SIMULATION-TESTING MODEL-BASED AND DESIGN-BASED BYCATCH ESTIMATORS

Elizabeth A. Babcock[1] , William J. Harford[2], Todd Gedamke[3], Sean Anderson[4], C. Phillip Goodyear[5]

SUMMARY

The BycatchEstimator tool developed by Babcock (2022) was used to estimate bycatch in fisheries simulated using the species distribution model and longline simulator (LLSIM) developed by Goodyear (2021). To compare the effectiveness of several design-based and model-based estimators that are used to estimate bycatch in a realistic context, an observer program similar to the US pelagic longline observer program (USPLL) was simulated. The estimates of total bycatch were precise and unbiased for all methods during recent years with high observer coverage. However, in the early years with lower observer coverage, the design-based methods (delta lognormal and ratio) performed worse than the delta lognormal model. The results were sensitive to how observers were allocated to trips. A geostatistical model showed that total bycatch estimates were more precise when spatial and/or spatiotemporal random effects were included. The BycatchEstimator tool was also applied to the real data from the USPLL. The tool was able to recreate the US Task I estimates in recent years, but when observer coverage was lower, estimates were sensitive to how strata with low sample sizes were pooled.

KEYWORDS

Bycatch, catch statistics, simulation, model testing, LLSIM, blue marlin, geostatistical models

1. Introduction

The BycatchEstimator tool of Babcock (2022) and Babcock et al. (2022) uses model-based and design-based procedures to estimate total annual bycatch by expanding the data from an observer program to the total effort from logbooks or landings records. This framework can also be used to estimate an annual index of abundance, calculated only from the observer data. Model-based bycatch estimation uses generalized linear models (GLM) based on the user's choice of observation error models (e.g. delta-lognormal, negative binomial) and predictor variables (e.g., year, season, depth). Information criteria (e.g. Akaike Information Criteria: AIC or Bayesian Information Criterion: BIC) can be used to find the best set of predictor variables, and cross-validation can be used to compare observation error models (e.g. negative binomial, delta lognormal). The selected GLMs are used to predict total bycatch in all logbook trips (or only unsampled trips, if desired) and total bycatch is estimated by summing across trips. The design-based methods include a stratified ratio estimator, and the delta-lognormal estimator of Pennington (1983), and the user may specify the stratification variables for these estimators (e.g. seasons, spatial areas). For the design-based estimators, if any strata have less than a user-specified number of observations (e.g. sets), estimates for those strata are made by pooling across the user's choice of stratification variables. Pooling can be done either across adjacent years, or by defining a more aggregated variable to be used in pooling (e.g. seasons rather than months), or by pooling across all levels of a variable.

The objectives of this analysis were (1) to add functionality to the bycatch estimation tool so that it can reproduce more of the methods that are used or are being considered for use for bycatch estimation in ICCAT CPC fisheries, (2) to improve functionality of the tool for abundance index standardization, (3) to consider how

[1]Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Cswy., Miami, Florida, 33149, USA. ebabcock@miami.edu
[2] Nature Analytics, 551 Lakeshore Rd E, Suite 105, Mississauga, Ontario, L5G 0A8, Canada. bill@natureanalytics.ca
[3] MER Consultants, 5521 SE Nassau Terrace, Stuart, Florida, 34997, USA. todd@merconsultants.org
[4] 220 Canterbury Cres, Nanaimo, British Columbia, V9T 4S4, Canada. sean@seananderson.ca
[5] 686 Hickory LN, Havana, Florida 32333, USA. phil_goodyear@msn.com

geostatistical models could be included in methods for bycatch estimation, and (4) to test the method on real CPC observer data. For the first objective, as a test case, a simulated observer program was developed that mimicked the design of the observer program for the US pelagic longline fishery (USPLL) and this was applied to the LLSIM's USA-like fleet. For this simulated fishery, bycatch of blue marlin was estimated using multiple model-based and design-based methods, including the method currently being used for the USPLL, which is the Pennington (1983) method, with pooling (as needed) across years, seasons and areas (Brown 2001). For objective 2, the ability to include random effects was added, so that commonly used methods such as random effects for year:area interactions could be modeled (Ortiz and Arocha 2004). For objective 3, the sdmTMB R package (Anderson et al. 2022) was used to estimate total bycatch for the simulated US longline fishery and for all three simulated fleets together. Finally, for objective 4, the bycatch estimation tool was applied with the same methods to the real US pelagic longline data, and the results were compared to the blue marlin discard estimates submitted to ICCAT by the USA.

2. Methods

2.1 Bycatch estimation tool

The functionality of the bycatch estimation tool is described in detail in Babcock et al. (2022) and the User Guide at Babcock (2022). Recently added functionality includes the Pennington (1983) delta lognormal estimator. This estimator for the mean CPUE in each stratum is:

$$E(CPUE) = \begin{cases} \frac{m}{n} e^{\bar{y}} G_m\left(\frac{1}{2}s^2\right) & m > 1 \\ \frac{x_1}{n} & m = 1 \\ 0 & m = 0 \end{cases}$$

where m is the number of positive observations out of n samples, $\bar{y}$ and s^2 are the mean and variance of the log of the positive observations, x_1 is the positive CPUE if only one value is positive (m=1), and G_m is a bias correction function defined as:

$$G_m(t) = 1 + \frac{m-1}{m}t + \sum_{j=2}^{\infty} \frac{(m-1)^{(2j-1)}}{m^j(m+1)(m+3)\ldots(m+2j-3)} \cdot \frac{t^j}{j!}$$

The variance of the mean CPUE is:

$$V_E(CPUE) = \begin{cases} \frac{m}{n} e^{2\bar{y}} \left\{ \frac{m}{n} G_m^2\left(\frac{1}{2}s^2\right) - \frac{m-1}{n-1} G_m\left(\frac{m-2}{m-1}s^2\right) \right\} & m > 1 \\ \left(\frac{x_1}{n}\right)^2 & m = 1 \\ 0 & m = 0 \end{cases}$$

In each stratum, the mean CPUE is multiplied by total effort to get total bycatch, and the totals are summed across strata. The variances are multiplied by effort squared and summed to calculate the total variance.

For both the Pennington estimator and the ratio estimator, the user specifies which variables define the strata. If desired, the user can specify a minimum number of sample units below which strata should be pooled. If pooling is needed, variables will be pooled in the order in which the design variables are entered (e.g. year, season, area). For strata over the cutoff of effort, no pooling is done. For each stratum (i) that requires pooling, the algorithm is: (1) identify the strata that need to be included to reach the minimum sample size, by pooling variables in the order specified by the user. For example, the pool could include adjacent years, and if the number of samples is still below the threshold could include all seasons; (2) estimate the design-based estimates across the observer and logbook data in the pool; (3) calculate the total bycatch of stratum i as the total bycatch in its pool times the fraction of the total effort in the pool that is in stratum i. The variance is also allocated to stratum i based on the fraction of effort in each stratum in the pool.

The new version of the tool also has the capacity to include random effects, such as a vessel effect, or random interactions (Ortiz and Arocha 2014) as are commonly used for CPUE index standardization. If random effects are specified, they are included in all models for both bycatch estimation and index standardization. In practice, it may be useful to run the model with fixed effects only to identify interactions that improve prediction, and then include these interactions as random effects in a later run of the model. Note that it is possible to run the model for index standardization only without estimating total bycatch. In this case, no logbook data set is needed.

The updated user's guide and information on the GitHub site (Babcock 2022) now include detailed information on how to install the package and debug common problems that come up, as well as advice on how to use the tool. This material is intended to make the package easier for analysts from multiple CPCs to learn and apply.

2.2 Longline fishery simulation testing

LLSIM was used to simulate a USA-like fleet as described by Goodyear (2021), with data from 1990 to 2015. The species distribution model (SDM) generates a 3-dimensional distribution of blue marlin and swordfish throughout the Atlantic Ocean based on the habitat preferences of the species (Goodyear 2016; Goodyear et al. 2017, Forrestal and Schirripa 2019). LLSIM then simulates longline sets by distributing hooks throughout the habitat of the species, consistent with the distribution, gear, depth of fishing, use of light sticks and other characteristics of historical longline fishing fleets. The probability of each hook capturing a blue marlin or swordfish is then determined by the three-dimensional location of the hook and the probability of fish presence from the SDM (Goodyear 2021).

While LLSIM initially produces set-level catches, sets were allocated to simulated trips to more accurately reflect the fact that observers are allocated randomly by trips rather than sets. The method of Grüss et al. (2019), which has been used in previous simulation studies from LLSIM, was modified to more closely mimic the distribution of predictor variables within trips in the US fishery. Sets were allocated to the same trip if they were in the same gear, month and spatial area (5 x 5 squares) and had the same number of hooks between floats (hbf), since these variables tended to be similar within trips in the USPLL fishery. Trips with more than 7 sets were randomly allocated to different trips so that the median trips had about 7 sets and the distribution of sets per trip was roughly comparable to the real USPLL fishery. This algorithm allowed for correlation among sets in the same trip to introduce potential clustering bias into the simulated observer data.

The datasets simulated using LLSIM were then provided to the observer program sub-model. The observer coverage level each year was as close as possible to the actual coverage level, based on those reported by Diaz et al. (2009), and the 2017 -2019 annual reports to ICCAT by USA, matching the "realistic" scenario from Babcock et al. (2022). In the US pelagic observer program, observers are randomly assigned to trips based on the effort (in sets) in the previous year in each stratum, defined by season (3 month period) and the US pelagic longline spatial areas (Brown 2001), and vessels are selected with probability proportional to their effort in the previous year (Keene et al. 2010). The simulated data has no way to match trips from the same vessel. Thus, to simulate a process similar to the US sampling plan, the following algorithm was used. Within each stratum, defined by season and USA pelagic longline area, in each year, the desired sampling effort, in sets, was calculated as the coverage level times the number of sets in the season-area stratum in the previous year. The probability of trips being sampled was roughly proportional to the number of sets being sampled. If the preliminary sample had more sets than were needed, later trips could be removed from the sample. This method gave coverage roughly similar to the true coverage levels over time. This algorithm was used to generate 100 random draws of the USA-like fishery.

The bycatch estimation tool was used to estimate total bycatch using for each of the 100 draws for the USA-like fleet only. Sets were used as the sample unit, in contrast to Babcock et al. (2022), which used trips as the sample unit. Potential predictor variables for GLM included year, the ICCAT billfish areas, season, and hooks between floats (hbf), and the best set of predictor variables was selected using the Bayesian Information Criterion (BIC). The delta-lognormal model-based estimator was used for all draws, and the bycatch in all trips was predicted from the model, rather than predicting only the unobserved effort. Variances were calculated by the Monte Carlo simulation method, which involves drawing 1000 values of each of the model coefficients and then drawing simulated values of the predicted catches in each logbook trip. The stratified ratio estimator (Rao 2000), and the Pennington (1983) delta lognormal estimator were also used, with the stratification variables of year, ICCAT billfish area and season (Brown 2001). The minimum sample size needed to avoid pooling was either 30, consistent with Brown (2001), or 5 for comparison. Years were pooled with the adjacent years so that the pool

for each stratum would include 3 years. If the year pooling did not reach the minimum sample sizes, all seasons were combined, then all areas. For the delta-lognormal model, design-based delta-lognormal and ratio estimator, the bias in annual total bycatch estimates was calculated. The accuracy of the variance estimates for both model and design-based estimates, was evaluated by calculating the coverage for each observation error and model year, where coverage is defined as the fraction of the 100 draws in which the estimated 95% confidence interval contained the true value.

Abundance indices were calculated using the Tweedie, delta-lognormal, negative binomial 1 and negative binomial 2 models for simulated USA-like observer data. Potential variables are year, the ICCAT billfish areas, season, and hooks between floats (hbf). A year-area interaction as a random effect was included in some runs for comparison.

For the purpose of testing the geostatistical models, the simulator was also used to generate a single draw with 5 percent observer coverage across all three fleets, including the one based on the USA, and the ones based on Japan and Brazil. Trips were sampled randomly for this draw. For comparison, both the geostatistical model and the bycatch estimation tool were applied to this dataset.

2.3 Geostatistical models

We fit geostatistical models using the R package sdmTMB (Anderson et al. 2022), which implements spatial and spatiotemporal predictive-process GLMMs using the SPDE (stochastic partial differential equation) approximation to Gaussian random fields. sdmTMB fits geostatistical models with maximum marginal likelihood calculated with TMB (Template Model Builder, Kristensen et al. 2016) and the Laplace approximation and uses a "mesh" constructed with the R package INLA (Lindgren and Rue 2015) for the SPDE approximation and bilinear interpolation. Previous work has shown that such an approach can improve the accuracy and precision of population indices from fisheries independent surveys (e.g., Thorson et al. 2015) and commercial CPUE (e.g., Grüss et al. 2019).

The general structure of the most complex models fit was

$$\begin{aligned} \mathbb{E}[y_{\mathbf{s},t}] &= \mu_{\mathbf{s},t}, \\ \mu_{\mathbf{s},t} &= f^{-1}\left(\mathbf{X}_{\mathbf{s},t}\boldsymbol{\beta} + O_{\mathbf{s},t} + \omega_{\mathbf{s}} + \epsilon_{\mathbf{s},t}\right), \\ \boldsymbol{\omega} &\sim \text{MVNormal}(\mathbf{0}, \boldsymbol{\Sigma}_{\omega}), \\ \boldsymbol{\epsilon}_{\mathbf{t}} &\sim \text{MVNormal}(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon}), \end{aligned}$$

where $y_{\mathbf{s},t}$ represents blue marlin bycatch counts at coordinates $\mathbf{s}$ in space and time t, μ represents the mean, f^{-1} represents an inverse link function, $\mathbf{X}$ represents a design matrix, $\boldsymbol{\beta}$ represents a vector of main-effect coefficients, $O_{\mathbf{s},t}$ represents an offset of log hook count, $\boldsymbol{\omega}$ represents a spatial Gaussian random field with mean zero and covariance $\boldsymbol{\Sigma}_{\omega}$, and $\boldsymbol{\epsilon}_{\mathbf{t}}$ represents a spatiotemporal Gaussian random field with mean zero and covariance $\boldsymbol{\Sigma}_{\epsilon}$ that is independent each year. For main effects, we used factor (categorical) predictors for year, season, and light-stick presence as well as a linear predictor for log hooks between floats (hbf).

There are several ways these models can be configured. We evaluated four families: NB2, NB1, delta-gamma, and delta-lognormal. We initially tested the Tweedie distribution but had challenges with model convergence unless the model was fit to CPUE instead of bycatch with an offset. We therefore excluded it here, but future work could explore it. We evaluated three random field configurations: (1) no random fields, (2) spatial random fields, and (3) spatial and spatiotemporal random fields with the spatiotemporal random fields being independent each year. Based on initial testing, we let the range parameter, which defines the distance at which spatial correlation has decayed to about 13%, be independent between the spatial and spatiotemporal random fields. We also allowed for spatial anisotropy (Fuglstad et al. 2015; Thorson et al. 2015): spatial correlation that varies with direction. Here, we shared the anisotropy properties between the spatial and spatiotemporal fields since they can be challenging to estimate. For sdmTMB, this means we used the arguments anisotropy = TRUE, share_range = FALSE, spatial = "on", and spatiotemporal = "iid" for our full spatiotemporal model; the anisotropy is by default shared but can be configured within sdmTMBcontrol() via the 'map' argument.

Geostatistical models are best fit in a coordinate space where distance is constant. For smaller spatial areas, UTMs are commonly used, but these datasets spanned well beyond one UTM zone introducing considerable distortion. Instead, we chose a custom Albers projection with reference longitudes at approximately 1/6 from the bottom and top of the data—future work could investigate the impact of this decision, but our initial observations suggest this had a minimal impact on model predictions. We fit out models with coordinates in 100 km units so

that the spatial parameters were on an appropriate scale for estimation. We configured the mesh to have a 'cutoff' of 500 km for the full dataset and 300 km for the USA-like data set, which means no triangle edge was allowed to be smaller than 500 or 300 km, respectively. This mesh has approximately 250 triangle vertices or 'knots' for both datasets. Model convergence can be sensitive to the exact mesh configuration—a problem that can often be alleviated by increasing or reducing the mesh resolution or, more elegantly, applying penalized complexity priors, which could be explored in future work (available within sdmTMBpriors() via pc_matern()).

sdmTMB uses a projection matrix **A**, calculated through R-INLA, to bilinearly interpolate random field values at triangle vertex locations ($\boldsymbol{\omega}^*$ and $\boldsymbol{\epsilon}^*$) to the values at the locations of the observed or predicted data (Lindgren and Rue 2015): $\boldsymbol{\omega}^* = \mathbf{A}\boldsymbol{\omega}$ and $\boldsymbol{\epsilon}^* = \mathbf{A}\boldsymbol{\epsilon}$. The matrix **A** has a row for each data point and a column for each vertex defining the weight of the neighboring three vertices.

sdmTMB minimizes the negative marginal log likelihood—calculated via TMB—with the non-linear optimization routine stats::nlminb() in R (Gay 1990; R Core Team 2022) followed by a Newton optimization routine stats::optimHess() in R (R Core Team 2022) to further reduce the likelihood gradients with respect to fixed effects if needed. Using sanity() function within sdmTMB, we assessed convergence by checking that the Hessian matrix was positive definite, that all gradients with respect to fixed effects were < 0.001, and that no random field marginal standard deviations were too small (< 0.01; suggesting the parameter had 'collapsed' to zero) among other checks.

To derive predicted total bycatch on the logbook data, we predicted from our model on the logbook data and summed the predicted bycatch each year. We calculated standard errors on the log value of this total via the generalized delta method (as implemented in TMB, Kristensen et al. 2016) and calculated 95% Wald confidence intervals as $\pm$ 1.96 the standard error in log space. We evaluated model fit via randomized quantile residuals and calculated summary statistics of MARE (median absolute relative error), MRE (mean relative error), and coverage (the proportion of years in which the true value was within the confidence interval). See https://github.com/seananderson/llsim-geostat for the code used in this analysis.

2.4 USA Pelagic longline fishery

To demonstrate the application and behavior of the Bycatch Estimator R package in real world situations, the tool was used to conduct a preliminary evaluation of blue marlin bycatch in the USA Atlantic pelagic longline fishery. Raw data from the USA pelagic observer program (POP) and the pelagic longline logbook reporting program (USPLL) were provided by NOAA's Southeast Regional Office. To be as consistent as possible with previous studies, analysts from the Southeast Fisheries Science Center (i.e. those that provided the US Task 1 estimates) were contacted directly to obtain both length-weight conversion parameters and estimates of Blue Marlin total bycatch for the most recent years of data (Anonymous 2018; SEFSC, pers comm). Data were provided beginning in 1992, which corresponds to the start of the observer program, through 2021. The raw data files contained just under 425,000 records/sets in logbook program and just under 22,000 observed sets.

Data were documented well and provided in raw form, which required a full evaluation and processing prior to application of the bycatch estimation package. Records which did not contain fundamental information (e.g. year, gear) were removed and then decisions had to be made about retaining records which contained questionable entries for a variable (i.e. hooks, hbf, latitude, and/or longitude) or for outliers that were outside the expected range.

For example, in logbook cases where latitude or longitude was missing, an average fishing location for each vessel over all years was used to designate a US or ICCAT area (Version: 2016.02). In other logbook records, total hooks (i.e. effort/set) or hbf was missing and averages or calculations from other variables (e.g. for missing hbf, total hooks per set was divided by floats per set) was applied. The observer data had very little missing data, but an average weight was used in cases where unrealistic fish weights were calculated or length was not recorded. Although raw data manipulation was not necessary in more than ~5% of the data, this was an important step given the low sample sizes used to calculate CPUE for a stratum. It is important to note here and throughout that the results presented in this document are only intended to illustrate how the BycatchEstimator package can be applied to real data and the results should not be used in place of the US Task 1 estimates.

3.0 Results

3.1 Simulation testing

Using simulated USA-like data, the delta lognormal model, delta lognormal estimator and stratified ratio estimator all performed similarly, although there was a slight tendency for the design delta lognormal to estimate higher bycatch than the stratified ratio estimator (**Table 1, Figure 1**). The minimum cutoff below which the design-based estimators were pooled (5 sets vs. 30 sets) did not make much difference in the median estimates. Because the simulated observer program had coverage less than 1% in the early years increasing to more than 8% in 2015, the bias in the estimates decreased over time for all estimation methods (**Figure 2**).

As the observer sample size increased over time, the fraction of confidence intervals that included the true value increased for all three methods (**Figure 3**). When observer coverage was low in the early years, the fraction of confidence intervals including the true value was higher for the model-based estimate than the design-based estimates. Within individual simulations, the widths of the confidence intervals were similar for all three methods, even with different pooling cutoffs in the design-based methods (**Figure 4**). The lower fraction of true values in the confidence intervals for the design-based methods in the early years seems to be caused by the slight overestimate in the design-based estimates in the early years in some simulations. Increasing the amount of pooling somewhat reduced the fraction of confidence intervals including the true value in the early years for design-based estimators. This underestimation was not seen when observers were allocated randomly rather than in a stratified system (e.g. **Figure 4**, top panel). The data were pooled in the simulation much more than in the real US data (**Figure 5**). This indicates that either the simulated allocation of sets to trips or the simulated allocation of observers to trips was not representative of the real US fishery (Brown 2001, Diaz et al. 2009).

When the model was used to generate abundance indices, there was not much difference in either trends or confidence intervals among the observation error models (delta-lognormal, negative binomial 1, negative binomial 2, Tweedie), and whether or not a random year:area effect was included (**Figure 6**). However, different random draws from the observer sampling program gave highly variable year to year estimates in the early years when observer coverage was low. The similarity among observation error models is not surprising, as all performed well according to model diagnostics. For example, the DHARMa QQ plots show the quantiles of the DHARMa scaled residuals against a uniform distribution (**Figure 7**), and the points for all distributions fall roughly along the line, indicating that the observation error model adequately describes the distribution of the data (Hartig 2020). Also, within a random draw, the same set of predictor variables were generally chosen for all observation error models (e.g. **Table 2**).

3.2 Geostatistical models

The geostatistical models were applied using a mesh with edges of at least 500 km, corresponding to about 250 vertices throughout the Atlantic (**Figure 8**). We observed considerable anisotropy with correlation decaying more quickly in an approximately latitudinal direction compared to an approximately longitudinal direction. The exact level of anisotropy may be sensitive to the projection used. We also estimated a larger spatial range than spatiotemporal range (**Figure 9**). Randomized quantile residuals suggested the models were consistent with the distribution of the data (**Figure 10**). The models applied to the full Atlantic simulated dataset and with the specified level of mesh resolution fit reasonably quickly (about six minutes for the most complex model) but were slower to fit than GLMs without random fields. The models with only spatial random fields fit considerably faster than models with both spatial and spatiotemporal random fields.

Models were applied with no spatial random fields, with spatial random fields (e.g., **Figure 11**) and with spatiotemporal random fields (e.g., **Figure 12**). These random field deviations represent spatially correlated effects from latent variables not included in the model that are static through time (spatial) and that change from year to year (spatiotemporal). The projected maps of predicted bycatch across space and time were consistent with the expected distribution of blue marlin habitat and a decreasing trend in bycatch over time (**Figure 13**). The variability in the predicted bycatch followed the expected pattern with higher CVs at the north and south range edges where the data were sparser (**Figure 14**).

When the models were used to predict bycatch for the simulated logbook sets, the resulting predictions were quite accurate, particularly when spatial and spatiotemporal random fields were included (**Figure 15, Figure 16**). In general, the models with spatial and spatiotemporal random fields performed best in terms of MARE, MRE, and confidence interval coverage, with spatial fields a close second, and models without random fields third (**Table 3**). For the USA-fleet-only dataset, the marginal standard deviation of the delta-gamma and delta-lognormal spatiotemporal random fields collapsed to zero suggesting that the models with only spatial random

fields were sufficient. This result and which specific observation error model is preferred may be sensitive to the mesh set up. Nevertheless, all the observation error models performed adequately. The two negative binomial models consistently performed best in confidence interval coverage.

3.3 US Pelagic longline

The initial evaluation of the data showed some differences in the observer coverage available for the analysis in comparison to the coverage presented through 2007 in Diaz et al. (2009). In 2001, 2002, 2003, and 2005, the observed effort (hooks) in this analysis represent 64%, 38%, 43% and 73% respectively of those reported (**Figure 17, Figure 18**). Consultation with the data providers suggested that the inclusion of experimental trips or records outside the standard observer program may have resulted in these differences and for the purposes of this study no additional data were requested.

The analysis was initially set up according to the observer allocation procedure described in Brown (2001), which utilized US Pelagic Longline area; however, for comparison to the Task 1 data reported at BUM (2018) and those provided by the SEFSC (pers comm), the analysis utilized the ICCAT areas to estimate bycatch for all years from 1992 to present. Manual pooling of areas to obtain sufficient sample sizes was initially done and estimates were sensitive to this process. The BycatchEstimator package was updated during this work and areas are now pooled within the analysis as necessary which greatly simplifies the pre-processing of raw data necessary. A single lumped category was used for a small number of records (e.g. 0.6% of PLL, and 0.1% of POP) for which designating a clear ICCAT area was questionable.

Once the data were formatted and the BycatchEstimator package applied, a couple of unexpected real world data issues arose. In the raw data, some strata contained more observed sets than were reported in the logbook and the package failed. To resolve this issue a few observed sets were simply removed. Once the data were corrected the package was able to produce results; however, estimates were unreasonable from the delta-lognormal model-based estimator. In short, the logbook data contained values of hbf that were orders of magnitude greater than those in the observer database so the extrapolation of this linear term by multiple orders of magnitude gives a predicted variance that is extremely high. Because the formula to calculate the mean for the lognormal includes the variance term in the bias correction, the resulting estimates were not valid. The hbf values were capped at the maximum value in the observer database rather than simply excluded, but the lesson learned was that the numerical range for variables needs to be comparable between observer and logbook datasets.

The first direct comparison of bycatch estimates from the tool and those provided by SEFSC were made on with results from the Pennington (1983) delta lognormal estimator with the stratification variables of year, ICCAT billfish area and season to be consistent with the methodology in Brown (2001). The US Task 1 estimates for all but five years prior to 2010 fell within the confidence intervals of those calculated with the design-based delta (**Figure 19**). Estimates from the early part of the time series (prior to ~2010) were sensitive to data processing as would be expected given the low observer coverage rates and resulting extrapolations. Between 2000 and 2010, the years in which raw differed from that in Diaz et al. (2008) the bycatch estimator produced estimates which were consistently higher than the US Task 1 data. After 2010 estimates were very similar and stable to minor changes in raw data.

All of the observation error models selected from the BycatchEstimator package for this analysis successfully converged providing both estimates and confidence intervals (**Figure 20**). Estimates from 1996 were the most uncertain with confidence intervals from the TMBnbinomial2 model dwarfing the signal from the other methods. The pattern of annual estimates from all models was similar across methods with differing levels of uncertainty based on the approach (**Figure 21**).

4. Discussion

Compared to the simulations in Babcock et al. (2022), the estimated bycatch for the simulated USA-like fleet in this paper have somewhat higher bias and lower confidence interval coverage for equivalent methods. This may be because the observer sampling program was less random due to the stratified observer allocation scheme. The fact that the simulated data are generated by set rather than trip makes it difficult to recreate a real trip-based observer allocation process. Thus, these results are probably not representative of the real US fishery. However, they are informative about the kinds of biases that can be introduced by a sampling scheme that is not entirely random at the level of the sample unit. Future work might try a range of observer allocation schemes and pooling

algorithms to see if it is possible to overcome some of these biases. Also, the spatial allocation scheme was based on the US pelagic longline areas (Brown 2001), while the estimation method used the ICCAT billfish areas, and this mismatch may explain some of the bias in the estimates.

In addition to the model-based estimates, the BycatchEstimator tool can calculate bycatch using either the stratified ratio estimator or design-based delta lognormal estimator, with multiple options for how pooling should be set up to impute bycatch in under-sampled strata. This makes the tool useful for simulation testing and should make it possible to approximate the estimation methods used in multiple fisheries. However, it is likely that many real world fisheries will have nuances in their estimation methods that cannot easily be automated, such as adjusting the number of years over which data are pooled over time as observer coverage levels increase. The tool could potentially be applied in recent years only when estimation methods are more consistent.

The ability to add random effects to the model now makes it possible to recreate many of the abundance index estimations methods used by CPCs. The indices we estimated from the simulated data with and without random effects were very similar, but this was expected because the simulated data has no significant year:area interaction (results not shown). Further testing could evaluate other variables that do have significant interactions.

Although geostatistical models have not yet been added to the bycatch estimation tool, they can be set up for these data using sdmTMB. Model convergence can be sensitive to technical details such as how the spatial mesh is designed, and the models are slower to fit than GLMs without random fields. However, based on our preliminary work, random fields can substantially improve the estimates of total bycatch, so they should be the focus of future work. In the geostatistical models, as in the GLMs used in the BycatchEstimator, the specific choice of an observation error model (e.g. delta-lognormal vs. negative binomial 1 or 2) made minor differences in confidence interval coverage but all performed adequately and gave similar total bycatch estimates.

The overall application of the BycatchEstimator package to data from the US Pelagic Longline Fishery was successful and proves promising for applications to other data sets. As would be expected, in the early years of the time series when US observer coverage was below 5%, the results of this analysis were sensitive to how the raw data were prepared and how much pooling of strata (i.e. years, seasons, and areas), were necessary to meet minimum sample size requirements. A consistent and documented data processing approach should be the first step in moving forward with a powerful analysis package that can be applied to multiple CPC data streams. Decisions such as how to treat a ‘parted’ longline or whether to calculate ‘effective’ effort in terms of hooks set or hooks retrieved are universal questions which analysts will face regardless of country of origin. A standard approach for imputing missing effort or area variables (i.e. hooks, sets, trips, or ICCAT area) that borrows information from pooled strata, if necessary, to designate a ‘typical’ trip within a strata would be helpful to many analysts attempting to apply the BycatchEstimator tool or any other standardized method for bycatch estimation. Limiting the number of decisions that have to be made by analysts in the data processing step would allow for greater comparisons of BycatchEstimator results between CPC countries. Further work could explore an extension of the BycatchEstimator package or a stand-alone module which would allow for all CPC datasets to be pre-processed (i.e. quality control and filling in missing data when necessary) as input files in a consistent fashion.

As this process moves forward the use of additional factors and how they may be reliably collected should be considered, as additional variables that are correlated with bycatch could greatly improve the precision and accuracy of model-based estimates. The most obvious factor that was not considered in this application but was shown to be important in the Babcock et al. (2022) study, is the use of lightsticks. In the LLSIM data this factor serves as a proxy for targeting and the intended depth of the sets. Although the US logbook data contained a field for the use of lightsticks it was not populated with enough information to use. The use of additional information such as stated target species, time of set, bait, and soak time could be used to infer target species. Similarly, future work should consider using factors which are currently available (e.g. habitat suitability scoring scheme) or those that could be generated from satellite information that can be applied to the raw data based on time, date and location (e.g. temperature, frontal boundaries, chlorophyll). Additional variables could potentially make model-based bycatch estimates much more accurate, with the caveat that variables are only useful if they are also present in the logbook data for the prediction step.

Finally, the ability to match observer trips and/or sets to the logbook data in all CPC’s should be made a priority. Effort, and thus the extrapolations of observer data, are generally made on self-reported effort metrics and validation of these data by cross referencing sets will allow for the calculation of correction factors, and the

ability to include observed bycatch as a known constant, which improves precision in model-based estimates. Finally, in this analysis, the model-based and design-based estimators were quite similar, for the most part, implying that either approach can yield valid estimates. However, it should be noted that model-based methods generally require data from the whole time series, or at least a large part of it, to estimate the model coefficients, while design-based estimates can be calculated for only recent years, because the estimates are not influenced by data from other time periods beyond the range of years that are pooled together to deal with sparse data.

In conclusion, this set of analysis found that the BycatchEstimation tool is able to estimate bycatch accurately with a variety of methods. However, results can be sensitive to apparently small differences in how observers are allocated to trips, and the decisions made in data cleaning and missing data imputation, implying that these elements may be more important than statistical methodology in the attempt to standardize bycatch estimation across fisheries. Finally, adding geostatistical methods, in cases where location data are available from both observer and logbook data, has the potential to improve estimates substantially.

5. Acknowledgements

This work was developed with financial support from the ICCAT Science Budget and the European Union Grant Agreement SI2.839159 - Strengthening the scientific basis for decision-making in ICCAT. E. A. Babcock's work was also partly supported by NOAA, via the Cooperative Institute for Marine and Atmospheric Science and Goodyear's work was supported by The Billfish Foundation, Ft. Lauderdale, FL, USA.

6. References

Anderson, S.C., E.J. Ward, P.A. English, L.A.K. Barnett. 2022. sdmTMB: an R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. bioRxiv 2022.03.24.485545; doi: https://doi.org/10.1101/2022.03.24.485545

Anon. 2018. Collect. Report of the 2018 ICCAT Blue Marlin Stock Assessment Meeting (Miami, United States, 18-22 June 2018) Vol. Sci. Pap. ICCAT, 75(5): 813-888.

Babcock, E.A. 2022. Bycatch Estimator R library. https://ebabcock.github.io/BycatchEstimator/

Babcock, E.A., W. J. Harford, T. Gedamke, D. Soto, and C. P. Goodyear. 2022. Collect. Vol. Sci. Pap. ICCAT, 79(5): 304-339.

Brown, C.A. 2001. Revised estimates of bluefin tuna dead discards by the U.S. Atlantic pelagic longline fleet, 1992-1999. Collect. Vol. Sci. Pap. ICCAT 52(3):1007-1021.

Diaz, G.A., L. R. Beerkircher & V. R. Restrepo. 2009. Description of the U.S. Pelagic Observer Program (POP). Collect. Vol. Sci. Pap. ICCAT 64(7): 2415-2426

Forrestal, F.C., and Schirripa, M.J. 2020. Addition of Swordfish distribution model to longline simulator study. Collect. Vol. Sci. Pap. ICCAT, 77(5): 37-66.

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. 2015. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. Stat. Sinica 25(1): 115–133. Institute of Statistical Science, Academia Sinica.

Gay, D.M. 1990. Usage summary for selected optimization routines. Computing Science Technical Report 153: 1–21.

Goodyear, C.P. 2021. Development of new model fisheries for simulating longline catch data with LLSIM. Collect. Vol. Sci. Pap. ICCAT, 78(5): 53-62.

Goodyear, C. P., M. Schirripa, and F. Forrestal. 2017. Longline data simulation: a paradigm for improving CPUE standardization. Collect. Vol. Sci. Pap. ICCAT 74:379-390.

Goodyear, C.P., 2016. Modeling the time-varying density distribution of highly migratory species: Atlantic blue marlin as an example. Fish. Res. 183, 469–481.

Grüss, A., J. F. Walter, E. A. Babcock, F. C. Forrestal, J. T. Thorson, M. V. Lauretta, and M. J. Schirripa. 2019. Evaluation of the impacts of different treatments of spatio-temporal variation in catch-per-unit-effort standardization models. Fish. Res. 213:75-93.

Hartig, F. 2020. DHARMa: Residual Diagnostics for Hierarchical (multi-Level / Mixed) Regression Models. https://CRAN.R-project.org/package=DHARMa.

Keene, K. F., L. R. Beerkircher and D. W. Lee. 2010. SEFSC pelagic observer program data summary for 2005 & 2006. NOAA Tech. Memo. NMFS-SEFSC-603

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: Automatic differentiation and Laplace approximation. J. Stat. Softw.70 (5) 1-21.

Lindgren, F., and Rue, H. 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63(1): 1–25.
Ortiz, M. and F. Arocha. 2004. Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish species in the Venezuelan tuna longline fishery. Fish. Res. 70, 275–297.
Pennington, M. 1983. Efficient Estimators of Abundance, for Fish and Plankton Surveys. Biometrics 39 (1): 281–86. https://www.jstor.org/stable/2530830.
R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Rao, P. S. R. S. 2000. Sampling Methodologies with Applications. Texts in Statistical Science. Chapman and Hall/ CRC.
Thorson, J.T., Shelton, A.O., Ward, E.J., and Skaug, H.J. 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICES J. Mar. Sci. 72(5): 1297–1310.

Table 1. Summary of percent bias for estimates of bycatch from the USA-like simulated fleet. Estimation models are: stratified ratio estimator, design-based delta-lognormal and delta-lognormal model. Shown are results across all years 'All', and for years 2000 and 2010 across 100 simulation runs.

				Centered 95%	
Year	**Sets needed to avoid pooling**	**Estimation model**	**Median**	**Lower**	**Upper**
All	30 sets	Stratified ratio	0.74	-40.67	54.70
All	30 sets	Design delta	9.63	-37.82	67.92
All	5 sets	Stratified ratio	1.01	-44.00	64.53
All	5 sets	Design delta	9.30	-41.12	77.92
All	Model	Model delta	5.84	-40.77	66.83
2000	30 sets	Stratified ratio	-3.88	-35.00	37.17
2000	30 sets	Design delta	5.18	-30.93	44.59
2000	5 sets	Stratified ratio	-5.03	-46.98	50.74
2000	5 sets	Design delta	0.52	-43.36	56.40
2000	Model	Model delta	-2.55	-43.21	53.56
2010	30 sets	Stratified ratio	-5.39	-47.93	46.14
2010	30 sets	Design delta	1.07	-46.32	53.34
2010	5 sets	Stratified ratio	-5.09	-47.01	53.39
2010	5 sets	Design delta	2.82	-44.46	60.32
2010	Model	Model delta	19.08	-24.32	65.67

Table 2. Variables selected by BIC for each observation error, for one random draw of the USA-like simulated data with realistic observer coverage, with and without a Year:area random effect.

Model	No interaction	Random interaction
TMBbinomial	hbf + Year	hbf + Year + (1 | Year:area)
TMBdelta-Lognormal	area + hbf + season + Year	area + hbf + season + Year + (1 | Year:area)
TMBnbinom2	area + hbf + Year + offset(log(Effort))	area + hbf + Year + (1 | Year:area) + offset(log(Effort))
TMBnbinom1	area + hbf + Year + offset(log(Effort))	area + hbf + Year + (1 | Year:area) + offset(log(Effort))
TMBtweedie	area + hbf + Year	area + hbf + Year + (1 | Year:area)

Table 3. Median absolute relative error (MARE), mean relative error (MRE), and 95% confidence interval coverage for geostatistical models with four families and two random field configurations. The models are sorted by MARE. The non-spatial models ('Tweedie', 'NB2', 'NB1', and 'Delta-lognormal' below) are from the versions fit in Section 2.2.

(a) For all three fleets

Model	MARE	MRE	Coverage
NB2 sdmTMB spatial + spatiotemporal fields	0.04	0.02	0.93
Delta-lognormal sdmTMB spatial + spatiotemporal fields	0.04	0.04	0.90
NB1 sdmTMB spatial + spatiotemporal fields	0.04	0.02	0.93
Delta-lognormal sdmTMB spatial fields	0.05	0.05	0.79
NB1 sdmTMB spatial fields	0.05	0.03	0.76
NB2 sdmTMB spatial fields	0.05	0.02	0.76
Delta-gamma sdmTMB spatial fields	0.06	0.05	0.72
Delta-gamma sdmTMB spatial + spatiotemporal fields	0.06	0.04	0.83
Tweedie	0.08	0.01	0.62
NB2	0.10	0.03	0.55
Delta-lognormal	0.10	0.03	0.52
NB1	0.11	0.06	0.52

(b) For the USA-like fleet only

Model	MARE	MRE	Coverage
Delta-lognormal sdmTMB spatial fields	0.09	-0.01	1.00
Delta-gamma sdmTMB spatial fields	0.09	-0.01	1.00
NB1 sdmTMB spatial fields	0.11	-0.01	0.96
NB1 sdmTMB spatial + spatiotemporal fields	0.11	-0.01	0.96
NB2 sdmTMB spatial + spatiotemporal fields	0.12	-0.01	0.96
NB2	0.12	0.01	0.96
NB2 sdmTMB spatial fields	0.12	-0.01	0.96
Tweedie	0.12	0.10	1.00
Delta-lognormal	0.13	0.06	1.00
NB1	0.15	0.01	1.00

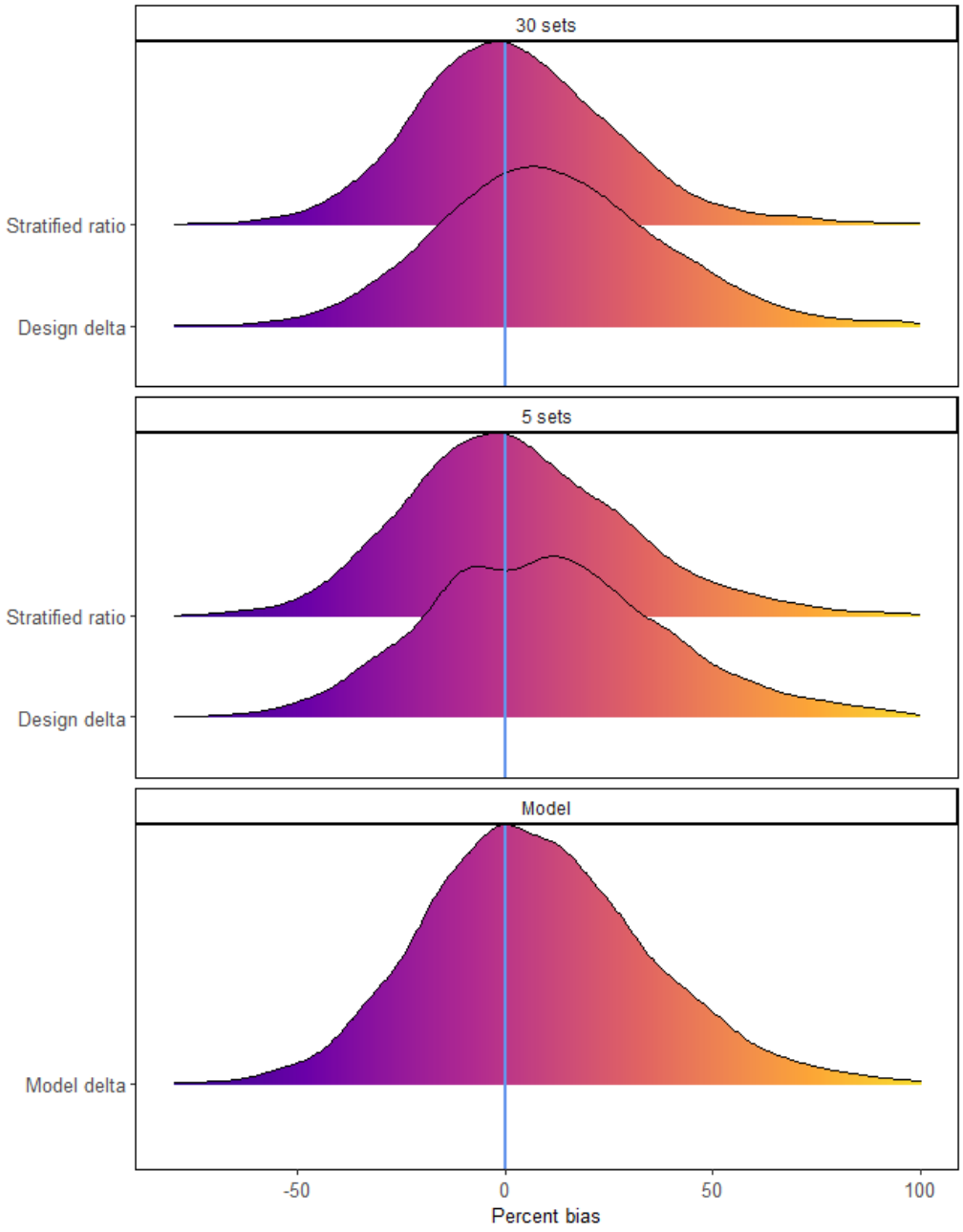

Figure 1. Percent bias across all years and runs by estimator for 100 simulations from the USA-like fleet. Panels indicate whether the design estimators were pooled below a minimum size of 30 sets or 5 sets, or whether the method is model-based.

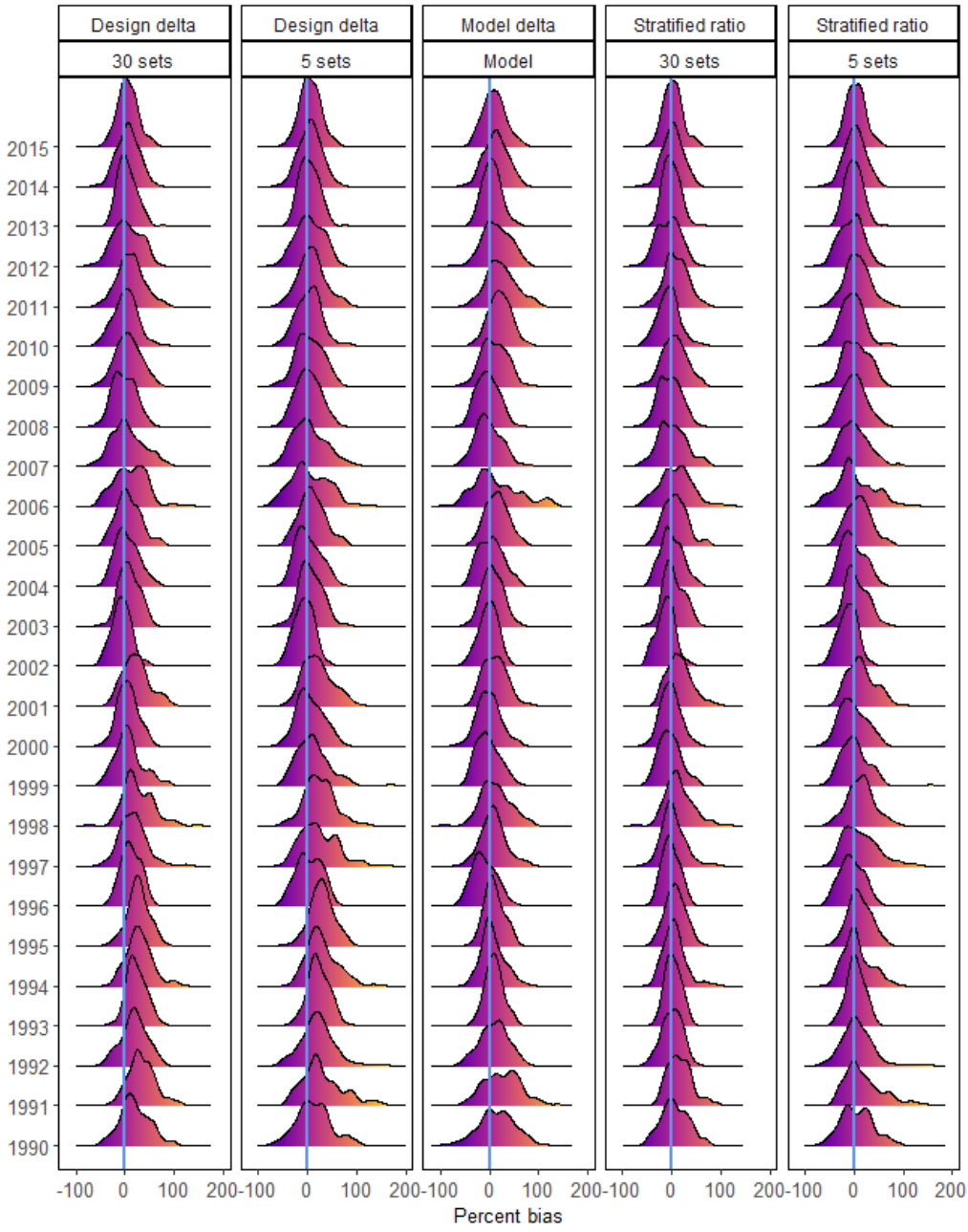

Figure 2. Percent bias across runs by estimator for 100 simulations from the USA-like fleet. Panels indicate whether the design estimators were pooled below a minimum size of 30 sets or 5 sets, or whether the method is model-based.

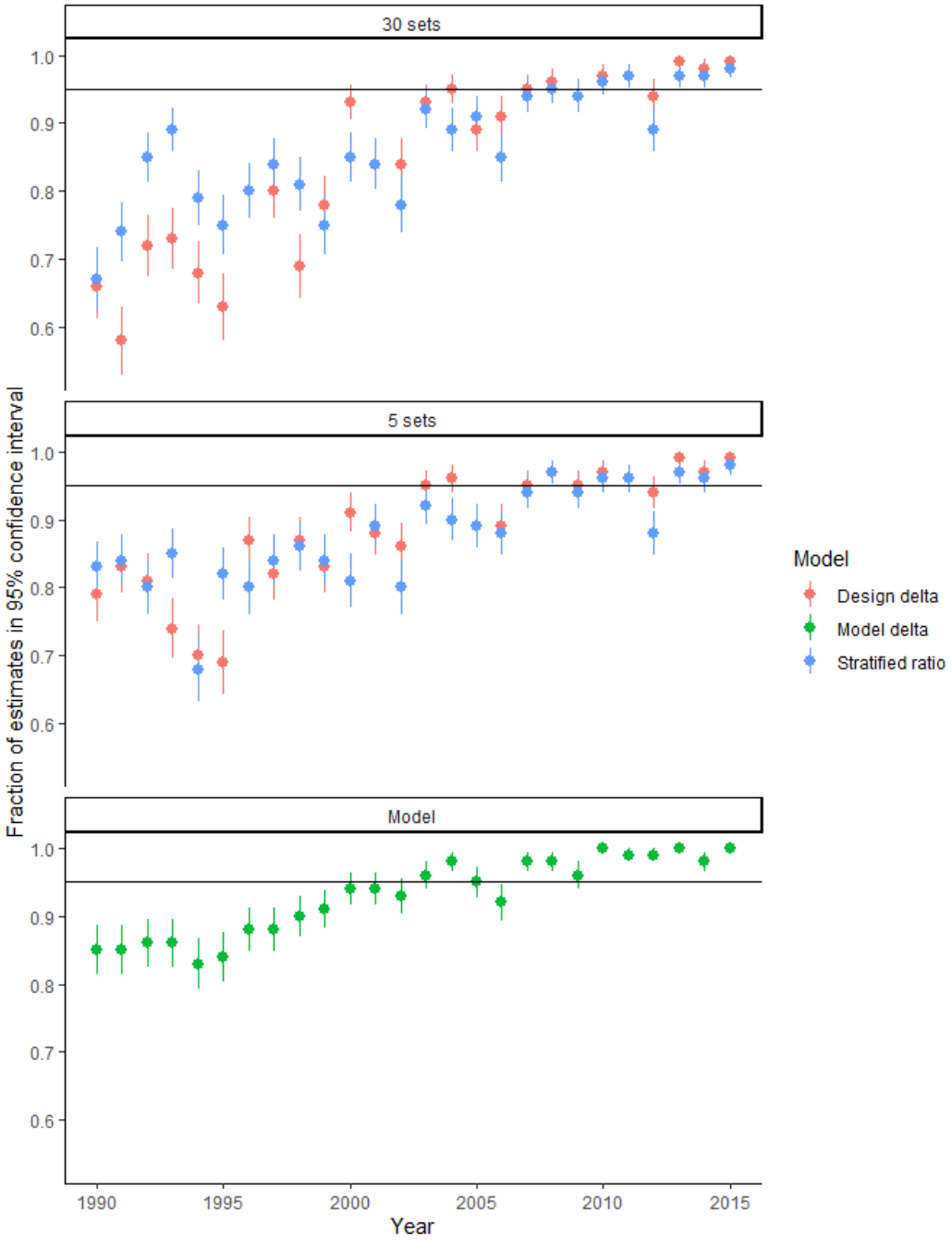

Figure 3. Fraction of the draws that included the true value in the confidence interval in each year for each method. Panels indicate whether the design estimators were pooled below a minimum size of 30 sets or 5 sets, or whether the method is model-based.

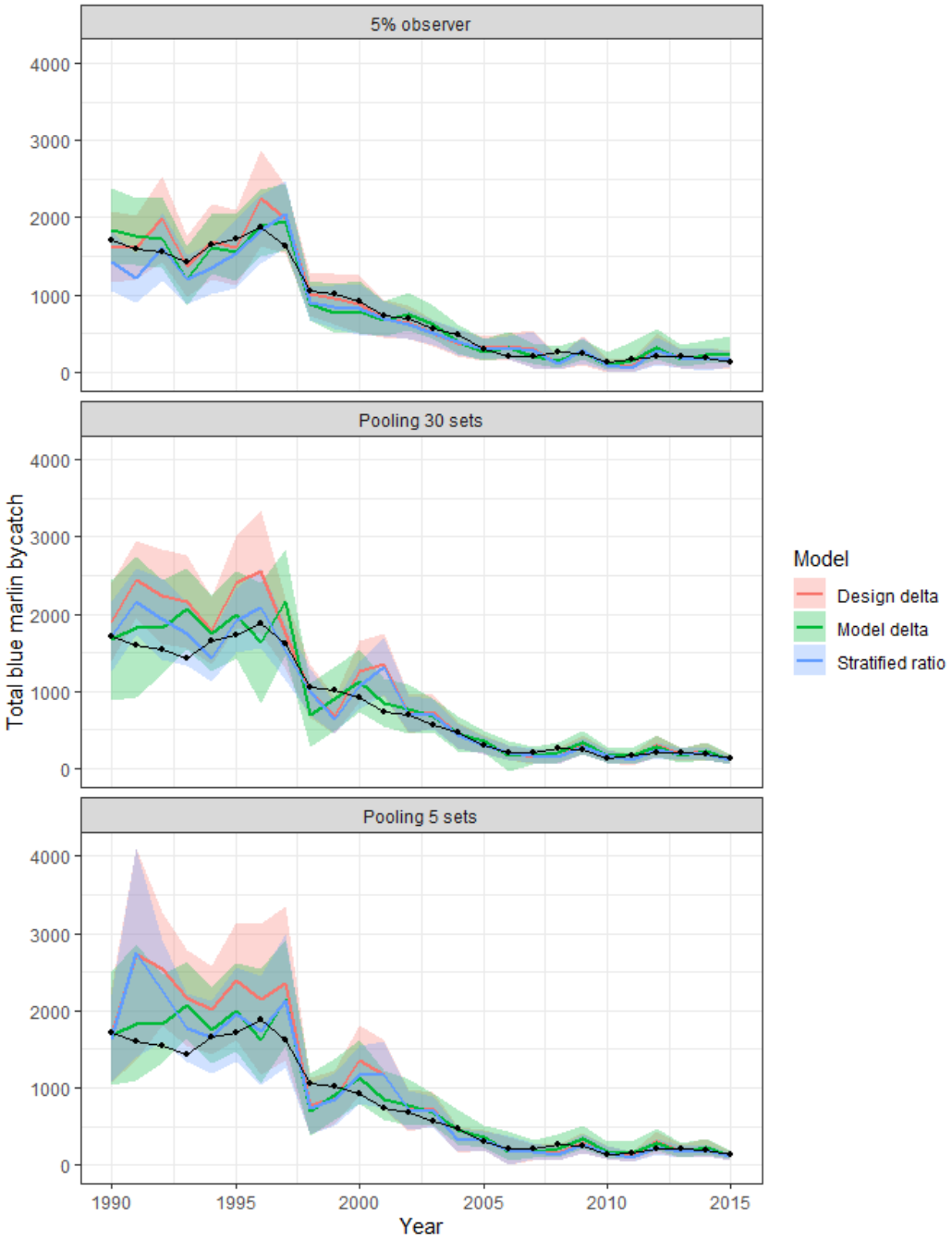

Figure 4. Example total bycatch estimates from one run of the estimator for the USA-like simulations, for a random 5% observer coverage, and for the realistic stratified observer allocation, pooling if the sample in a stratum is less than either 30 or 5 sets. True total bycatch is the black line.

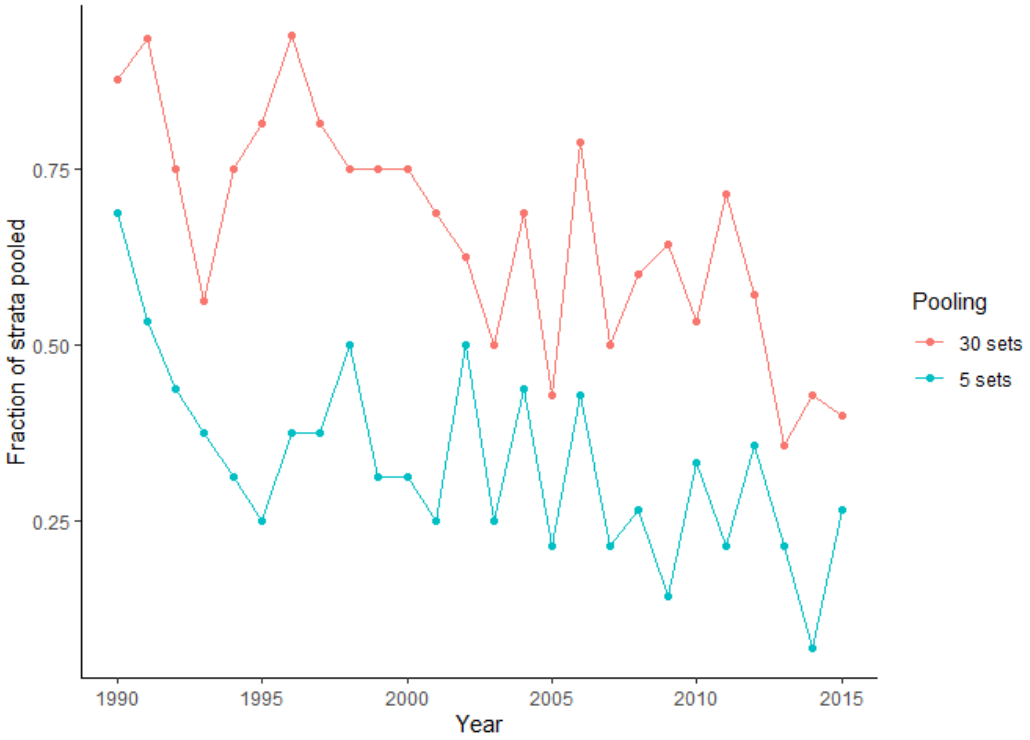

Figure 5. Fraction of strata that had to be pooled when the minimum sample size cutoff was 30 sets vs. 5 sets.

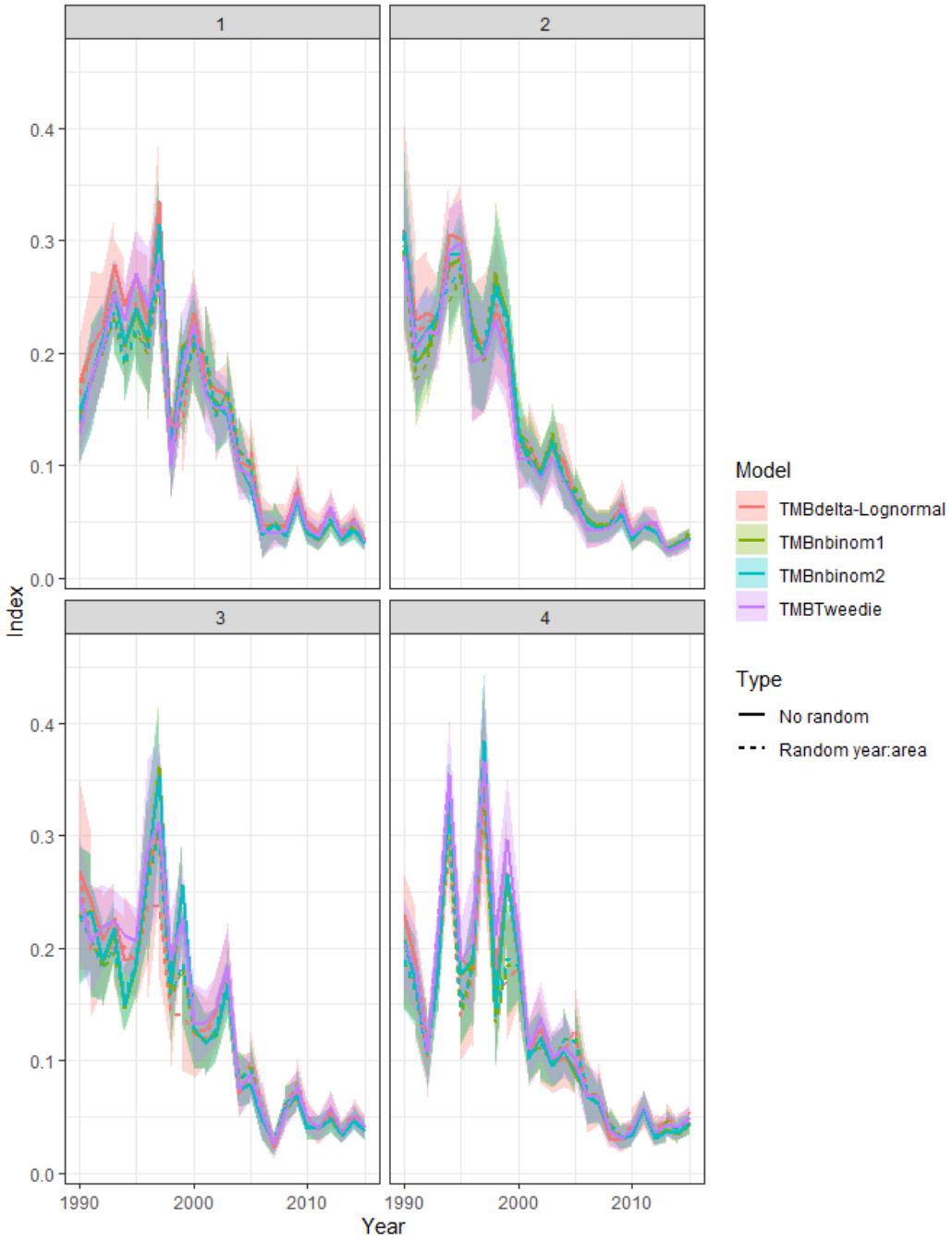

Figure 6. Abundance indices calculated by multiple methods from the USA-like simulation observer data. Panels are 4 random draws of the realistic observer sampling program.

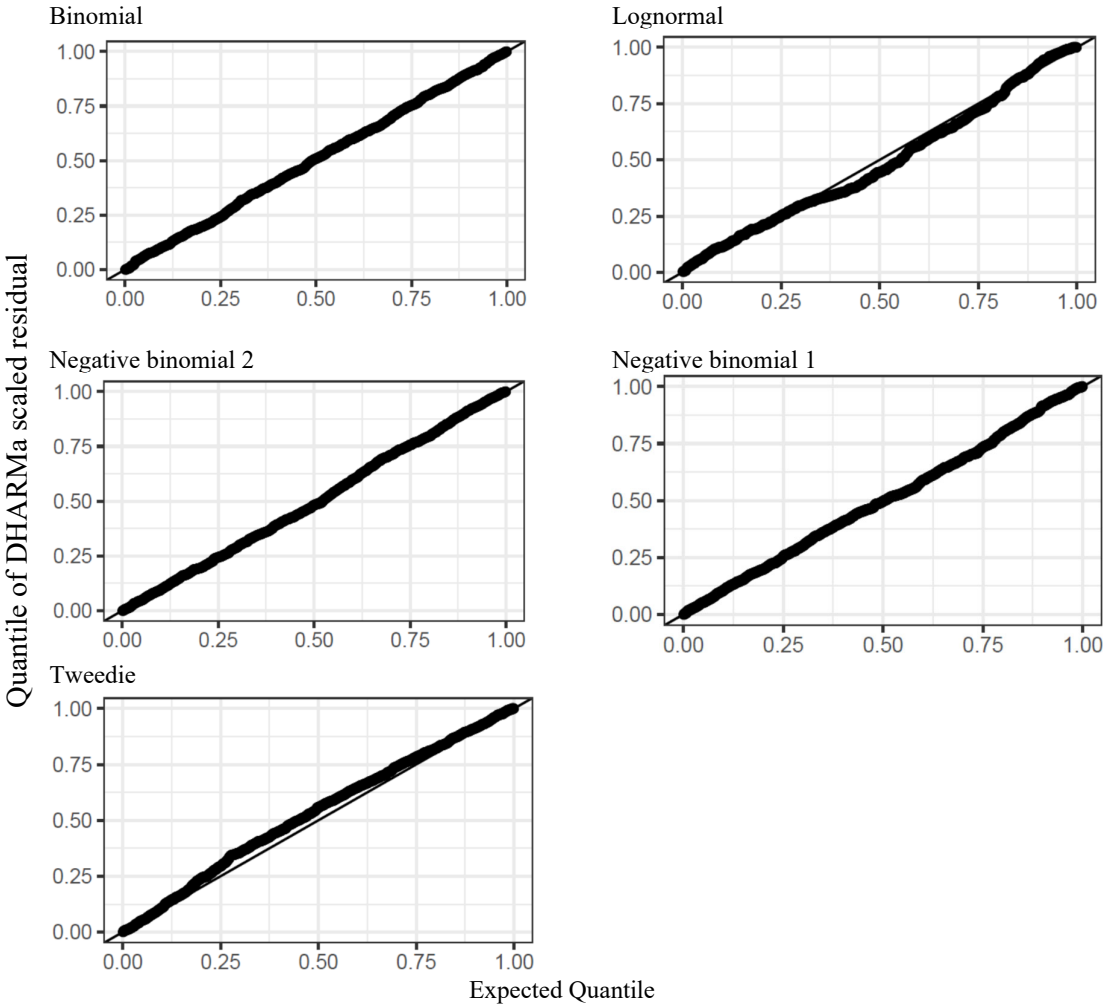

Figure 7. Residual QQ plots for one draw from the USA-like simulation with realistic observer coverage.

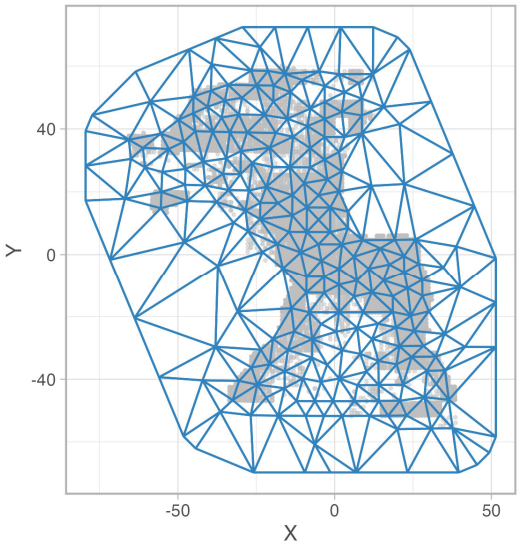

Figure 8. Mesh used in the SPDE approximation and bilinear interpolation for the full-dataset simulation across all three fleets. The grey dots in the background represent the simulated observer data locations. This mesh uses a cutoff of 5 with 100 km X-Y units, which means no triangle edge is allowed to be smaller than 500 km. This mesh has approximately 250 vertices or knots.

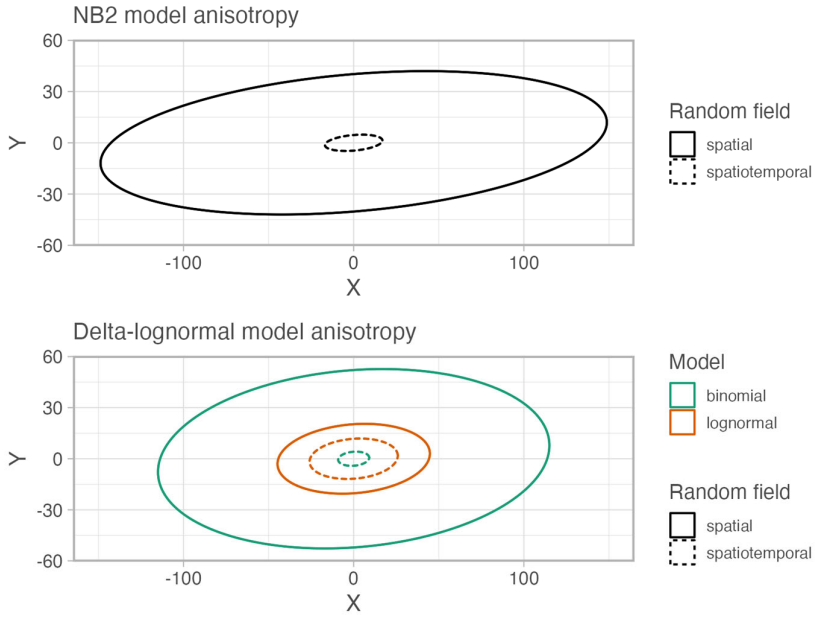

Figure 9. Visualization of estimated anisotropy for the NB2 and delta-lognormal models. The ellipses indicate spatial and spatiotemporal range parameters in any direction from zero in the middle. The range is the distance at which two data points are effectively independent (about 0.13 correlation). The units are the units of the X and Y coordinates, which here are 100 km within the Albers projection.

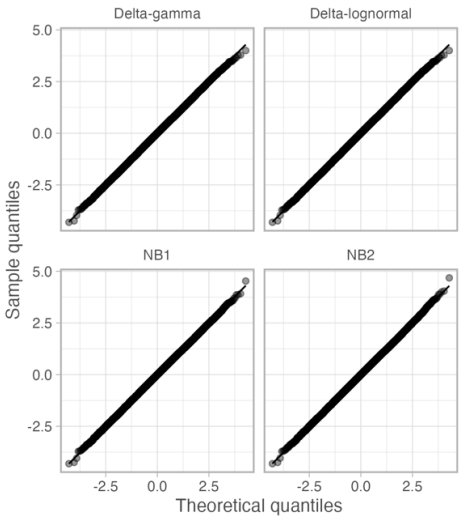

Figure 10. Randomized quantile residuals for the four models with spatial and spatiotemporal random fields.

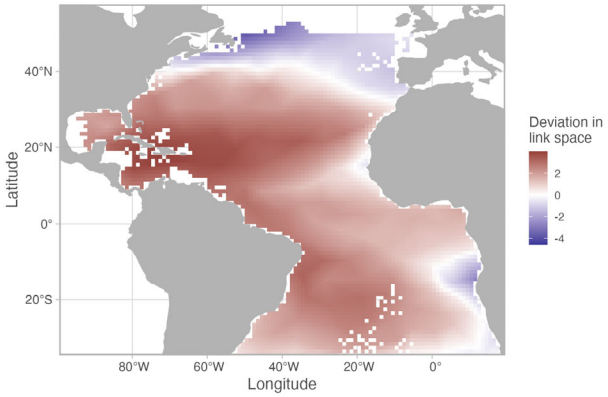

Figure 11. Spatial random field values for the NB2 model. These are deviations in link space and represent spatially correlated effects from latent variables not included in the model.

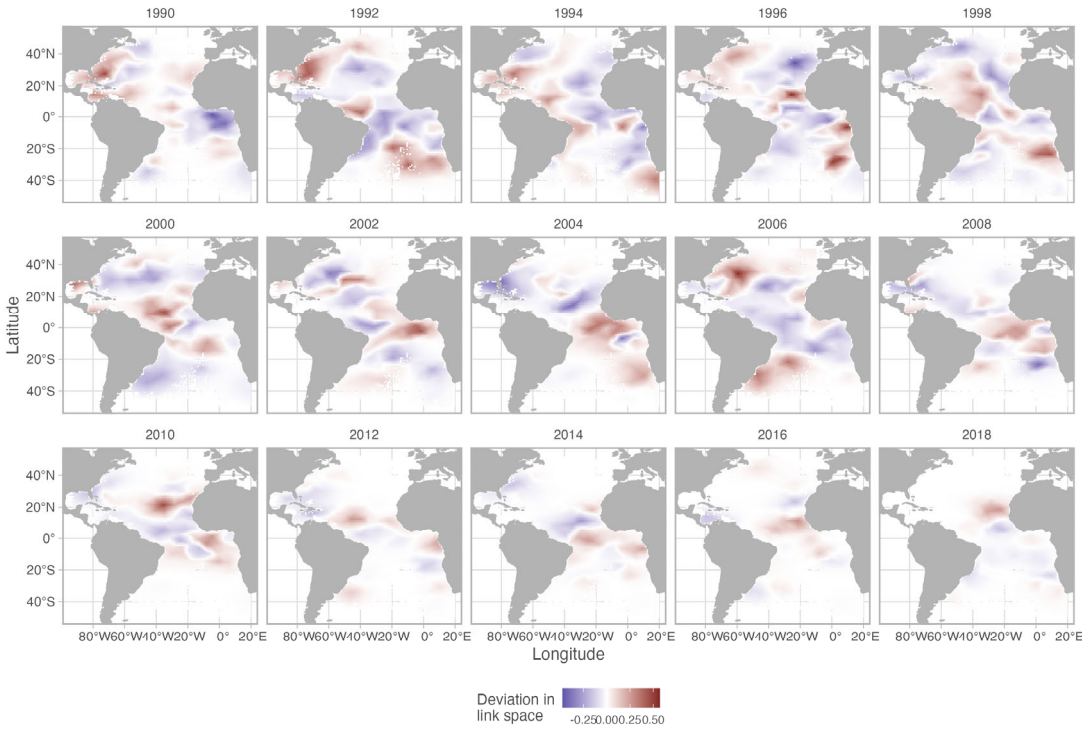

Figure 12. Spatiotemporal random field values for the NB2 model. These are deviations in link space and represent spatially correlated effects from latent variables not included in the model that change each year. Every second year is omitted here to save space. The model assumes these random effects to be independent each year.

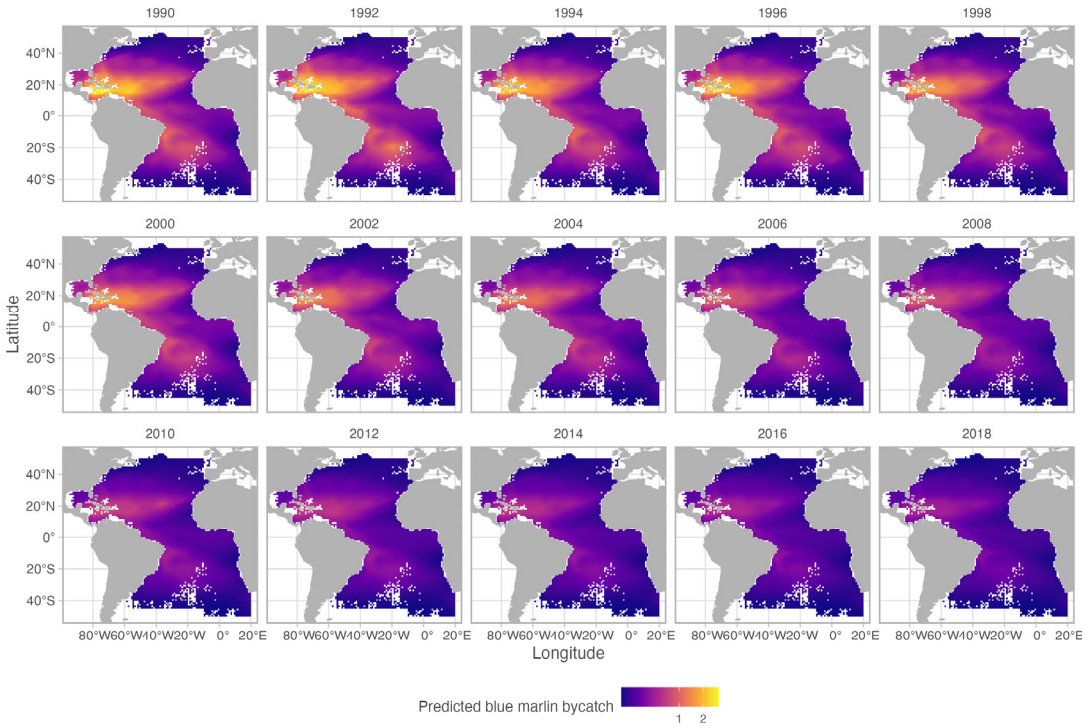

Figure 13. Overall predictions in space from the NB2 model shown for every second year. These predictions are created by summing all fixed and random effect components of the model.

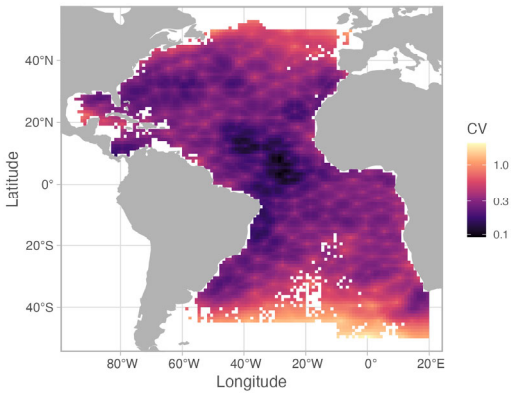

Figure 14. The coefficient to variation (CV) on predicted bycatch in space from the NB2 spatiotemporal model for an example year. The spatial uncertainty looks similar across years. The mesh-pattern observed in the CV is a known artifact of the bilinear interpolation—the prediction uncertainty is most accurate at the triangle vertices and their connecting lines.

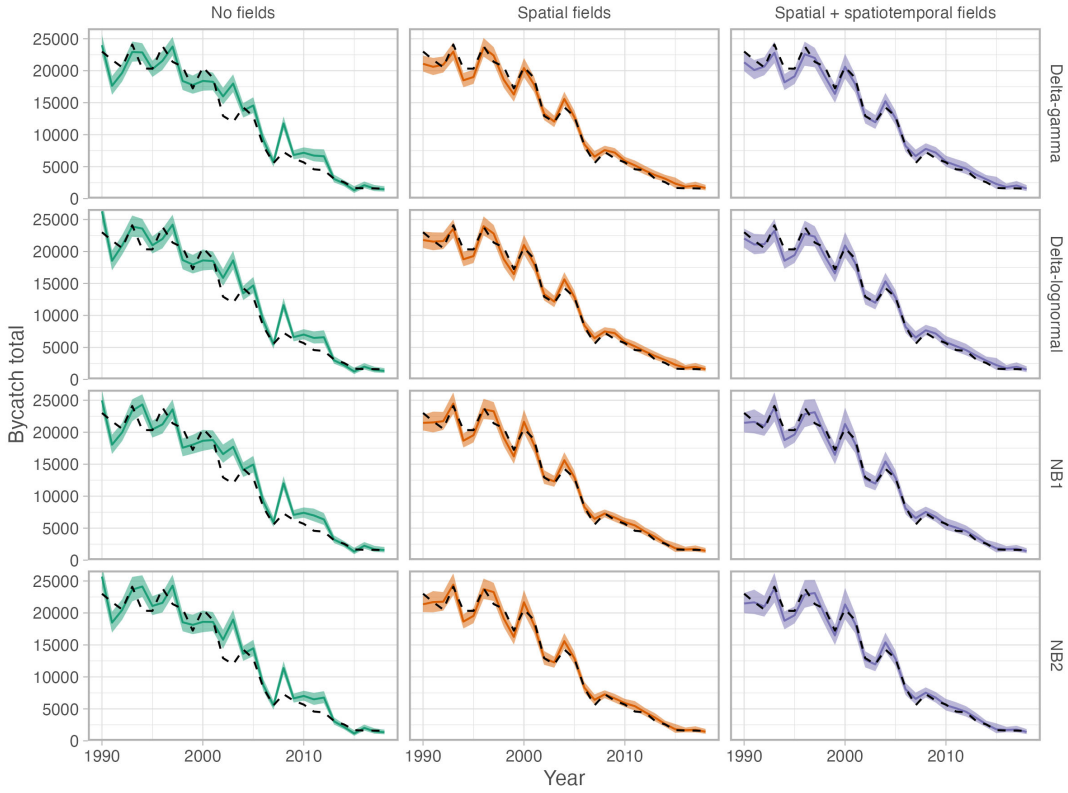

Figure 15. Total predicted blue marlin bycatch for four families (rows) and three random field configurations (columns) for the full dataset, summed across all three fleets. The true total is shown with a dashed black line. For the model estimates, lines represent mean estimates and ribbons represent 95% confidence intervals.

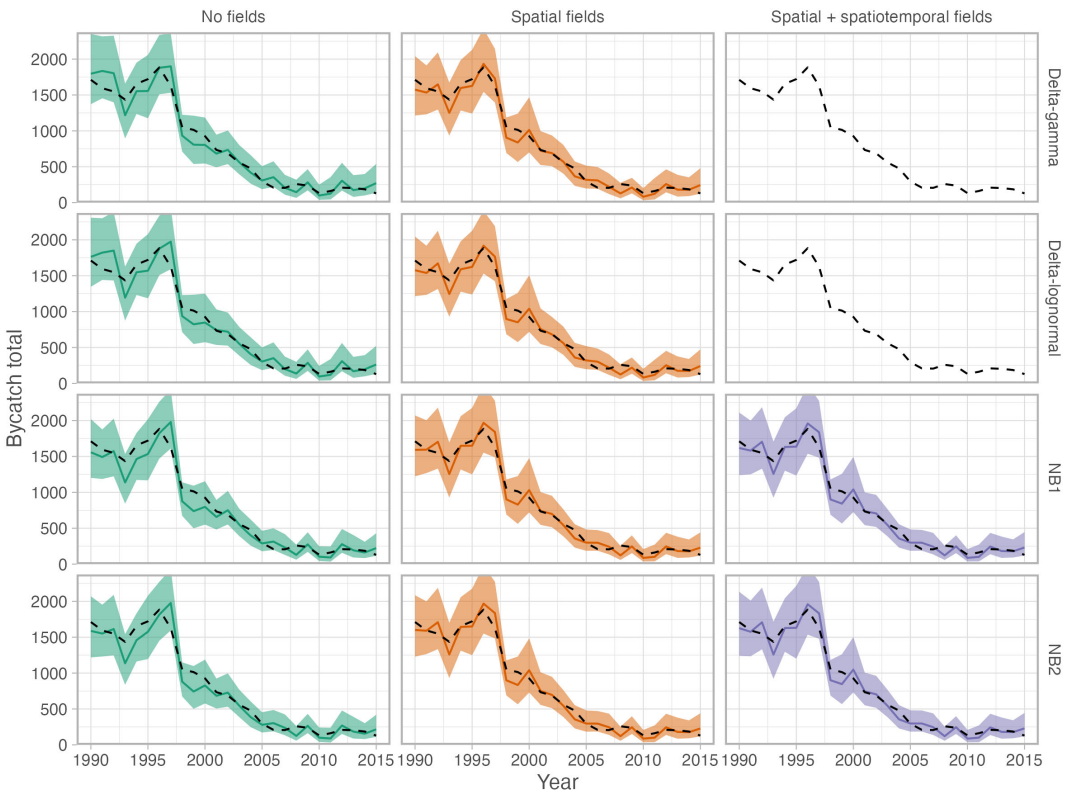

Figure 16. Total predicted blue marlin bycatch for four families (rows) and three random field configurations (columns) **for the USA-like fleet only**. The true total is shown with a dashed black line. For the model estimates, lines represent mean estimates and ribbons represent 95% confidence intervals. The spatial + spatiotemporal fields delta-lognormal and delta-gamma models are omitted here because the spatiotemporal standard deviation collapsed to zero suggesting the models with only spatial fields were appropriate.

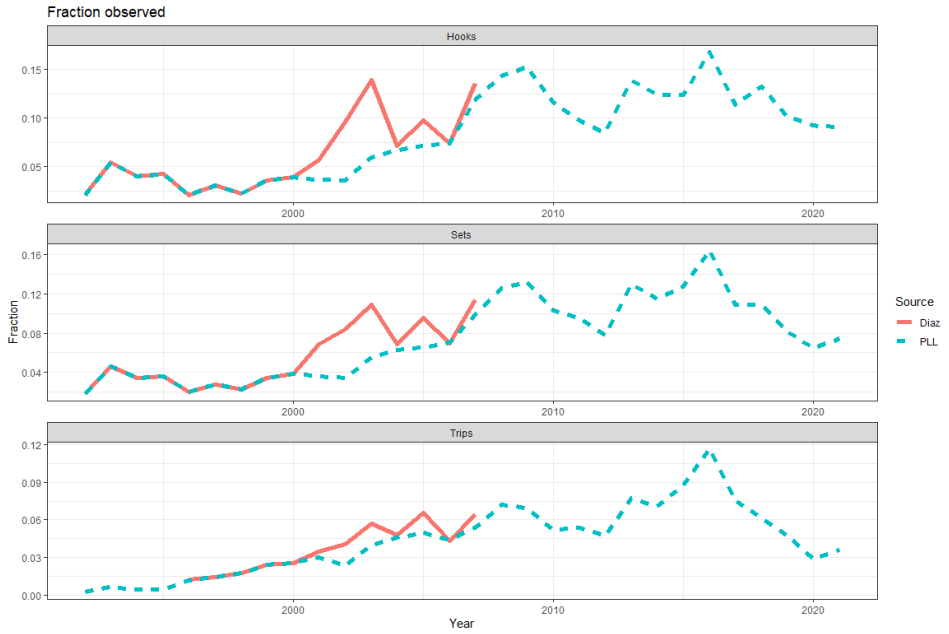

Figure 17. Comparison of fraction of Observed to total Pelagic Longline hooks, sets, and trips, as contained in the datasets used in this analysis and those reported by Diaz et al. (2009).

Figure 18. Comparison of total effort in the Pelagic Longline logbook data as contained in the datasets used in this analysis and those reported by Diaz et al. (2009).

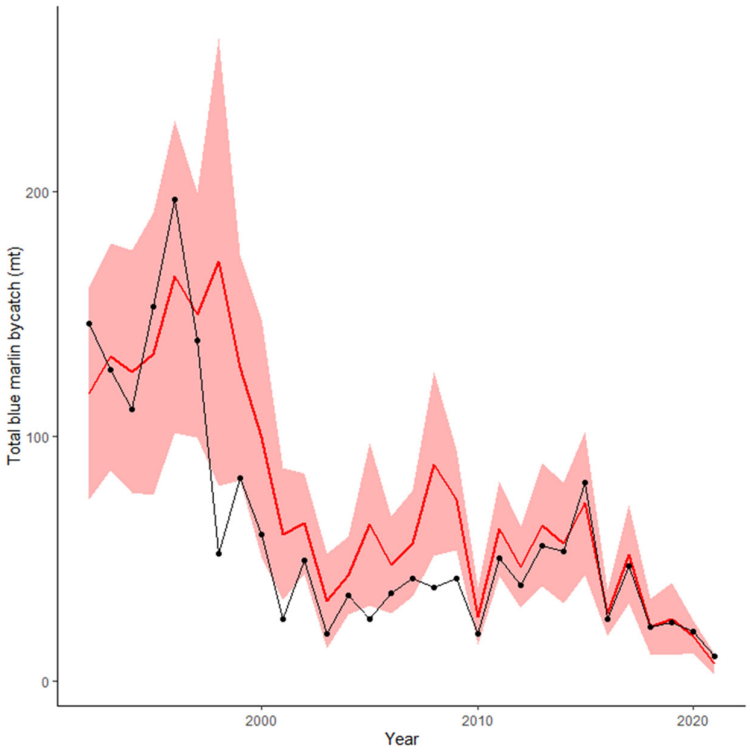

Figure 19. Delta lognormal estimates with confidence intervals (red line and shading) as compared to the US Task 1 reported bycatch estimates (solid black line). Minimum sample of 30.

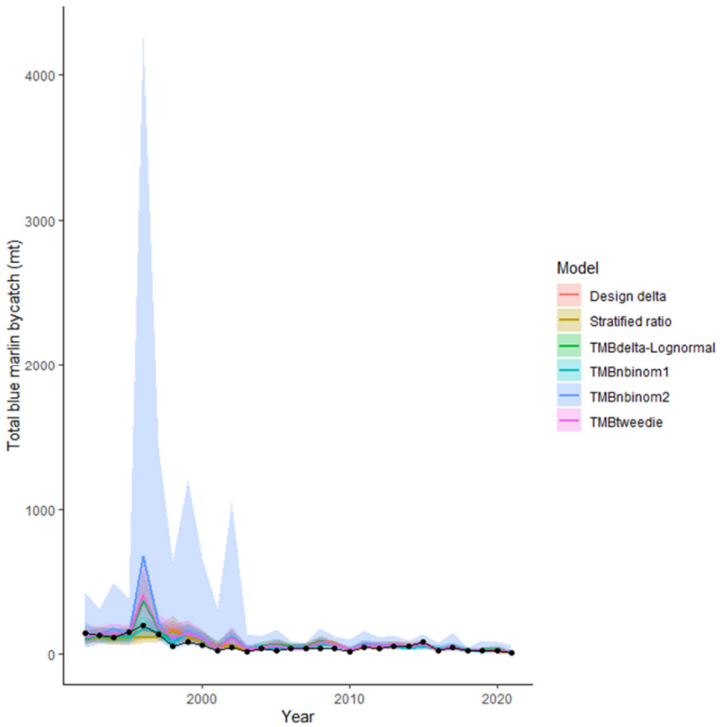

Figure 20. Bycatch estimates as calculated by multiple methods. USA Task 1 data are presented as the solid black line.

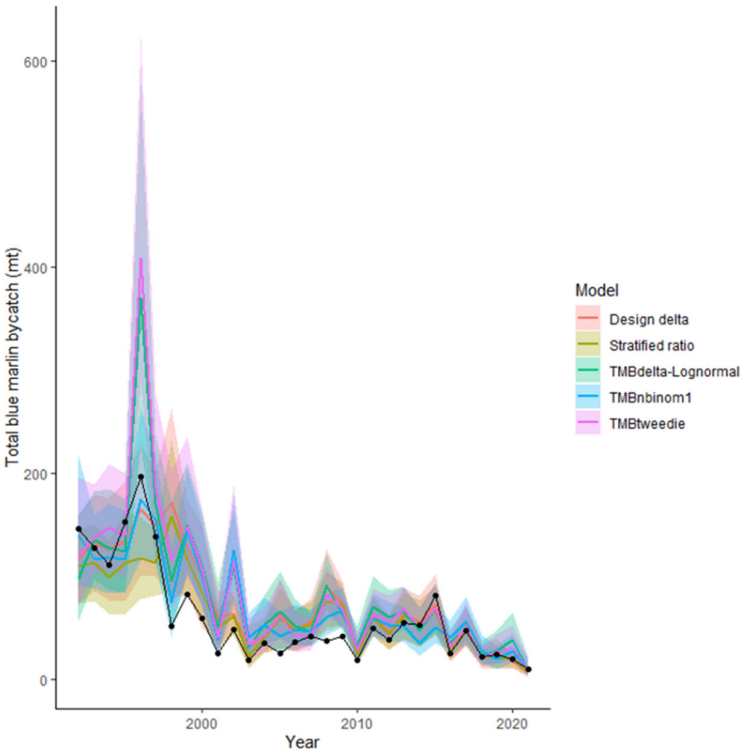

Figure 21. Bycatch estimates as calculated by multiple methods (TMB Binomial2 model results removed). USA Task 1 data are presented as the solid black line.