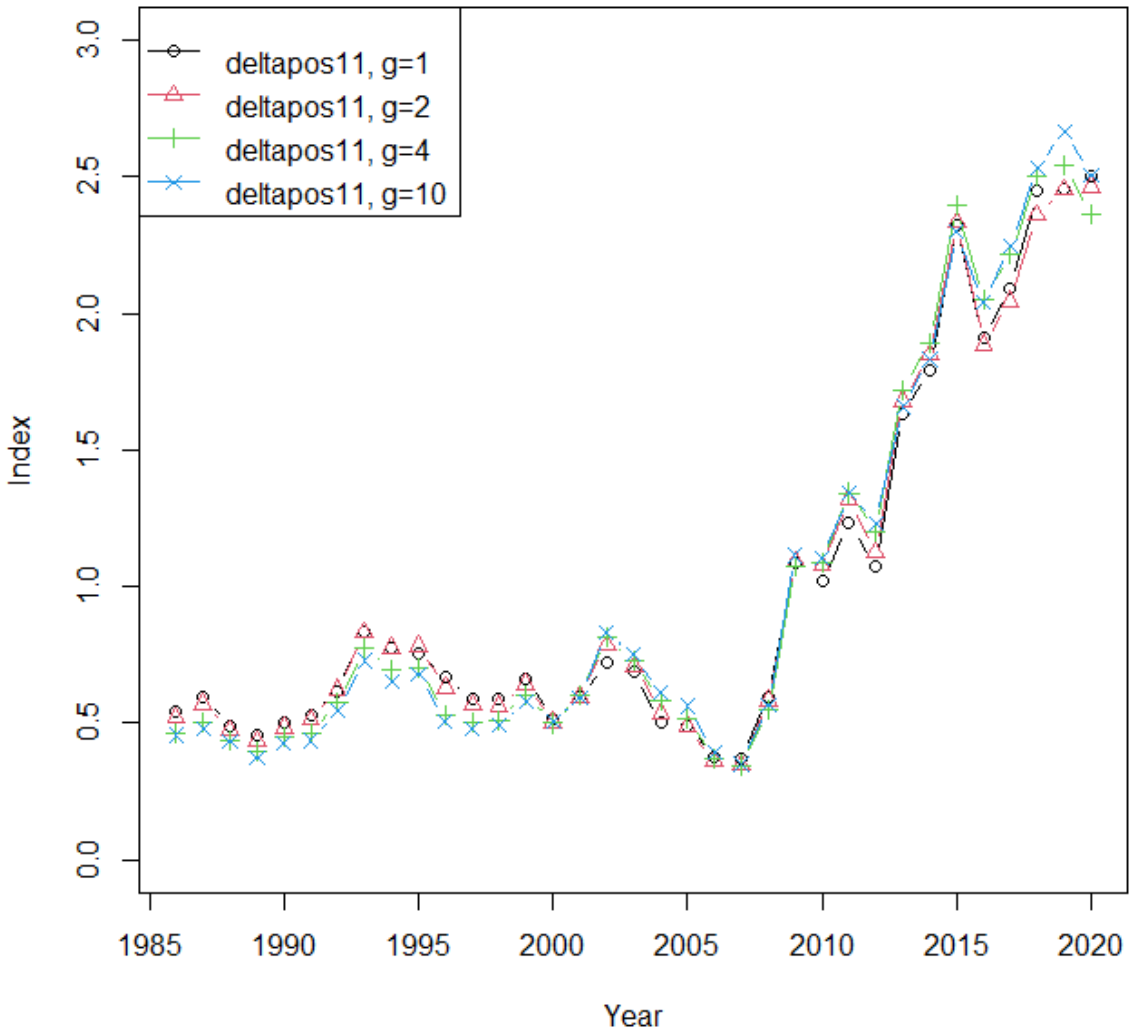

Figure 16: Delta lognormal model 11 indices with different levels of gamma.

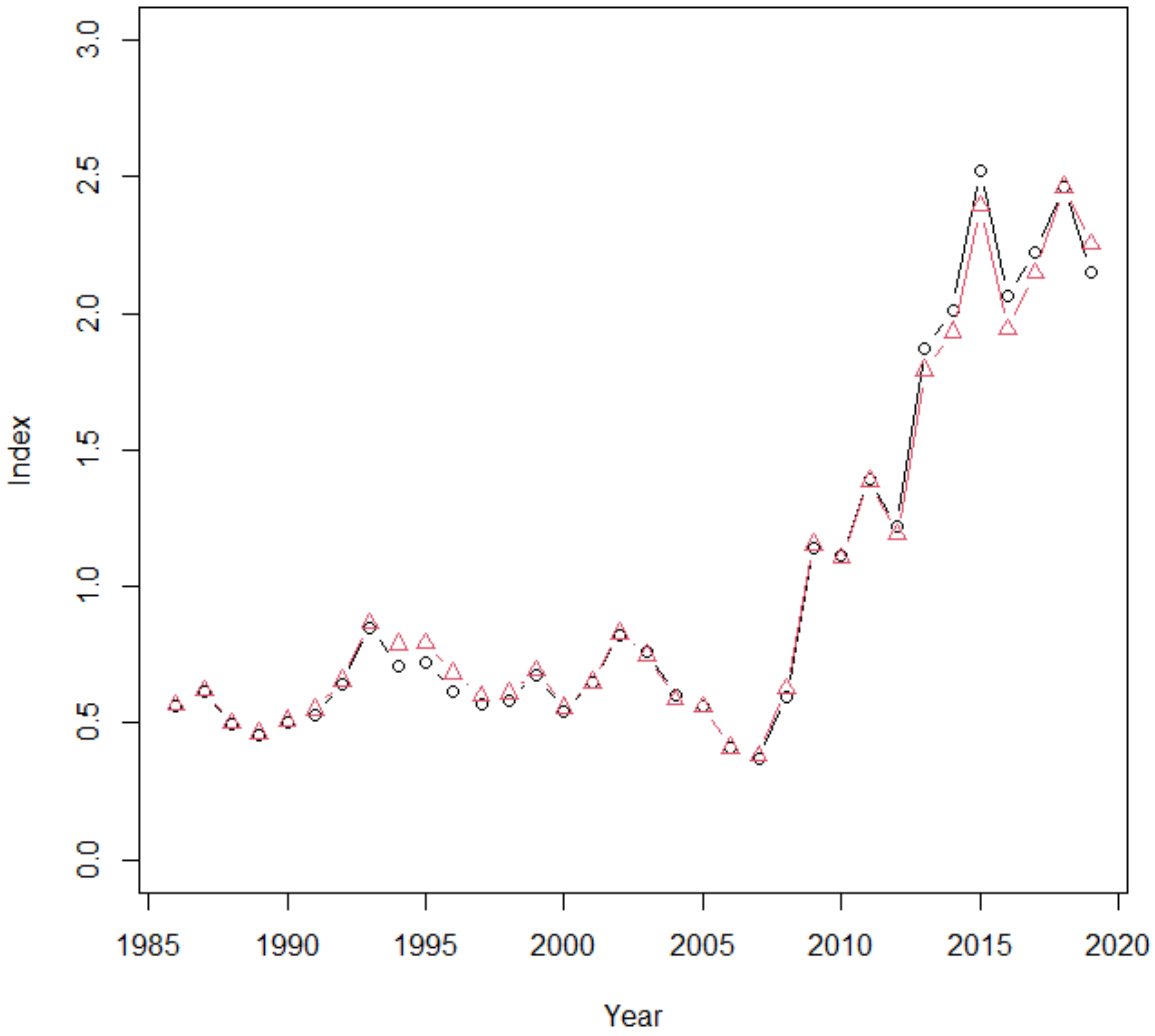

Figure 17: Delta lognormal model 11 indices with either te() smoother on all variables, or the ti() smoother used for interaction terms.

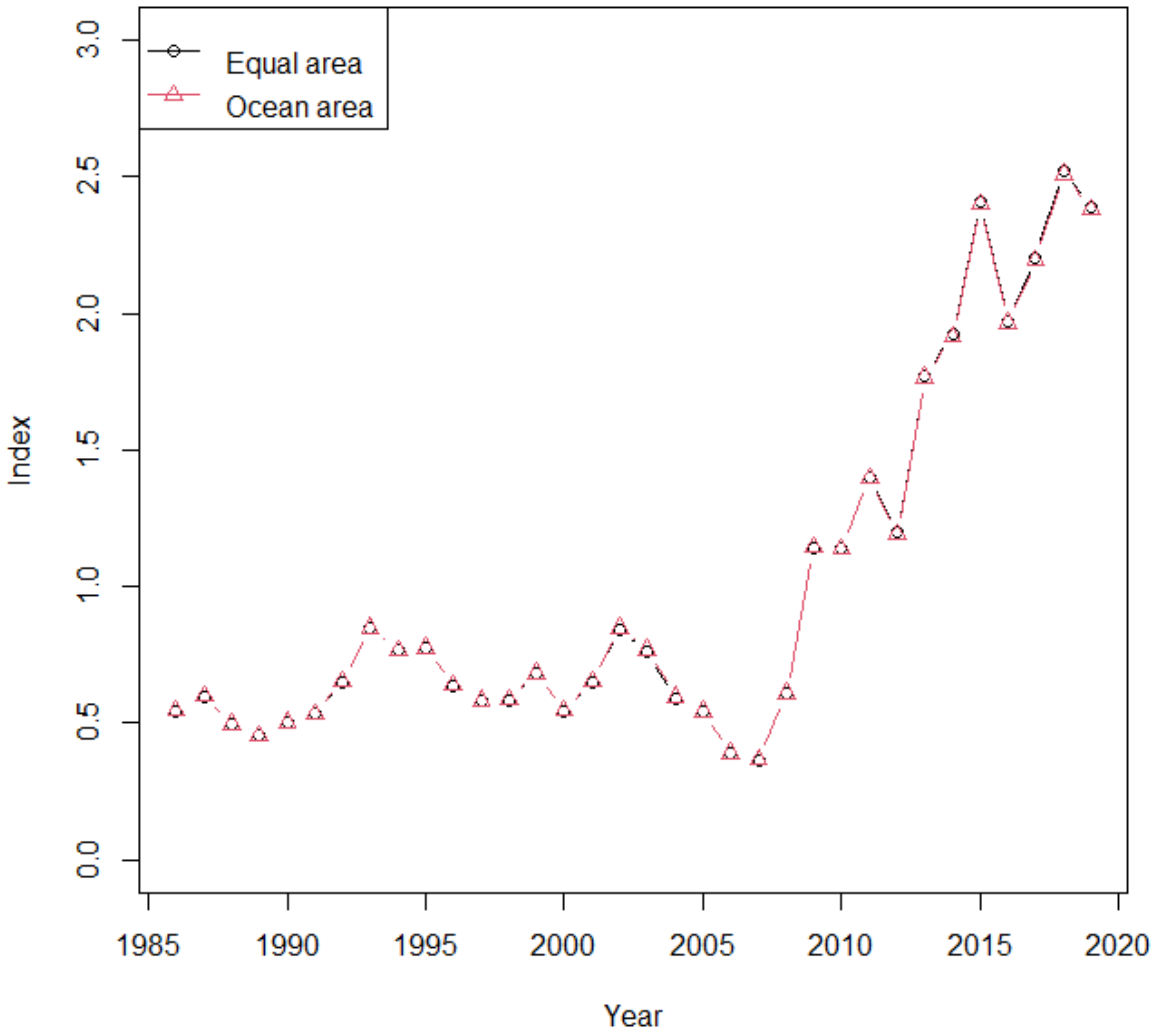

Figure 18: Delta lognormal model 11 Indices both with and without accounting for the differences among spatial cells in ocean area.

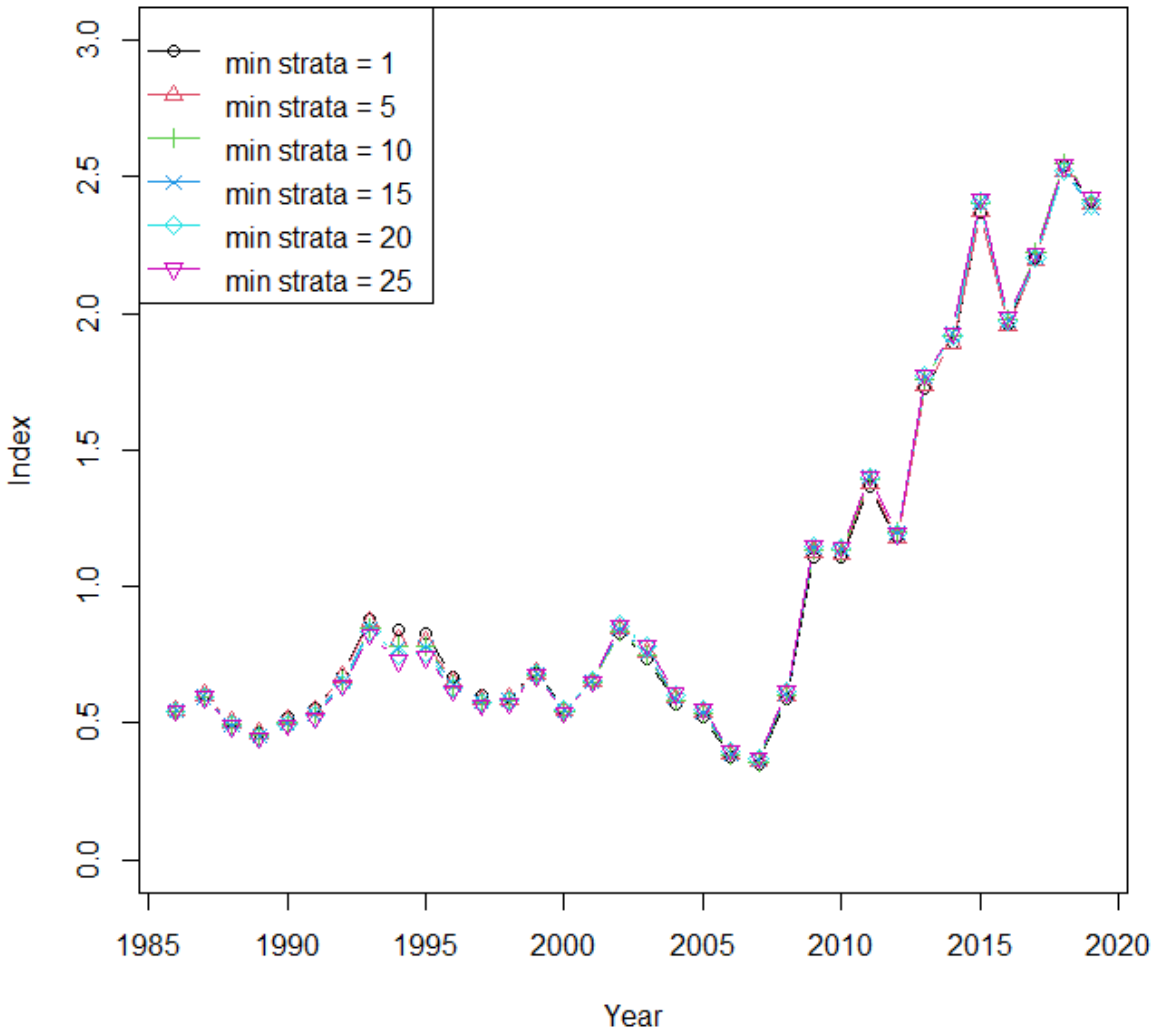

Figure 19:Delta lognormal model 11 Indices with different criteria for the minimum number of samples per spatial stratum.

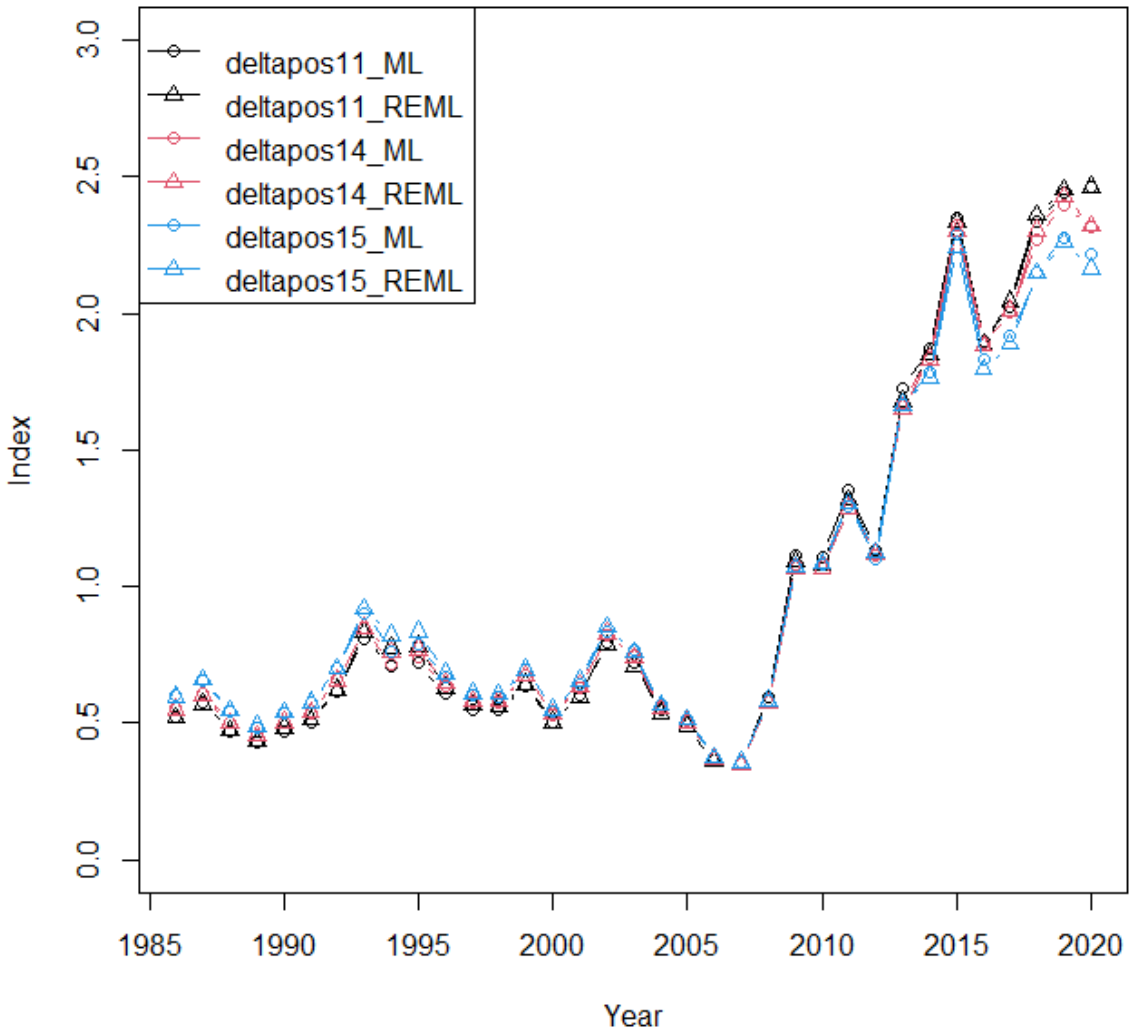

Figure 20: Indices from three different model structures fitted with either maximum likelihood (ML) for the delta component and GCV for the positive component, or restricted maximum likelihood (REML) for both components.

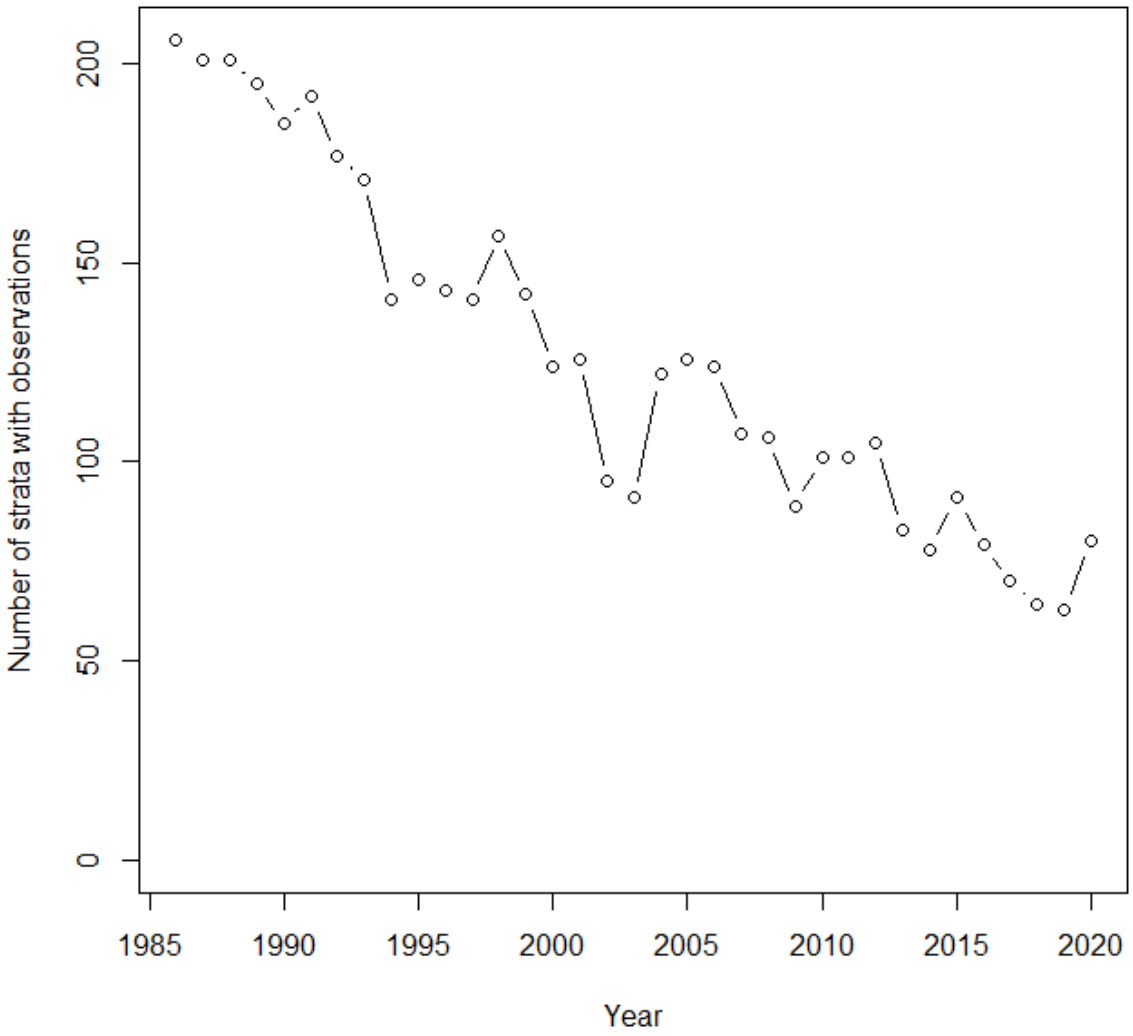

Figure 21: Number of strata per year with observations, including zero observations.

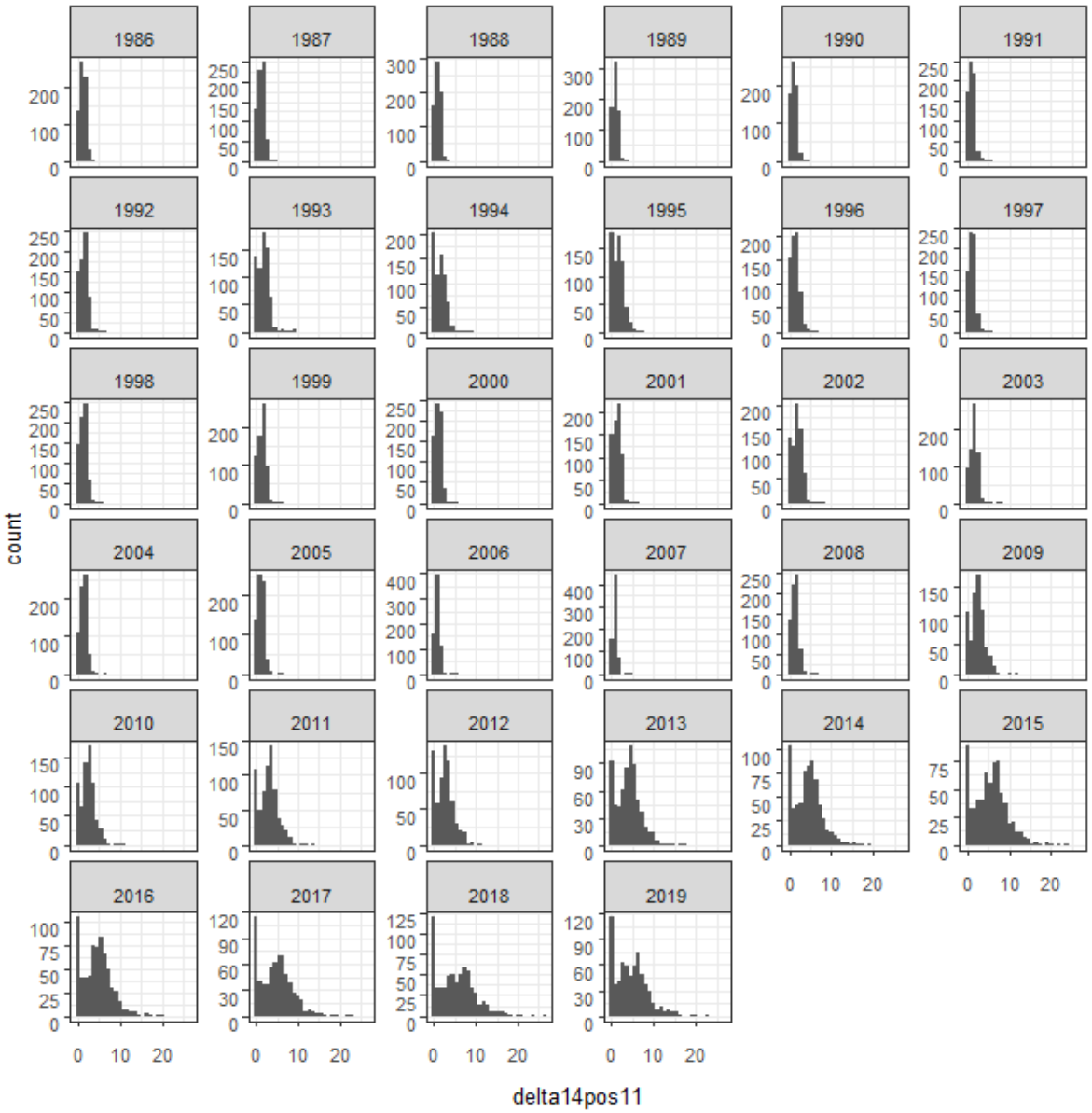

Figure 22: Histogram of predicted stratum values by year that result from combining models delta14 and pos11.

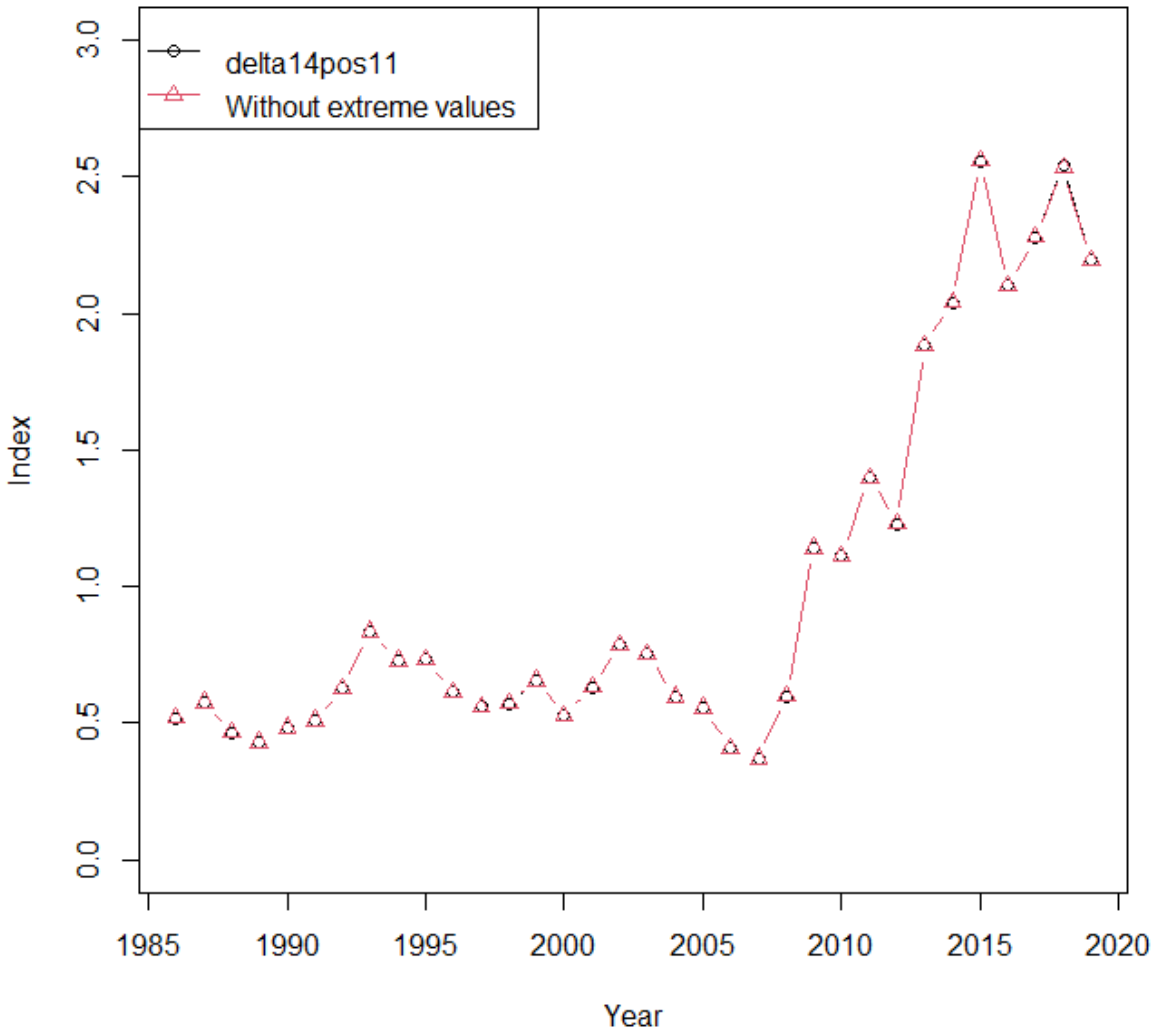

Figure 23: Comparison of indices with and without the extreme values. Extreme values are removed by limiting the maximum predicted values of all strata by year to the maximum observed CPUE in the year.

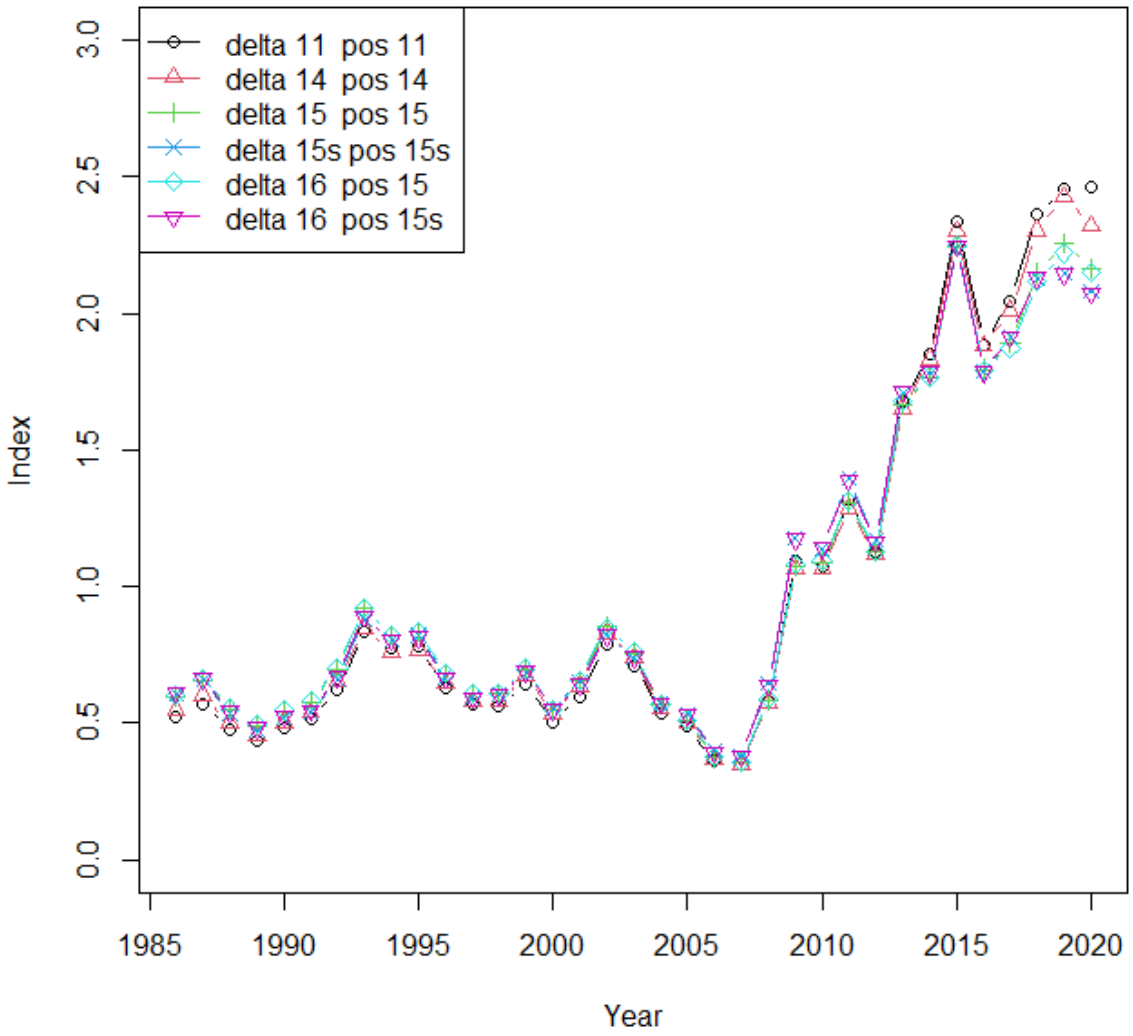

Figure 24: Indices for the six models with the lowest total AIC values.