Evaluation of the impacts of different treatments of spatio-temporal variation in catch-per-unit-effort standardization models

Arnaud Grüss[1, 2, a*], John F. Walter III[3, b], Elizabeth A. Babcock[1, c], Francesca C. Forrestal[4, d], James T. Thorson[5, e], Matthew V. Lauretta[3, f], Michael J. Schirripa[3, g]

[1]Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL, 33149, USA

[2]School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA, 98105-5020, USA

[3]Southeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 75 Virginia Beach Drive, Miami, FL, 33149-1099, USA

[4]Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL, 33149, USA

[5]Habitat and Ecosystem Process Research program, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 7600 Sand Point Way N.E., Seattle, WA 98115, USA

Author email addresses
[a]gruss.arnaud@gmail.com
[b]john.f.walter@noaa.gov
[c]ebabcock@rsmas.miami.edu
[d]fforrestal@rsmas.miami.edu
[e]James.Thorson@noaa.gov
[f]matthew.lauretta@noaa.gov
[g]michael.schirripa@noaa.gov

Keywords: Catch-per-unit-effort (CPUE); standardization methods; indices of relative abundance; simulation-testing; spatio-temporal models

Funding: This work was supported in part by a NOAA grant through the Cooperative Institute for Marine and Atmospheric Studies at the University of Miami [grant number NA150AR4320064]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Corresponding author
Dr. Arnaud Grüss
School of Aquatic and Fishery Sciences
University of Washington
Box 355020
Seattle, WA, 98105-5020
United States of America
Telephone: (01) 305 606 5696
Email: gruss.arnaud@gmail.com

ABSTRACT

Many stock assessments heavily rely on indices of relative abundance derived from fisheries-dependent catch-per-unit-effort (CPUE) data. Therefore, it is critical to evaluate different CPUE standardization methods under varying scenarios of data generating processes. Here, we evaluated nine CPUE standardization methods offering contrasting treatments of spatio-temporal variation, ranging from the basic generalized linear model (GLM) method not integrating a year-area interaction term to a sophisticated method using the spatio-temporal modeling platform VAST. We compared the performance of these methods against simulated data constructed to mimic the processes generating fisheries-dependent information for Atlantic blue marlin (*Makaira nigricans*), a common bycatch population in pelagic longline fisheries. Data were generated using a longline data simulator for different population trajectories (increasing, decreasing, and static). These data were further subsampled to mimic an observer program where trips rather than sets form the sampling frame, with or without a bias towards trips with low catch rates, which might occur if the presence of an observer alters fishing behavior to avoid bycatch. The spatio-temporal modeling platform VAST achieved the best performance in simulation, namely generally had one of the lowest biases, one of the lowest mean absolute errors (MAEs), and 50% confidence interval coverage closest to 50%. Generalized additive models accounting for spatial autocorrelation at a broad spatial scale (one of the lowest MAEs and one of the lowest biases) and, to a lesser extent, non-spatial delta-lognormal GLMs including a year-area interaction as a random effect (one of the lowest MAEs and one of the best confidence interval coverages) also performed adequately. The VAST method provided the most comprehensive and consistent treatment of spatio-temporal variation, in contrast with methods that simply weight predictions by large spatial areas, where it is critical, but difficult, to get the *a priori* spatial stratification correct before weighting. Next, we applied the CPUE standardization methods to

real data collected by the National Marine Fisheries Service Pelagic Observer Program. The indices of relative abundance predicted from real observer data were relatively similar across CPUE standardization methods for the period 1998-2017 and suggested that the blue marlin population of the Atlantic declined over the period 1998-2004 and was relatively stable afterwards. As spatio-temporal variation related to environmental changes or depletion becomes increasingly necessary to consider, greater use of spatio-temporal models for standardizing fisheries-dependent CPUE data will likely be warranted.

1. Introduction

Stock assessments, and subsequent fisheries management advice, rely largely on fisheries-dependent data, i.e., data that are collected with the assistance of fishers (Maunder and Punt, 2004). Many stock assessment models use indices of relative abundance to fit predicted fish abundances or biomasses and to estimate stock parameters (Maunder and Starr, 2003; Lynch et al., 2012). Nearly all of the indices of relative abundance employed in the stock assessments of highly migratory populations and other fish populations lacking fisheries-independent surveys are derived from fisheries-dependent catch-per-unit-effort (CPUE) data (Bishop, 2006; Maunder et al., 2006; Walter et al., 2014a). However, as fisheries do not randomly sample fish stocks, it is necessary to "standardize" fisheries-dependent CPUE data to account for confounding factors that influence catchability which, if not accounted for, could result in a non-proportional relationship between fisheries-dependent CPUE and true stock abundance (Walters, 2003; Maunder and Punt, 2004; Ye and Dennis, 2009). Various methods have been developed to perform CPUE standardization (Maunder and Punt, 2004). To improve confidence in stock assessment outcomes and the fisheries management decisions based on these outcomes, it is critical to evaluate and compare CPUE standardization methods under different scenarios about fish abundance trends and the distribution of fish and fishing effort across time and space (Bigelow and Maunder, 2007; Goodyear, 2003; Lynch et al., 2012; Campbell, 2015).

Conventional methods for standardizing CPUE data consist of fitting generalized linear models (GLMs; McCullagh and Nelder, 1989), generalized additive models (GAMs; Wood, 2006) or generalized linear mixed models (GLMMs; Breslow and Clayton, 1993) integrating covariates influencing catchability to CPUE data. Often, the GLMs used for CPUE standardization simply include fixed year and area effects (e.g., the GLMs employed for standardizing the CPUE data of highly migratory species such as blue marlin (*Makaira*

107 *nigricans*); Forrestal et al., 2017). Hereafter, this basic CPUE standardization method is
108 referred to as the "GLM" method (Table 1). The GLMs and GLMMs used for CPUE
109 standardization sometimes also include a year-area interaction term when it is thought that
110 annual trends in abundance may differ among areas of the study region (e.g., Nakano, 1989;
111 Chang, 2003; Miyabe and Takeuchi, 2003; Forrestal et al., 2017). In their seminal paper,
112 Maunder and Punt (2004) emphasized that the appropriate way to deal with year-area
113 interactions is either to employ GLMMs integrating the year-area interaction term as a
114 random effect (henceforth the "GLMMint" method), or to use GLMs integrating the year-area
115 interaction term as a fixed effect and then weight GLM predictions for the individual area
116 strata by the surface area of these area strata (see below). Employing GLMs integrating a
117 year-area interaction term as a fixed effect and not weighting GLM predictions for the
118 individual area strata by the surface area of these area strata (henceforth the "GLMint"
119 method) negates the interest of the year-area interaction term, as the index of relative
120 abundance will then be dependent and vary upon the specific area stratum chosen (Maunder
121 and Punt, 2004; Lynch et al., 2012; Campbell, 2015).

122 The CPUE standardization methods that take into account the surface area of the areas
123 making up the study region to weight CPUE observations have been studied in detail in
124 Campbell (2004, 2015). Hereafter, we refer to these methods as the "GLMwt" and
125 "GLMwt.int" methods, depending on whether they integrate a fixed year-area interaction term
126 or not. With the GLMwt and GLMwt.int methods, first, CPUE data are standardized for
127 individual areas and years, then they are multiplied by the surface areas of their respective
128 areas and, finally, an index of relative abundance is computed as the sum of the products of
129 standardized CPUE data and surface areas (Campbell, 2004, 2015; Maunder and Punt, 2004).
130 In addition to promoting the weighting of the year-area interactions by the surface area of
131 each area of the study region, Campbell (2004, 2015) argued that weights should be assigned

to raw CPUE data based on the year-area stratum to which they belong when the number of observations in each year-area stratum varies substantially. Assigning prior weights to raw CPUE data allows for a balanced dataset for GLM-parameter estimation (Campbell, 2015). Hereafter, we refer to the CPUE standardizing methods assigning prior weights to raw CPUE data as the "GLMprwt" and "GLMprwt.int" methods, depending on whether they integrate a fixed year-area interaction term or not.

Some CPUE standardization methods offer a more sophisticated treatment of spatio-temporal variation by accounting for spatial and/or spatio-temporal autocorrelation. GAMs can account for spatial autocorrelation at a broad spatial scale through the integration of an interaction term between eastings and northings (i.e., longitude and latitude expressed in UTM coordinates), and for spatio-temporal autocorrelation at a broad spatial scale by nesting the year effect within the interaction term between eastings and northings (Su et al., 2011; Grüss et al., 2016, 2019). Hereafter, we refer to the CPUE standardizing methods using GAMs accounting for spatial and/or spatio-temporal autocorrelation as the "GAM" and "GAMint" methods, depending on whether they account for spatio-temporal autocorrelation at a broad spatial scale or not. Spatio-temporal models take a step further and exploit the property of spatial and spatio-temporal structure at a fine spatial scale to then predict variables of interest (Thorson et al., 2015; Grüss et al., 2017). Recent years have seen the emergence of spatio-temporal modeling methods for standardizing CPUE data (e.g., Pereira et al., 2012; Berg et al., 2014; Walter et al., 2014b; Thorson et al., 2015; Cao et al., 2017). Due to their properties, spatio-temporal models are particularly compelling for standardizing the CPUE data obtained from observers, i.e., the trained personnel placed on fishing boats to collect data. In fact, the data collected by observers are often clustered since they tend to be repeated samples from the same fishing boats at similar sites, and they cover only a limited spatial and temporal extent of the fishery of interest (Beerkircher et al., 2002; Walter et al., 2014b).

157 Furthermore, observer data could have sampling bias, as fishing boats with observers on
158 board may try to avoid locations where bycatch is high (Benoît and Allard, 2009; Walter et
159 al., 2014b), above and beyond the inherent potential biases of fisheries-dependent data.

160 Data simulators are valuable tools for evaluating CPUE standardization methods as
161 they allow for a known true annual trend in fish abundance (Lynch et al., 2012; Forrestal et
162 al., 2019b). Over recent years, several simulation analyses have been carried out for
163 evaluating and comparing CPUE standardization methods (e.g., Carruthers et al., 2010, 2011;
164 Lynch et al., 2012; Pereira et al., 2012; Ono et al., 2015; Thorson et al., 2016; Forrestal et al.,
165 2017, 2019b). For example, Carruthers et al. (2011) employed spatial production models to
166 simulate theoretical commercial fisheries, and then compared the performance of variants of
167 the GLM method applied to CPUE data from the theoretical commercial fisheries. Another
168 example is that of Lynch et al. (2012), who developed a data simulator for running a
169 comparison of the accuracy of the GLM method and an habitat-based standardization method
170 applied to CPUE data from the Atlantic Japanese longline fishery. No published study has
171 utilized simulation analysis to compare the performance of CPUE standardization methods
172 offering contrasting treatments of spatio-temporal variation (e.g., GLM vs. GLMMint vs.
173 GAM vs. spatio-temporal method).

174 In this study, we evaluated and compared nine CPUE standardization methods offering
175 contrasting treatments of spatio-temporal variation (Table 1), ranging from the basic GLM
176 method to a sophisticated method using the Vector Autoregressive Spatio-temporal Model
177 (henceforth the "VAST" method; Thorson, 2019). We applied these nine CPUE
178 standardization methods to Atlantic blue marlin CPUE data collected by fisheries observers.
179 Blue marlin is a large, highly migratory species of substantial importance to recreational and
180 artisanal fisheries and primarily a bycatch species of open-ocean longline fleets (Sharma et
181 al., 2017). Firstly, we evaluated the nine CPUE standardization methods utilizing simulated

182 data from the U.S. pelagic longline fishery developed with the LLSIM data simulator
183 (Forrestal et al., 2017; Goodyear et al., 2017). Next, the CPUE data from the simulated
184 pelagic longline fishery were subsampled to mimic sampling by an observer program. We
185 either randomly subsampled 10% of the trips, or we selected 10% of the trips such that trips
186 with lower than average catch rate were selected in a higher proportion, resulting in a biased
187 sample that might reflect the process of an observer bias, where fishing trips with observers
188 tend to avoid locations with high bycatch rates. We then applied the CPUE standardization
189 methods to the subsampled CPUE data in a design where the model developer (Arnaud Grüss)
190 did not know any details regarding the LLSIM simulations and the environmental conditions
191 in the system simulated in LLSIM. Secondly, we applied the contrasting CPUE
192 standardization methods to CPUE data collected by the National Marine Fisheries Service
193 (NMFS) Pelagic Observer Program (Beerkircher et al., 2002) over the period 1992-2017.

194

195 **2. Material and methods**

196 ***2.1. Study region***

197 Our study region is the portion of the North Atlantic shown in Fig. 1. This region
198 encompasses the ten NMFS areas defined for stock assessments of the International
199 Commission for the Conservation of Atlantic Tunas (ICCAT) (Fig. 1): (1) the Gulf of Mexico
200 (GOM); (2) the Mid Atlantic Bight (MAB); (3) the South Atlantic Bight (SAB); (4) Florida
201 East Coast (FEC); (5) the Caribbean (CAR); (6) the Northeast Coastal area (NEC); (7) the
202 Sargasso area (SAR); (8) the Northeast Distant area (NED); (9) the North Central Atlantic
203 (NCA); and (10) the Offshore South area (OFS). To be able to utilize the GAM and VAST
204 methods, we produced a 1° x 1° spatial grid covering the entire study region, and we

estimated the surface area of the cells of that spatial grid. The spatial grid for the North Atlantic includes 3,079 cells.

2.2. LLSIM data

In the present study, we employed the longline CPUE data simulator LLSIM (Forrestal et al., 2017, 2019a; Goodyear, 2017; Goodyear et al., 2017). In brief, the core of LLSIM is the computation of the catch of the U.S. pelagic longline fishery on a single hook of a longline set (Forrestal et al., 2017). Each hook is characterized by a depth distribution and a geographical position (latitude, and longitude) and is associated with a specific year, month, fraction of daylight and position along the longline. All the characteristics of the hook are associated with the individual longline set. LLSIM simulates the catch of the pelagic longline fishery as a stochastic process for each of the hooks of each longline set. The region covered by LLSIM extends from -35° latitude to 55° and from -95° longitude to 20°; however, only LLSIM data for the portion of that region shown in Fig. 1 were considered in this study. The region covered by LLSIM is broken down into 1° x 1° cells, which each includes 46 depth data. To make computations, LLSIM integrates fish population size, a gear coefficient and a habitat coefficient for each longline set. In each of the 1° x 1° cells, the habitat coefficient integrates the hook-depth probabilities with fish relative density in each of the 46 depth strata apportioned by the fraction of the longline sets that operate in hours of daylight and darkness. The hook-depth probabilities are derived from the measurements made by time-depth recorders attached to longlines of the U.S. pelagic longline fleet (Goodyear, 2017). The three-dimensional patterns of fish density considered by LLSIM come from a volume weighted habitat suitability model developed in Goodyear (2016). Goodyear (2016)'s habitat suitability model uses information on blue marlin oxygen tolerance from Brill (1994)'s study, as well as

temperature utilization and diel ΔT patterns from tagged blue marlins, to determine the three-dimensional patterns of blue marlin density from environmental data from a coupled ocean-biogeochemical model.

The LLSIM data employed in the present study were for three virtual blue marlin populations that had the exact same characteristics except that one maintained a constant abundance over time (Population 1), one was generally declining (Population 2) and the third one was generally increasing (Population 3) (Figs. 2a-c). LLSIM provided us with data for 294,305 longline sets for the U.S. pelagic longline fishery for each of the three populations, which covered the period 1987-2015. Catch was expressed as the number of blue marlins caught during the longline set, and fishing effort was expressed as the number of hooks in the set. CPUE was then the number of blue marlins caught per 1,000 hooks. In addition to catch and fishing effort data, LLSIM provided values for a number of parameters, including year, season, the type of hook used, the number of light sticks used, the type of bait used, and the number of hooks between floats (Table 2). NMFS areas were assigned based on the latitude and longitude associated with each simulated longline set.

LLSIM offers some advantages over data simulators employed in previous CPUE standardization studies. Previous CPUE standardization studies generally used simplified data simulators that closely resembled the mechanics of the CPUE standardization models (e.g., Lynch et al., 2012; Carruthers et al., 2010). By contrast, LLSIM is based on conditioning of observed catch rates to complex layers of oceanographic data, real-world fleet dynamics and fisheries-dependent variables. Thus, the underlying dynamics of LLSIM are governed by "unobservable", non-linear environmental processes that are far more complex than the subset of information that is communicated to CPUE standardization models, making the simulation-evaluation process with LLSIM less idealized. Furthermore, the common challenges, such as violation of independence between fishing sets, are captured (at least spatially) by LLSIM.

The evaluation component of the simulation-evaluation process conducted in this study was such that the model developer (Arnaud Grüss) did not know any details regarding the LLSIM simulations and the environmental conditions in the system simulated in LLSIM.

2.3. CPUE standardization methods considered in this study

In this study, we considered nine CPUE standardization methods (Table 1), which we describe below. The raw CPUE data from LLSIM included many zeros. In this context, it was appropriate to fit delta GLMs, GAMs and GLMMs (Lo et al., 1992; Stefánsson, 1996; Barry and Welsh, 2002). The delta approach involves modeling the probability of encounter of a fish population assuming a binomial distribution, and the mean CPUE when fish are encountered assuming a lognormal distribution, and then multiplying the results together to obtain an overall standardized CPUE (Lo et al., 1992; Grüss et al., 2014). Future studies could explore other variants of the delta approach (e.g., Thorson, 2017), though we hypothesize that any improvements in statistical efficiency will affect CPUE standardization methods similarly and will not affect relative performance among the nine standardization methods explored in this study. Moreover, for all CPUE standardization methods, no model selection was conducted as all the covariates influencing catchability (henceforth "catchability covariates") were deemed likely to influence CPUE.

2.3.1. The GLM, GLMint and GLMMint methods

The delta GLMs we developed for the GLM method estimated terms for year and area as fixed effects and integrated the fixed effects of catchability covariates and no year-area

interaction term. We fitted both the binomial GLMs and the lognormal GLMs making up these delta GLMs in the R environment, following the equation:

$$g(\eta) = year + season + area + hook + bait + light + hbf \quad (1)$$

where η is either the probability of encounter when given binomial response data, or an estimate of CPUE when given non-zero CPUE data; g represents the link function between η and each covariate (logit in the case of the binomial GLMs, and log in the case of the lognormal GLMs); $hook$ is the type of hook used; $bait$ is the type of bait used; $light$ is the number of light sticks used expressed as a categorical variable; and hbf is the number of hooks between floats expressed as a categorical variable; $season$, $hook$, $bait$, $light$ and hbf are all catchability covariates.

The delta GLMs we developed for the GLMint method were similar to those developed for the GLM method, except that they also included a year-area interaction term ($year * area$) as a fixed effect. We fitted both the binomial GLMs and the lognormal GLMs making up these delta GLMs in the R environment, following the equation:

$$g(\eta) = year + season + area + hook + bait + light + hbf + year * area \quad (2)$$

The delta GLMMs we developed for the GLMMint method were similar to the delta GLMs developed for the GLMint method, except that the $year * area$ term was included in the binomial and lognormal GLMMs as a random rather than as a fixed effect. The binomial and lognormal GLMMs developed for the GLMint method were fitted using the "glmer" function in the "lme4" library for R (Bates et al., 2015).

For the GLM, GLMint and GLMMint methods, following Punt et al. (2000) and Ono et al. (2015), we predicted mean annual probability of fish encounter and mean annual CPUE when fish are encountered with the fitted binomial and lognormal GLMs or GLMMs, using the levels of the season, area, hook, bait, light and hbf factors with the largest sample size

(Table 2). Then, the predicted mean probability of fish encounter was multiplied by the predicted mean annual CPUE when fish are encountered to generate the predicted total CPUE in each year. The standard errors of the predictions of the delta GLMs or GLMMs were computed from the standard errors of the predictions of the binomial and lognormal GLMs or GLMMs using the formula presented in Lo et al. (1992).

It is worth reiterating that employing the GLMint method negates the interest of the year-area interaction term, as a specific area stratum then needs to be chosen to construct an index of relative abundance (Maunder and Punt, 2004; Lynch et al., 2012; Campbell, 2015). Therefore, it would not be relevant to use the GLMint method evaluated here in the real world; we considered the GLMint method here solely to explore the consequences of integrating the year-area interaction effect in a GLM as a fixed vs. as a random effect.

2.3.2. The GLMwt, GLMwt.int, GLMMprwt and GLMprwt.int methods

The GLMwt and GLMwt.int methods consisted of fitting binomial and lognormal GLMs with and without a fixed $year * area$ term following, respectively, Eqs. (1) and (2), and then making a series of calculations rather than solely multiplying the predictions of binomial and lognormal GLMs together (see below). The GLMprwt and GLMprwt.int methods were similar, except that they assigned prior weights to the data based on the year-area stratum to which the data belonged. Following Campbell (2015), when the GLMprwt and GLMprwt.int methods were employed, a weight $weight_{y,a}$ was assigned to an observation for year *y* and area *a* as follows:

$$weight_{y,a} = \frac{Nobs}{Nstrata} . \frac{1}{n_{y,a}} \quad (3)$$

where $n_{y,a}$ is the number of observations for year *y* and area *a*; $Nobs$ is the total number of observations; and $Nstrata$ is the total number of strata considered, with $Nstrata = Ny \times Na$, where Ny is the number of years considered (29 when working with LLSIM data; 26 when working with real observer data) and Na is the number of areas considered (10).

When the GLMwt, GLMwt.int, GLMprwt and GLMprwt.int methods are utilized, the estimation of annual CPUEs takes place in three steps (Campbell, 2004, 2015). First, probabilities of encounter are predicted with fitted binomial GLMs and CPUEs when fish are encountered are predicted with fitted lognormal GLMs for each year, each season and each area, using the levels of the hook, bait, light and hbf factors with the largest sample size (Table 2; Punt et al., 2000; Ono et al., 2015). Second, CPUE for year *y* and season *s* is estimated as follows:

$$CPUE_{y,s} = \sum_{a=1}^{Na} SA_a \, prob_{y,s,a} u_{y,s,a} \tag{4}$$

where $prob_{y,s,a}$ is the probability of encounter in year *y*, season *s* and area *a* predicted by the binomial GLM; $u_{y,s,a}$ is the CPUE when fish are encountered in year *y*, season *s* and area *a* predicted by the lognormal GLM; and SA_a is the surface area (in km²) of area *a*. Third and lastly, annual CPUEs are computed from CPUE estimates for each year and season as follows:

$$CPUE_y = \frac{1}{Ns} \sum_{s=1}^{Ns} CPUE_{y,s} \tag{5}$$

where Ns is the number of seasons (4). We computed the standard errors of these annual CPUEs using the formula developed in Campbell (2015). Note that a geometric mean could be employed in lieu of the arithmetic mean in Eq. (5); the advantage of geometric mean is that it is scale invariant and less sensitive to outliers (Campbell, 2015).

341

342 *2.3.3. The GAM method*

343 Regarding the GAM method, we fitted both the binomial GAMs and the lognormal
344 GAMs making up the delta GAMs using the R package "mgcv" (Wood and Augustin, 2002;
345 Wood, 2006), following the equation:

$$g(\eta) = year + s(X,Y) + season + hook + bait + light + hbf \quad (6)$$

346 where $s(X,Y)$ is product smooth fitted to eastings (X) and northings (Y), which replaces the
347 fixed effect of area and represents spatial autocorrelation at a broad spatial scale (Grüss et al.,
348 2016, 2019).

349 As for the previous models, we predicted annual probability of fish encounter and
350 annual CPUE when fish are encountered for the cells of the spatial grid for the North Atlantic
351 with the fitted binomial and lognormal GAMs, using the levels of the season, hook, bait, light
352 and hbf factors with the largest sample size (Table 2; Punt et al., 2000; Ono et al., 2015; Grüss
353 et al., 2018b, 2018c). We then calculated mean annual probabilities of fish encounter over all
354 cells of the spatial grid for the North Atlantic and mean annual CPUEs when fish are
355 encountered over all cells of the spatial grid. Finally, these two results were multiplied
356 together to predict total CPUEs in each year. We computed the standard errors of mean
357 annual probabilities of fish encounter and mean annual CPUEs when fish are encountered
358 using Marra and Wood (2012)'s method, which accounts for covariance between predictions
359 for the individual cells of the spatial grid. We then employed the formula presented in Lo et
360 al. (1992) to compute the standard errors of delta GAM predictions from the standard errors
361 of mean annual probabilities of fish encounter and mean annual CPUEs when fish are
362 encountered.

363

2.3.4. The VAST method

The VAST method consisted of developing spatio-temporal delta GLMMs implemented using the R package "VAST" (Thorson, 2019), which is publicly available online (https://github.com/James-Thorson/VAST). Below, we describe the estimation of probabilities of encounter and CPUEs when fish are encountered with VAST. Additional details can be found in Appendix A1. One detail to highlight here is that, for computational reasons, 250 knots were defined in VAST via the application of a *k*-means algorithm (Thorson et al., 2015) to the locations of raw (observed) CPUE data. These knots are allocated spatially with a density proportional to sampling intensity, and indices of relative abundance are obtained by summing over the annual standardized CPUEs estimated for each knot. Another detail to highlight is that VAST integrates across the coefficients of the catchability covariates by implementing restricted maximum likelihood (REML) estimation (Grüss et al., 2018a, 2018d).

With VAST, probability of encounter was approximated using a spatio-temporal binomial GLMM with a logit link function and linear predictors, including a Gaussian Markov random field representing spatio-temporal variation in probability of encounter and another Gaussian Markov random field representing spatial variation in probability of encounter. The spatio-temporal binomial GLMM predicts probability of encounter p_i at site $s(i)$ as follows:

$$p_i = logit^{-1}\left(\sum_{y=1}^{Ny}\beta_y^{(p)}YEAR_{i,y} + \sum_{season=1}^{Nseasons}\gamma_{season}^{(p)}SEASON_{i,season} + \sum_{hook=1}^{Nhooks}\delta_{hook}^{(p)}HOOK_{i,hook} + \sum_{bait=1}^{Nbaits}\zeta_{bait}^{(p)}BAIT_{i,bait} + \sum_{light=1}^{Nlights}\eta_{light}^{(p)}LIGHT_{i,light} + \sum_{hbf=1}^{Nhbfs}\theta_{hbf}^{(p)}HBF_{i,hbf} + \varepsilon_{J(i),Y(i)}^{(p)} + \omega_{J(i)}^{(p)}\right) \quad (7)$$

where $YEAR_{i,y}$ is a design matrix where $YEAR_{i,y}$ is one for the year y during which sample i was collected and zero otherwise; $\beta_y^{(p)}$is an intercept that varies among years; $SEASON_{i,season}$ is a design matrix where $SEASON_{i,season}$ is one for the season level associated with sample i and zero otherwise; $\gamma_{season}^{(p)}$ is a season effect on probability of encounter (where $\gamma_{season}^{(p)} = 0$ for the season level with the largest sample size for a population, where this constraint is imposed for identifiability of all year effects $\beta_y^{(p)}$); $Nseasons$ is the number of season levels (4); $HOOK_{i,hook}$ is a design matrix where $HOOK_{i,hook}$ is one for the hook level associated with sample i and zero otherwise; $\delta_{hook}^{(p)}$ is a hook effect on probability of encounter (where $\delta_{hook}^{(p)} = 0$ for the hook level with the largest sample size for a population, where this constraint is imposed for identifiability of all year effects $\beta_y^{(p)}$); $Nhooks$ is the number of hook levels (3); $BAIT_{i,bait}$ is a design matrix where $BAIT_{i,bait}$ is one for the bait level associated with sample i and zero otherwise; $\zeta_{bait}^{(p)}$ is a bait effect on probability of encounter (where $\zeta_{bait}^{(p)} = 0$ for the bait level with the largest sample size for a population, where this constraint is imposed for identifiability of all year effects $\beta_y^{(p)}$); $Nbaits$ is the number of bait levels (4 when working with LSSIM data); $LIGHT_{i,light}$ is a design matrix where $LIGHT_{i,light}$ is one for the light level associated with sample i and zero otherwise; $\eta_{light}^{(p)}$ is a light effect on probability of encounter (where $\eta_{light}^{(p)} = 0$ for the light level with the largest sample size for a population, where this constraint is imposed for identifiability of all year effects $\beta_y^{(p)}$); $Nlights$ is the number of light levels (4); $HBF_{i,hbf}$ is a design matrix where $HBF_{i,hbf}$ is one for the hbf level associated with sample i and zero otherwise; $\theta_{hbf}^{(p)}$ is an hbf effect on probability of encounter (where $\theta_{hbf}^{(p)} = 0$ for the hbf level with the largest sample size for a population, where this constraint is imposed for identifiability of all year effects $\beta_y^{(p)}$); $Nhbfs$ is the number of hbf levels (5 when working

with LSSIM data; 4 when working with real observer data); $\varepsilon^{(p)}_{J(i),Y(i)}$ is the spatially correlated variability in probability of encounter at the knot *J(i)*, which is the nearest knot to sample *i*, in year $Y(i)$ in which sample *i* was collected; and $\omega^{(p)}_{J(i)}$ is the spatially correlated variability in probability of encounter at the knot *J(i)* that is persistent among years. Both $\varepsilon^{(p)}_{J(i),Y(i)}$ and $\omega^{(p)}_{J(i)}$ are random effects.

Similarly, with VAST, positive catch rate was approximated using a spatio-temporal lognormal GLMM with a log link function and linear predictors, including a Gaussian Markov random field representing spatio-temporal variation in positive catch rate and another Gaussian Markov random field representing spatial variation in positive catch rate. The spatio-temporal lognormal GLMM predicts positive catch rate λ_i at site *s(i)* as follows:

$$\lambda_i = exp\left(\sum_{y=1}^{Ny}\beta_y^{(\lambda)}YEAR_{i,y} + \sum_{season=1}^{Nseasons}\gamma_{season}^{(\lambda)}SEASON_{i,season} + \sum_{hook=1}^{Nhooks}\delta_{hook}^{(\lambda)}HOOK_{i,hook} + \sum_{bait=1}^{Nbaits}\zeta_{bait}^{(\lambda)}BAIT_{i,bait} + \sum_{light=1}^{Nlights}\eta_{light}^{(\lambda)}LIGHT_{i,light} + \sum_{hbf=1}^{Nhbfs}\theta_{hbf}^{(\lambda)}HBF_{i,hbf} + \varepsilon^{(\lambda)}_{J(i),Y(i)} + \omega^{(\lambda)}_{J(i)}\right) \quad (8)$$

where the parameters on the right side of Eq. (8) have the same meaning and characteristics as the parameters on the right side of Eq. (7), except that they apply to log-catch rate.

To make predictions with fitted spatio-temporal GLMMs, we assumed that the Gaussian Markov random field in each cell of the spatial grid for the North Atlantic was equal to the value of the random field at the closest knot. Consequently, the surface area SA_j associated with knot *j* was calculated as the number of cells of the spatial grid for the North Atlantic associated with knot *j* times the surface areas of these cells. It was then possible to calculate total CPUE in year *y* across our entire study region as follows:

$$\widehat{CPUE}_y = \sum_{j=1}^{n_j} SA_j \; logit^{-1}\left(\hat{\beta}_y^{(p)} YEAR_{j,y} + \hat{\varepsilon}_{j,y}^{(p)} + \hat{\omega}_j^{(p)}\right). \\ exp\left(\hat{\beta}_y^{(\lambda)} YEAR_{j,y} + \hat{\varepsilon}_{j,y}^{(\lambda)} + \hat{\omega}_j^{(\lambda)}\right) \quad (9)$$

where $\hat{\beta}_y^{(p)}$ and $\hat{\beta}_y^{(\lambda)}$ are fixed effects of year estimated through maximum likelihood estimation; and $\hat{\varepsilon}_{j,y}^{(p)}$, $\hat{\omega}_j^{(p)}$, $\hat{\varepsilon}_{j,y}^{(\lambda)}$ and $\hat{\omega}_j^{(\lambda)}$ are random effects set to the value that maximizes the joint likelihood conditional on the estimated value of fixed effects of year (Thorson et al., 2015). The standard errors of the annual CPUEs predicted by the spatio-temporal GLMMs were computed using a generalization of the delta method (Thorson et al., 2015; Thorson and Barnett, 2017).

2.4. Scenarios considered in this study

Three scenarios were considered for each of the three virtual blue marlin populations: (1) the "ALL" scenario, where all LLSIM data (i.e., the 294,305 simulated longline sets) were employed to standardize CPUE data; (2) the "10%" scenario, where 10% of the fishing trips simulated by LLSIM were randomly selected, and (3) the "10%BIAS" scenario, which consisted of selecting 10% of the fishing trips simulated by LLSIM such that trips with lower than average catch rate were selected in a higher proportion, resulting in a biased sample that might reflect the process of an observer bias where fishing trips with observers operate differently than ones without observers to avoid bycatch species (e.g., sea turtles). In the real world, the percentage of trips of the U.S. pelagic longline fishery sampled by observers each year varies from one year to the next, but is around 10% on average (Beerkircher et al., 2002).

To build the 10% and 10%BIAS scenarios, we needed to work with fishing trips. However, LLSIM provided us with simulated longline sets. Therefore, we needed to assign each of the LLSIM longline sets to fishing trips, such that each fishing trip would have

445 longline sets around the same time and location. To generate fishing trips with these
446 characteristics, we assigned fishing sets that were in the same year, month and NMFS area to
447 the same fishing trip. This yielded a total of 18,870 fishing trips, with a median of 6 sets per
448 trip and a maximum of 329. Since, in the real world, the number of longline sets per vessel
449 month in the U.S. pelagic longline fishery has a median of 8 (range 1-40; Beerkircher et al.,
450 2002), we broke up the fishing trips that had more than 40 longline sets into trips with 8 sets
451 each, counting from the first longline set in the dataset, so as to maintain any structure in the
452 data that might be incorporated in longline set order. The resulting dataset had a total of
453 37,327 fishing trips with a median of 8 longline sets each (range 1-40).

454 As mentioned above, the three virtual blue marlin populations had the exact same
455 characteristics except that one maintained a constant abundance over time, one was generally
456 declining and the third one was generally increasing. Therefore, with respect to the 10%
457 scenario, it was possible to generate subsamples for the three virtual blue marlin populations
458 together. Since the generation of subsamples for the 10% scenario is a stochastic process, we
459 produced five subsamples for the 10% scenario so as to run five replicates of the scenario.

460 To obtain subsamples to explore the 10%BIAS scenarios, we randomly drew 10% of
461 fishing trips with a probability of sampling a particular trip (*prob*) generated from the
462 equation:

$$logit(prob) = a + b \times C \quad (10)$$

463 where C is here the total catch of blue marlin in the fishing trip under consideration. The
464 parameters a and b were set so that the probability of sampling a given fishing trip was 0.1 at
465 the mean catch level and decreased to 0.01 at the maximum catch level. This gave an overall
466 sampling effort of around 10% of fishing trips, with a significantly lower probability of
467 sampling fishing trips that catch more blue marlins. Since the catches varied between the three

468 virtual blue marlin populations, we generated different samples for each population.
469 Furthermore, since the generation of subsamples for the 10%BIAS scenario is a stochastic
470 process, for each virtual blue marlin population, we produced five subsamples for the
471 10%BIAS scenario so as to run five replicates of the scenario.

472 Following the best practices provided in Campbell (2015), for the ALL scenario and
473 all replicates of the 10% and 10%BIAS scenarios, we constructed a "Walters' table" from the
474 raw CPUE data with a row for each year and a column for each area (Table 3 and Table A2).
475 The Walters' table for the ALL scenario showed that there were observations in all year-area
476 strata (Table 3). By contrast, there were missing observations in many year-area strata for all
477 replicates of the 10% and 10%BIAS scenarios (Table A2). Therefore, under the 10% and
478 10%BIAS scenarios, it was necessary to impute CPUE values in unobserved year-area strata
479 when working with the GLMwt.int and GLMprwt.int methods (Walters, 2003; Carruthers et
480 al., 2010). There is no standard method for imputing CPUE values in unobserved year-area
481 strata (Walters, 2003; Carruthers et al., 2010, 2011; Campbell, 2015). In this study, we used
482 one of the imputation methods employed in Campbell (2015). This method consisted of
483 imputing CPUE values in unobserved year-area strata by directly using the predictions made
484 for those year-area strata by simpler GLMs not integrating a year-area interaction term.

485 To illustrate the usefulness of spatio-temporal models beyond CPUE standardization,
486 we estimated the eastward and northward centers of gravity (COGs) of the virtual blue marlin
487 populations and their effective area occupied with VAST when considering the ALL scenario
488 (which uses all of the LLSIM data). The computation of COGs and effective areas occupied is
489 described in Appendix A1.

490

491 ***2.5. Evaluation and comparison of the CPUE standardization methods***

492 The first step in evaluating and comparing CPUE standardization methods was to plot
493 the normalized estimated annual trend in CPUE for each method. Normalized CPUEs
494 estimated for each standardization method were then compared amongst one another, as well
495 as to the normalized virtual blue marlin population abundance (true abundance) (Figs. 2a-c)
496 and to the normalized CPUEs calculated directly from the LLSIM data (nominal CPUEs) for
497 each virtual blue marlin population. Normalization was carried out in all cases by dividing
498 mean annual CPUEs or abundance by their mean value over the period from 1987-2015.
499 Then, we assessed the performance of the CPUE standardization methods for each virtual
500 blue marlin population and scenario based on three metrics: (1) a bias metric described below;
501 (2) mean absolute error (MAE), which quantifies error in the estimated CPUEs; and (3) a
502 confidence coverage metric described below.

503 The bias metric we considered was the coefficient d of the following linear model
504 (Thorson et al., 2015):

$$\widehat{CPUE}_y = c + d \times I_y \qquad (11)$$
$$\varepsilon_y \sim Normal(0, \sigma_\varepsilon^2)$$

505 where c is an intercept; $\widehat{CPUE}_y$ is the normalized estimated CPUE in year *y*; I_y is the
506 normalized true abundance in year *y*; ε_y is the "estimation error" in the normalized estimated
507 CPUE; and σ_ε^2 is the variance of ε. A d of 1 is indicative that changes in true abundance are
508 reflected accurately by the estimated CPUE, while a d greater than 1 (lower than 1) indicates
509 that $\widehat{CPUE}_y$ underestimates (overestimates) changes in true abundance (Wilberg et al., 2010;
510 Thorson et al., 2015). It was not possible to calculate bias for Population 1, whose true
511 abundance is constant over time (Fig. 2a).

512 MAE was calculated for each virtual blue marlin population and scenario as (Willmott
513 and Matsuura, 2005; Stow et al., 2009):

$$MAE = \sum_{y=1}^{Ny} \frac{\left|\widehat{CPUE}_y - I_y\right|}{Ny} \quad (12)$$

where Ny is the number of years considered (29). The higher the MAE, the greater the error in the estimated CPUEs (Stow et al., 2009).

Finally, for each virtual blue marlin population, scenario and standardization method, coverage was calculated as the percentage of years over the period 1987-2015 that the 50% confidence interval of the normalized estimated CPUE index contained the normalized true abundance (Agresti and Coull, 1998; Newcombe, 1998; Brown et al., 2001). We chose a nominal probability of 50% rather than 90 or 95% confidence intervals to provide greater contrast in performance. Well-performing confidence intervals are ones where the nominal (predetermined) probability equals the actual proportion of replicates where the confidence interval contains the true value. In our case, coverage values >50% indicate that the confidence intervals are too wide and coverage values <50% indicate that the confidence intervals are too narrow (Bolker, 2008; Johnson et al., 2016).

2.6. Application of the CPUE standardization methods to real observer data

All CPUE standardization methods with the exception of the GLMint method were also applied to real observer data collected by the NMFS Pelagic Observer Program (Beerkircher et al., 2002) over the period 1992-2017. We did not consider the GLMint method, since, as explained earlier, this method is not relevant for standardizing CPUE data in the real world (Maunder and Punt, 2004; Lynch et al., 2012; Campbell, 2015). As was the case for the analysis conducted with LLSIM data, we worked with CPUE per set expressed as the number of blue marlins caught per number of hooks set. The catchability covariates considered for the application to real observer data were identical to those considered when

536 working with LLSIM data, except bait, which was excluded as a factor as more than 99% of
537 the observations were with dead bait (Table 4). The "Walters' table" we constructed from the
538 raw NMFS Pelagic Observer Program CPUE data showed that were missing observations in
539 40 year-area strata (i.e., in around 15.4% of the year-area strata; Table 5). Therefore, when
540 working with the GLMwt.int and GLMprwt.int methods, we used one of the imputation
541 methods employed in Campbell (2015), which consisted of imputing CPUE values in the
542 unobserved year-area strata by directly using the predictions made for those year-area strata
543 by simpler GLMs not integrating a year-area interaction term. When working with VAST, we
544 also estimated the eastward and northward COGs and the effective area occupied of the blue
545 marlin population, following the methodology described in Appendix A1.

546

547 **3. Results**

548 ***3.1. COGs and effective area occupied of the virtual blue marlin populations***

549 The eastward and northward COGs and the effective area occupied of virtual blue
550 marlin populations 1, 2, and 3 were estimated under the ALL scenario via the spatio-temporal
551 GLMMs computed using VAST (Figs. 2d-l). This analysis suggested that Population 1, which
552 maintained a constant abundance over the period 1987-2015 (Fig. 2a), also had a constant
553 effective area occupied between 1987 and 2015 (Fig. 2f) and that, between 1996 and 2015,
554 the COG of Population 1 moved northward (Fig. 2e). The spatio-temporal GLMMs indicated
555 that the COG of Population 2, whose abundance generally declined over the period 1987-
556 2015 (Fig. 2b), moved northward between 1996 and 2015 but also westward in 2006 (Figs.
557 2g-h). Moreover, the predicted effective area occupied of Population 2 decreased between
558 2000 and 2006 and then stabilized (Fig. 2i). Finally, the spatio-temporal GLMMs indicated
559 that the COG of Population 3, whose abundance generally increased over the period 1987-

2015 (Fig. 2c), moved northward between 1996 and 2015 (Fig. 2k). The predicted effective area occupied of Population 3 increased slightly between 1987 and 1995 and was stable afterwards (Fig. 2l).

3.2. Analyses conducted with LLSIM data

We considered three virtual blue marlin populations, three sampling scenarios (ALL, 10%, and 10%BIAS) and nine standardization methods (GLM, GLMint, GLMMint, GLMwt, GLMwt.int, GLMprwt, GLMprwt.int, GAM, and VAST). Moreover, for each virtual blue marlin population, we ran five replicates of the 10% scenario and five replicates of the 10%BIAS scenario. Therefore, we estimated a total of 3 * (1 + 5 + 5) * 9 = 297 indices of relative abundance. Under the 10% and 10%BIAS scenarios, there were instances where inclusion of the fixed year-area interaction term led to convergence issues with the binomial GLMs; convergence issues arise when any year-area stratum has 0% or 100% encounter rates, as noted in previous studies (Lynch et al., 2012; Campbell, 2015). When binomial GLMs integrating a fixed year-area interaction effect did not converge, we combined the predictions of a binomial GLM without a year-area interaction effect with the predictions of a lognormal model integrating a fixed year-area interaction effect.

The relative sample size of the levels of the area factor and catchability covariates varied largely over the period 1987-2015 (Fig. 3 and Fig. A3), justifying the standardization of the LLSIM CPUE data. Notably: (1) the “unknown” hook type was dominant until 2004, after what virtually all the hooks used were circle hooks; and (2) the “unknown” bait type was employed in 1987 and 1988, while the “dead” bait type was dominant between 1989 and 2015 (Fig. 3 and Fig. A3).

583 Overall, the indices of relative abundance estimated by all CPUE standardization
584 methods matched true abundances well (Fig. A4). However, under the 10% and 10%BIAS
585 scenarios, there were several instances where the standardization methods relying on GLMs
586 integrating a fixed year-interaction effect (i.e., the GLMint, GLMwt.int and GLMprwt.int
587 methods) resulted in poorly estimated indices of relative abundance (Figs. 4-5 and Figs. A4).
588 We examine some of these instances in detail below.

589 In general, CPUE standardization methods had relatively little bias (Fig. 6). An
590 exception to this general pattern was Population 3 under the 10%BIAS scenario, for which
591 two CPUE standardization methods (GLMprwt and GLMMint) noticeably underestimated the
592 true changes in abundance. Under all scenarios and for all populations, generally, the
593 GLMMint method had the strongest negative bias (representing hyperstability in the estimated
594 index of abundance), while the GLMint method had the strongest positive bias. Under the
595 ALL and 10% scenarios, the GLMprwt.int, VAST and GLMprwt methods had the lowest
596 biases. Under the 10%BIAS scenario, the GLM and GAM methods had the lowest biases, the
597 GLMprwt.int method had a relatively low negative bias similar to that of the VAST method
598 for all populations combined and Population 2, and the GLMwt method had a relatively low
599 negative bias similar to that of the VAST method for Population 3 (Fig. 6).

600 MAE showed great variation among CPUE standardization methods (Fig. 7). Under
601 all scenarios and for all populations, the GAM method was usually the CPUE standardization
602 methods with the lowest MAE, followed closely by the VAST, GLMwt and GLMMint
603 methods, in this order. Under the ALL scenario, the GLMint method was the method with the
604 largest MAE, usually followed by the GLMwt.int and GLMprwt.int methods. Under the 10%
605 scenario, the GLMprwt and GLMprwt.int methods, which both assigned prior weights to data
606 based on the year-area stratum to which the data belonged, had, in general, the largest MAEs,
607 followed by the GLMint method. An exception to this general pattern was Population 2, for

608 which the GLMint method had the largest MAE under the 10% scenario. Under the 10%BIAS
609 scenarios, the GLMint, GLMprwt and GLMprwt.int methods had the largest MAEs (Fig. 7).

610 Coverage also showed great variation among CPUE standardization methods,
611 particularly under the ALL scenario (Fig. 8). Under all scenarios and for all populations, the
612 VAST method had the coverage the closest to 50%, often followed by the GLMMint method.
613 The good coverage of the GLMMint method was in great part due to the fact that its
614 predictions were associated with large standard errors (Fig. A5). Under the ALL scenario, the
615 GLMwt and GLMprwt methods had the coverages the farthest to 50% (Fig. 8), due to the fact
616 that they predicted standard errors that were anomalously low (Fig. A5). Under the 10% and
617 10%BIAS scenarios, the GLMprwt and GLMprwt.int methods had the lowest coverages (Fig.
618 8). Moreover, under the 10% and 10%BIAS scenarios, the GAM method often had coverages
619 that were much greater than 50%, indicating that this method often have confidence intervals
620 that are too wide (Fig. 8). The results for the GAM method were due to the fact that its
621 predictions were associated with large standard errors (Fig. A5).

622 To understand why, in some instances, the standardization methods relying on GLMs
623 incorporating a fixed year-interaction effect (i.e., the GLMint, GLMwt.int, and GLMprwt.int
624 methods) resulted in poorly estimated indices of relative abundance, we examined: (1) the
625 results obtained for Population 1 under the 10%BIAS scenario with Replicate 2 (Fig. 4 and
626 Figs. A6 and A7); and (2) the results obtained for Population 3 under the 10%BIAS scenario
627 with Replicate 1 (Fig. 5 and Figs. A8 and A9). Note that, in addition to the poorly estimated
628 indices of relative abundance obtained with the GLMint, GLMwt.int and GLMprwt.int
629 methods, Figs. 4 and 5 illustrate the low coverage of the GLMprwt and GLMprwt.int
630 methods. In the two cases examined here, the fixed year-area interaction term did not lead to
631 convergence issues with the binomial GLMs. For the two cases, we: (1) plotted the year-area
632 interaction coefficients of the binomial and lognormal GLMs developed for the GLMint

633 method (Figs. A6 and A8); and (2) produced maps showing the spatial distribution of
634 observer data for each year of the period 1987-2015 (Figs. A7 and A9).

635 In the first case examined (Population 1, 10%BIAS scenario, Replicate 2), while the
636 true abundance of the virtual blue marlin population was constant over the period 2000-2015,
637 the GLMint method predicted the index of relative abundance to increase over that period
638 (Fig. 4). This result is due to the fact that: (1) predictions were made with the GLMs
639 developed for the GLMint method using the NEC factor level (binomial model) and the FEC
640 factor level (lognormal model) (Table 2); and (2) the year-area interaction terms estimated for
641 the FEC and NEC areas for the GLMs developed for the GLMint method tended to increase
642 over the period 2000-2015 (Fig. A6). The GLMwt.int and GLMprwt.int methods, which
643 weighted year-area interactions by the surface area of each NMFS area, downweighted the
644 influence of the FEC and NEC areas and did not predict an increase in the index of relative
645 abundance over the period 2000-2015; yet, the indices of relative abundance estimated by the
646 GLMwt.int and GLMprwt.int methods fitted the true data more poorly than those estimated
647 by some of the other CPUE standardization methods such as the VAST, GAM and GLMwt
648 methods (Fig. 4). Almost all the year-area coefficients of the binomial and lognormal models
649 fitted by the GLMwt.int method were non-significant at the 5% level. To further gauge the
650 significance of the year-area interaction terms, for both the binomial and lognormal models
651 fitted by the GLMwt.int method, we performed stepwise model selection by the Akaike
652 Information Criterion (AIC), using the function "stepAIC" from the R package "MASS"
653 (Venables and Ripley, 2002). The stepwise model selection procedure resulted in the year-
654 area interaction term being dropped from both the binomial and lognormal models.

655 In the second case examined (Population 3, 10%BIAS scenario, Replicate 1), the
656 GLMint method predicted erroneous spikes over the most recent years (e.g., in 2008; Fig. 5).
657 These erroneous spikes were due to: (1) the fact that predictions were made with the GLMs

developed for the GLMint method using the NEC factor level (binomial model) and the FEC factor level (lognormal model) (Table 2); and (2) the year-area interaction coefficients estimated for the FEC and NEC areas for the GLMs developed for the GLMint method (e.g., which both peak in 2008; Fig. A8). Moreover, in the second case study examined, the GLMwt.int and GLMprwt.int methods estimated indices of relative abundance that fitted true abundances more poorly than those estimated by the GLMint method; the indices of relative abundance estimated by the GLMwt.int and GLMprwt.int methods exhibited additional erroneous spikes (e.g., in 2002 and 2006; Fig. 5). This result stems from the fact that the NED and NCA areas, which are located, respectively, in the northeast and the southeast of our study region, are associated with very large surface areas (Fig. A9) and high year-area interaction coefficients in some years (e.g., 2002 and 2006; Fig. A8). However, over the period 1996-2015, Population 3 was predicted to move northward (Fig. 2k). Consequently, the GLMwt.int and GLMprwt.int methods, which weight year-area interactions by the surface area of each NMFS area and give more weights to the NED and NCA areas than the GLMint method, overestimated relative abundance in some years (e.g., in 2002 and 2006; Fig. 5).

3.3. Application of the CPUE standardization methods to real observer data

The CPUE standardization methods applied to real observer data tended to predict similar patterns, particularly a decline in the blue marlin population over the period 1998-2004 followed by a relative stabilization of the population (Figs. 9 and 10). However, while the GLM, GLMMint GLMwt and GAM methods predicted a slight increase in blue marlin abundance in 1997-1998, the GLMwt.int, GLMprwt, GLMprwt.int and VAST methods predicted a marked peak in abundance for the same time period (Fig. 10). Moreover, the indices of relative abundance estimated with the GLMwt.int and GLMprwt.int methods were

682 more variable than those estimated with the other CPUE standardization methods, and they
683 exhibited lots of peaks and troughs (Figs. 9 and 10). VAST predicted that blue marlin COG
684 moved both eastward and southward in 1996-1997 and then moved slightly westward
685 between 1998 and 2014 (Figs. 11a-b). VAST also predicted that the effective area occupied
686 by blue marlin remained relatively constant over the period 1992-2017 (Fig. 11c).

687 To understand the estimated peak in relative abundance predicted for 1996-1997, we
688 generated maps showing: (1) the spatial distribution of observer data for each year of the
689 period 1992-2017 (which cannot be provided here or in the Supplementary data due to the
690 confidentiality of the observer data); and (2) the standard errors associated with the indices of
691 relative abundance estimated by the VAST method for each year of the period 1992-2017
692 (Fig. A10). The first maps suggested that the predicted peak in relative abundance for 1996-
693 1997 may be in part due to a few fishing trips with high catch rates made off the northeast
694 coast of Brazil, in an area where sets were not observed by the NMFS Pelagic Observer
695 Program outside of 1996 and 1997. The second maps showed that: (1) the locations of the
696 observer data collected in the area off the northeast coast of Brazil were used by VAST to
697 define a relatively large knot in the southeasternmost corner of our study region; but that (2)
698 despite the low number of samples and large surface area of that knot over which these
699 samples were extrapolated, the standard errors associated with the indices of relative
700 abundance predicted for that knot and adjacent knots were low, in 1996-1997, but also pre-
701 1996 and post-1997 (Fig. A10).

702

703 **4. Discussion**

704 In general, fisheries-independent surveys use sampling designs which on average
705 provide unbiased indices of relative abundance (Thompson, 2002). Unfortunately, because

706 fisheries-independent surveys are costly and time-consuming, they are generally conducted
707 during specific months and rarely entirely cover large marine regions such as the North
708 Atlantic (Lynch et al., 2012; Bourdaud et al., 2017). Consequently, many exploited fish
709 populations such as Atlantic blue marlin are not monitored by fisheries-independent surveys
710 (Lynch et al., 2012; Walter et al., 2014a). Instead, for these fish populations, indices of
711 relative abundance are derived from fisheries-dependent CPUE data, which are collected with
712 the assistance of fishers who adapt their fishing grounds and behavior based on prevailing
713 environmental and socio-economic conditions and, perhaps, the presence of observers
714 onboard (Walters, 2003; Maunder and Punt, 2004; Marchal et al., 2006; Walter et al., 2014a).
715 Under these circumstances, it is critical to assess the performance of methods for
716 standardizing fisheries-dependent CPUE data. In the present study, we evaluated and
717 compared nine CPUE standardization methods, which offered contrasting treatments of
718 spatio-temporal variation: (1) non-spatial methods that accounted or not for the interaction
719 between the year and area effects (GLM, GLMint, and GLMMint); (2) methods that
720 accounted or not for the interaction between the year and area effects, but also weighted or not
721 model predictions for individual areas by the surface area of each these areas and/or assigned
722 prior weights to raw CPUE data based on the year-area stratum to which the CPUE data
723 belonged (GLMwt, GLMwt.int, GLMprwt, and GLMprwt.int); (3) a method that accounted
724 for spatial autocorrelation at a broad spatial scale (GAM); and (4) a method that accounted for
725 spatial and spatio-temporal autocorrelation at a fine spatial scale (VAST).

726 Despite the substantial degradation of the simulated datasets by subsetting 10%
727 randomly and then 10% nonrandomly, across all of the virtual blue marlin populations, the
728 great majority of the CPUE standardization methods considered in this study managed to
729 extract a relatively unbiased trend in relative abundance. While we do not have unequivocal
730 evidence of observer effect bias occurring in the U.S. pelagic longline fishery, the commonly

731 employed GLM and GLMMint methods, the GAM method and the VAST spatio-temporal
732 method seem to be fairly robust to this potential problem. In this study, there were also virtual
733 populations for which other CPUE standardization methods (the GLMwt and GLMprwt.int
734 methods) had a relatively low negative bias similar to that of the VAST method when the
735 simulated datasets were nonrandomly subsampled to mimic observer bias (Fig. 6). We caveat
736 these ideas with the observation that simulated data rarely perform as poorly as true
737 observations, as it is difficult to mimic the full data generating process. Even 10% observer
738 coverage (as was assumed in this study) may not be possible in many fisheries (National
739 Marine Fisheries Service, 2016), and it is quite possible that the bias between what is
740 observed and what is caught in an overall fishery may change over time.

741 While the different CPUE standardization methods generally provided relatively
742 unbiased trends in relative abundance, with exceptions noted below, we found that the VAST
743 spatio-temporal method generally had one of the lowest biases, one of the lowest MAEs and
744 coverage closest to 50%. The strong performance in simulations of the VAST method argues
745 for greater consideration of spatio-temporal methods in standardization of fisheries-dependent
746 CPUE data. Additionally, spatio-temporal methods are particularly suited for working with
747 fisheries-dependent CPUE data, because they: (1) diminish the influence of repeated fishing
748 operations in sites, thus decreasing the influence of selection bias by fishers; and (2) allow for
749 imputation or extrapolation where CPUE is unknown (Walter et al., 2014a, 2014b). Moreover,
750 VAST is useful not only for standardizing CPUEs and can also be used, among other things,
751 for estimating COGs and effective areas occupied and conducting habitat and climate-
752 vulnerability assessments (see Thorson (2019) for a review). The spatio-temporal modeling
753 platform VAST has benefited from numerous recent developments, including a GitHub
754 repository enabling issue tracking (https://github.com/James-Thorson/VAST) and well-
755 documented example code accompanied by a detailed user guide (which can both be accessed

in GitHub). Also, VAST now has a fairly large and dynamic user community with numerous applications to fisheries-independent datasets, and the present study represents one of the very first applications of VAST to fisheries-dependent data (Thorson, 2019). However, when working with large datasets, VAST simulations can take a long computation time. For instance, it took us around four hours to run each of the VAST simulations under the ALL scenario with a laptop with a 2.6 GHz Intel Core i5-6440HQ processor, using single-thread.

Our results suggest that good alternatives to the VAST method are the GLMMint method, i.e., the variant of the basic GLM method incorporating a random year-area interaction effect, and the GAM method. The GLMMint method had one of the lowest MAEs and one of the best coverages, yet this method also had the strongest negative bias. The GLMMint method is practical in that it obviates the need for imputing CPUE values in unobserved year-area strata when working with unbalanced datasets (Campbell, 2015). The GLMMint method also performs reasonably well and is flexible in terms of fixed and random effects structure; for example, it would probably be feasible to extend the random effect term so that the season effect is nested within area and year. However, the GLMMint method should ideally be utilized only if year-area interactions can be fully explained as random effects (e.g., do not show a significant trend; Cooke, 1997; Maunder and Punt, 2004; Campbell, 2015). The GAM method, which accounts for spatial autocorrelation at a broad spatial scale, may also be a good alternative to the VAST method, because it had the lowest MAE among the nine CPUE standardization methods we tested, as well as the lowest bias under the observer bias scenario. On the other hand, we also found that the GAM method had confidence intervals that were often too wide. In this study, for computational reasons, we did not consider the GAMint method, which also accounts for spatial autocorrelation at a broad spatial scale by integrating a $s(X, Y, by = year)$ term (Wood, 2006). Had we used the

780 GAMint method in this study, we suspect that the GAMint method would have had a lower
781 MAE than the GAM method, at the expense of exceedingly wide confidence intervals.

782 It is important to note that the estimated coverages of the GLMMint and GAM
783 methods are in large part due to the very large standard errors associated with their predictions
784 (Fig. A5). As the CPUE standardization methods considered in this study rely on different
785 procedures for computing standard errors from two independent models (a binomial and a
786 lognormal models), some of the calculated standard errors may not be accurate. Therefore, to
787 some extent, the utility of the coverage metric is dependent on the relative accuracy of the
788 standard errors calculated by each CPUE standardization method. Thus, everything else equal,
789 one may be more confident in using a CPUE standardization method with one of the lowest
790 MAEs and one of the lowest biases (e.g., the GAM method) than a method with one of the
791 lowest MAEs and one of the best coverages (e.g., the GLMMint method).

792 In most cases, the year-area interaction effects in the simulated datasets were not very
793 strong, such that the CPUE standardization methods that either did not estimate them (GLM,
794 GLMwt, and GLMprwt) or estimated them as random effects (GLMMint) performed better
795 than the standardization methods that estimated them as fixed effects and did (GLMwt.int,
796 and GLMprwt.int) or did not (GLMint) use them in predictions. The GAM and the VAST
797 methods model the spatio-temporal effects as fixed and random, respectively, and uses them
798 in the predictions. Hence, the lack of substantial performance differences between including
799 or not including year-area interactions can likely be attributed to the simulated datasets having
800 year-area interaction effects that are not very strong. As no year-area interactions were
801 imposed on the simulated data and would only have been emergent properties of the
802 abundance trends, oceanography and habitat preferences of blue marlin, it is likely that any
803 induced year-area interactions were nor very strong or directional. In this study, our main goal
804 was to compare the performance of CPUE standardization methods integrating or not year-

805 area interaction terms. For this reason, we did not conduct any model selection procedure
806 (besides for understanding the results of one case study). Future studies interested in
807 estimating indices of relative abundance based on the most parsimonious models should
808 perform stepwise model selection by AIC (Venables and Ripley, 2002). This would allow
809 dropping the year-area interaction term from the binomial and/or lognormal models if this
810 interaction term is non-significant (along with non-significant catchability covariates), thereby
811 improving the predictions of the CPUE standardization process.

812 The greatest degradation in performance was with CPUE standardization methods
813 relying on GLMs incorporating a fixed year-area interaction effect (i.e., the GLMint,
814 GLMwt.int, and GLMprwt.int methods), which often resulted in poorly estimated indices of
815 relative abundance, the largest MAEs and the lowest coverages (though not necessarily the
816 largest biases). Similar results were obtained by Thorson and Ward (2013); using delta
817 GLMMs, the authors found that a random year-area interaction often had better performance
818 than a fixed year-area interaction when analyzing sparse fisheries-independent survey data.
819 The literature generally recommends to include year-area interactions as random effects (e.g.,
820 Lynch et al., 2012) where the effects are often constrained by distributional assumptions such
821 as to be normally distributed with a mean of zero. The main issue with the GLMint method is
822 that it gives too much weight to areas whose year-area coefficients hit bounds, are highly
823 erratic or have standard errors indicative of very poor estimation (Figs. A6 and A8). While the
824 spatial weighting employed in the GLMwt.int and GLMprwt.int methods could potentially
825 improve estimation by differentially weighting each year-area interaction coefficient, there is
826 no guarantee that a poorly estimated coefficient will get a small weight. Quite the opposite
827 happened in this study in some cases, where certain large spatial areas had very sparse
828 sampling.

829 Another notable result of the present study was the poor performance of the methods
830 assigning prior weights to data based on the year-area stratum to which the data belong (i.e.,
831 the GLMprwt and GLMprwt.int methods) under the 10% and 10%BIAS scenarios. When
832 dealing with subsamples of the LLSIM data that mimic sampling by observers, the GLMprwt
833 and GLMprwt.int methods often resulted in poorly estimated indices of relative abundance,
834 and they had among the largest MAEs and among the lowest coverages. (Yet, the
835 GLMprwt.int and GLMprwt methods had among the lowest biases; Fig. 6). This result was
836 relatively surprising, given that one would *a priori* expect that assigning prior weights to data
837 would compensate for a very unbalanced dataset by altering the relative influence of each data
838 point (Campbell, 2015). However, we observed virtually no differences between the indices
839 of relative abundance produced by the methods assigning prior weights to data and those not
840 assigning prior weights to data (Figs. 4-5). Using simulated CPUE data for Pacific broadbill
841 swordfish (*Xiphias gladius*) and the Australian pelagic longline fishery, Campbell (2015) also
842 found little differences between the predictions of the methods assigning vs. not assigning
843 prior weights to data. Furthermore, the author observed that assigning prior weights to data
844 resulted in slightly more biased predictions. Campbell (2015) discussed that the results he
845 obtained with the methods assigning or not assigning prior weights to data were likely due to
846 the fact that definition of areas in his study region appropriately stratified CPUE spatial
847 distribution.

848 Thus, neither of the instances of poor performance reported in this study reflect upon
849 the theory or merits of the GLMwt and GLMwt.int, GLMprwt and GLMprwt.int methods, but
850 rather relate to the nature and representativeness of the data relative to the NMFS areas to
851 which they are assigned. While the NMFS areas (Fig. 1) were chosen based on expert opinion
852 and generally reflect homogenous fishing regions, they are of very different sizes and have
853 very different sample coverage per unit area. This leads to correspondingly erratic estimates

of year-area interaction coefficients which may not be representative of the NMFS area to which they are assigned and, when weighted by the surface areas of the NMFS areas, can compound errors. For many fisheries, the area stratification chosen is not based on the biological characteristics of the fishery or the species of interest (such as homogeneity of fish density), but for other management-related reasons. Carruthers et al. (2011) found that GLMs with fixed year-area interaction terms performed better than GLMs without year-area interaction terms, which may have been because the authors employed a regular grid of cells to define areas, where each individual cell had a similar surface area such that no cells could dominate the predicted index of relative abundance. The fact that we relied on an irregular grid of cells to define areas where some cells had an extremely large surface area, combined with the fact that the LLSIM datasets were unbalanced spatially (Figs. A7 and A9), likely degraded the performance of the GLMwt.int and GLMprwt.int methods. Naturally, this raises the question: should we have developed better spatial stratification either by adopting a regular grid or by applying one of several algorithms that search for optimal partitioning to create homogenous spatial regions and minimize the strength of year-area interactions (Ichinokawa and Brodziak, 2010; Ono et al., 2015)?

While a better spatial partitioning might have improved the performance of the GLMwt.int and GLMprwt.int methods, our results support using spatio-temporal modeling to obviate the need to specify *a priori* spatial partitioning entirely. A regular grid would likely exacerbate issues of missing data and would not achieve homogenous stratification or minimization of year-area interactions, whereas an optimal partitioning *sensu* Ichinokawa and Brodziak (2010) and Ono et al. (2015) would likely result in disparate area sizes and sample coverage. It may be possible to restrict the data to a limited spatial area of inference where sampling is more uniform, but this may greatly reduce the sample size and can lead to problems when fishing fleets shift spatial locations (Campbell, 2004). The essential problem

879 is one of confounding where all further results depend critically on the initial spatial partition.
880 In situations where strong year-area interactions such as range contraction/range
881 expansion/spatial depletion are likely to occur, spatio-temporal modeling approaches provide
882 a consistent and compelling means of addressing them.

883 Overall, the application to real observer data collected by the NMFS Pelagic Observer
884 Program suggested that the relative abundance of the blue marlin population of the Atlantic
885 declined over the period 1998-2004 and was relatively stable afterwards. VAST also
886 suggested that Atlantic blue marlin COG may have moved slightly westward between 1998
887 and 2014 (Fig. 11a). Such a trend could potentially be indicative of a small spatial overlap
888 between the U.S. pelagic longline fishery and the expansion of the oxygen minimum zone in
889 the Eastern Atlantic (Stramma et al., 2012).

890 An issue observed with real observer data was the peak in abundance predicted in
891 1996-1997 by the GLMwt.int, GLMprwt, GLMprwt.int and VAST methods (Figs. 9 and 10).
892 An examination of the spatial distribution of observer data indicated that these peak CPUEs
893 occurred off the northeast coast of Brazil and were the result of only three trips by two fishing
894 vessels in 1996 and 1997 which had exceptionally high catch rates of blue marlin. Trips in
895 these locations were rarely ever observed in the remaining time series and these three trips
896 represent the only data for these southeasternmost spatial areas in 1996 and 1997, indicating
897 that they have substantial leverage on the estimations. However, the standard errors associated
898 with the predictions for the VAST knot defined from the observer data collected off northeast
899 Brazil and adjacent knots were low (Fig. A10), which is indicative that the model prediction
900 uncertainty was not increased by having very few samples to extrapolate over a large surface
901 area. This is in contrast to traditional geostatistical theory where, assuming stationarity of
902 spatial autocorrelation, having only a few samples to cover the entire southeastern prediction
903 region would result in very large standard errors relative to prediction regions in other areas

904 which were much more comprehensively sampled. This indicates that, in application, care in
905 developing the prediction knots is necessary to avoid overpredicting beyond the range of
906 spatial autocorrelation. While methods such as VAST can avoid the *a priori* specification of
907 spatial strata, they are not devoid of making some decisions regarding spatial structuring of
908 the prediction area. Hence, some greater curation of the placement of knots when setting up
909 VAST modeling approaches is recommended when working with spatially imbalanced
910 fishery-dependent datasets. This was not done in the present study, because we worked in a
911 design where the model developer was purposefully not provided with any details regarding
912 the datasets being analyzed, though this is an important issue to consider in future studies.
913 Contrary to the GLMwt.int, GLMprwt, GLMprwt.int and VAST methods, the GLMMint
914 method did not predict a peak in abundance in 1996-1997. This is because the GLMMint
915 method is a non-spatial method that models year-area interactions as random effects and that
916 does not assign weights to year-area strata; thus, the few trips off the northeast coast of Brazil
917 in 1996-1997 did not have a strong influence on the predictions made by the GLMMint
918 method.

919 The main avenues for future research we envision are the following ones: (1) an
920 analysis of the consequences of differential patterns of observer coverage, spatial sampling
921 distribution or observer bias; (2) improved consideration of spatial knot selection for VAST;
922 and (3) evaluating the performance of CPUE standardization methods under conditions of
923 stronger year-area interactions designed to mimic environmental changes and when area-
924 season and/or year-season interactions are considered. First, in the present study, we
925 developed algorithms to mimic sampling by an observer program, which allocate longline sets
926 to fishing trips so as to enable the application of CPUE standardization methods to 10% of the
927 fishing trips; this percentage was chosen because this is the average percentage of trips
928 sampled by the NMFS Pelagic Observer Program each year (Beerkircher et al., 2002).

However, future studies working with LLSIM data should take advantage of our algorithms to investigate whether sampling less or more than 10% of the fishing trips undertaken by the U.S. pelagic longline fishery would significantly alter the accuracy and precision of the indices of relative abundance estimated from CPUE data. Additionally, given the effect of a small number of spatial "outlier" trips in 1996 and 1997 on some indices using the real blue marlin data, it may be necessary to consider the potential influence of more isolated and sparse spatial samples which appear in the real data, as well as the impacts of variation in observer coverage across years. Second, given that sparse spatial samples can have undue influence on population trends and potentially, on COG inferences, it may be necessary to re-evaluate the methodology of knot allocation which allocates knots spatially with a density proportional to sampling intensity (Thorson et al., 2015). Hence there are few knots where sampling intensity is low so that these few knots represent a very large spatial area. Future studies could instead place knots with uniform spatial area (i.e., using a two-dimensional grid), and this would likely have better performance when applied to spatially unbalanced datasets like those explored in the present study. Finally, evaluating CPUE standardization methods under conditions of stronger year-area interactions would be valuable for informing climate-vulnerability assessments, and also for checking whether the GLMwt.int and GLMprwt.int methods would then perform better than the GLMwt and GLMprwt methods, as would be expected in theory (Campbell, 2015). We also recommend future studies to examine the performance of CPUE standardization methods when area-season and/or year-season interactions are considered. In the case of Atlantic blue marlin, these interactions terms should explain more variation in the CPUE standardization models and would likely provide more contrast in the performance evaluation, given that seasonal environmental changes are considered in LLSIM and also largely influence Atlantic blue marlin ecology in the real world (Goodyear, 2016). Future studies will also need to explore whether CPUE standardization

models which include year-area, area-season and/or year-season interactions as random effects terms are adequate when these interactions are strong.

In conclusion, the varying performance of the different CPUE standardization methods reflect their different treatments of spatio-temporal variation with the spatio-temporal method providing a more comprehensive and consistent treatment of this variation. This is in contrast with methods that simply weight predictions by large spatial areas, where it is critically important but particularly difficult to get the *a priori* spatial stratification correct before weighting. If year-area interactions are truly small in magnitude, random, spurious or ignorable, then the GLMMint method provides fairly good performance in CPUE standardization. The GAM method is another valuable alternative to spatio-temporal CPUE standardization methods. Moreover, some CPUE standardization methods not considered in this study, such as random forests (Li et al., 2015) or variants of the GLMwt.int and GLMprwt.int methods modeling year-area interactions as random effects (Campbell, 2015), could be employed in future studies. However, as issues of range contraction/expansion and shifts increase in frequency with environmental changes, evaluating them through the lens of arbitrary spatial strata will likely impede both detection and quantification of these phenomena. Hence, we encourage future studies to consider spatio-temporal modeling platforms such as VAST for standardizing fisheries-dependent CPUEs in different marine regions, so as to enable a generalization of the performance of spatio-temporal methods for standardizing fisheries-dependent CPUE data.

Authorship statement

AG, JFW, EAB and JTT designed and analyzed the models; AG, JFW, EAB, FCF, JTT, MVL and MJS conceived the models; AG, JFW, EAB, FCF, JTT and MVL wrote the paper; all authors have approved the final article.

Acknowledgments

This work was supported in part by a NOAA grant through the Cooperative Institute for Marine and Atmospheric Studies at the University of Miami [grant number NA150AR4320064]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are very grateful to the two reviewers (Robert Campbell and Henning Winker), as well as to two NOAA internal reviewers (Adyan Rios and David Hanisko), whose comments have greatly improved the quality and scope of our manuscript. We also thank Larry Beerkircher, Phil Goodyear and Clay Porch for their help or advice at different levels of this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version of the manuscript.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. The Am. Stat. 52, 119–126.

Barry, S.C., Welsh, A.H., 2002. Generalized additive modelling and zero inflated count data. Ecol. Model. 157, 179–188.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., Dai, B., Grothendieck, G., 2015. Package 'lme4.'

Beerkircher, L.R., Lee, D.W., Brown, C.J., 2002. SEFSC pelagic observer program data summary for 1992-2000. NOAA Technical Memorandum NMFS-SEFSC-486 (22 p).

1003 Benoît, H.P., Allard, J., 2009. Can the data from at-sea observer surveys be used to make
1004 general inferences about catch composition and discards? Can. J. Fish. Aquat. Sci. 66,
1005 2025–2039.
1006 Berg, C.W., Nielsen, A., Kristensen, K., 2014. Evaluation of alternative age-based methods
1007 for estimating relative abundance from survey data in relation to assessment models.
1008 Fish. Res. 151, 91–99.
1009 Bigelow, K.A., Maunder, M.N., 2007. Does habitat or depth influence catch rates of pelagic
1010 species? Can. J. Fish. Aquat. Sci. 64, 1581–1594.
1011 Bishop, J., 2006. Standardizing fishery-dependent catch and effort data in complex fisheries
1012 with technology change. Rev. Fish Biol. Fish. 16, 21.
1013 Bolker, B.M., 2008. Ecological models and data in R. Princeton University Press, Princeton,
1014 NJ.
1015 Bourdaud, P., Travers-Trolet, M., Vermard, Y., Cormon, X., Marchal, P., 2017. Inferring the
1016 annual, seasonal, and spatial distributions of marine species from complementary
1017 research and commercial vessels' catch rates. ICES J. Mar. Sci. 74, 2415–2426.
1018 Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed
1019 models. J. Am. Stat. Assoc. 88, 9–25.
1020 Brill, R.W., 1994. A review of temperature and oxygen tolerance studies of tunas pertinent to
1021 fisheries oceanography, movement models and stock assessments. Fish. Oceanogr. 3,
1022 204–216.
1023 Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval estimation for a binomial proportion.
1024 Stat. Sci. 101–117.
1025 Campbell, R.A., 2015. Constructing stock abundance indices from catch and effort data:
1026 Some nuts and bolts. Fish. Res. 161, 109–130.
1027 Campbell, R.A., 2004. CPUE standardisation and the construction of indices of stock
1028 abundance in a spatially varying fishery using general linear models. Fish. Res. 70,
1029 209–227.
1030 Cao, J., Thorson, J.T., Richards, R.A., Chen, Y., 2017. Spatiotemporal index standardization
1031 improves the stock assessment of northern shrimp in the Gulf of Maine. Can. J. Fsih.
1032 Aquat. Sci. 74, 1781–1793.
1033 Carruthers, T.R., Ahrens, R.N., McAllister, M.K., Walters, C.J., 2011. Integrating imputation
1034 and standardization of catch rate data in the calculation of relative abundance indices.
1035 Fish. Res. 109, 157–167.
1036 Carruthers, T.R., McAllister, M.K., Ahrens, R.N., 2010. Simulating spatial dynamics to
1037 evaluate methods of deriving abundance indices for tropical tunas. Can. J. Fish. Aquat.
1038 Sci. 67, 1409–1427.
1039 Chang, S.-K., 2003. Analysis of Taiwanese white marlin catch data and standardization of
1040 catch rates. Col. Vol. Sci. Pap. ICCAT 55, 453–466.
1041 Cooke, J.G., 1997. A procedure for using catch-effort indices in bluefin tuna assessments.
1042 Col. Vol. Sci. Pap. ICCAT 46, 228–232.
1043 Forrestal, F.C., Goodyear, C.P., Schirripa, M., 2019a. Applications of the longline simulator
1044 (LLSIM) using US pelagic longline logbook data and Atlantic blue marlin. Fish. Res.
1045 211, 331–337.
1046 Forrestal, F.C., Goodyear, C.P., Schirripa, M.J., Babcock, E.A., Lauretta, M., Sharma, R.,
1047 2017. Testing robustness of CPUE standardization using simulated data: Findings of
1048 initial blind trials. Col. Vol. Sci. Pap. ICCAT 74, 391–403.
1049 Forrestal, F.C., Schirripa, M., Goodyear, C.P., Arrizabalaga, H., Babcock, E.A., Coelho, R.,
1050 Ingram, W., Lauretta, M., Ortiz, M., Sharma, R., 2019b. Testing robustness of CPUE
1051 standardization and inclusion of environmental variables with simulated longline catch
1052 datasets. Fish. Res. 210, 1–13.

Goodyear, C.P., 2017. Simulating longine catch with LLSIM: a user's guide (version 2) (23 p).
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.
Goodyear, C.P., 2003. Tests of the robustness of habitat-standardized abundance indices using simulated blue marlin catch–effort data. Mar. Freshwater Res. 54, 369–381.
Goodyear, C.P., Schirripa, M.J., Forrestal, F.C., 2017. Longline data simulation: A paradigm for improving CPUE standardization. Col. Vol. Sci. Pap. ICCAT 74, 379–390.
Grüss, A., Biggs, C., Heyman, W.D., Erisman, B., 2018a. Prioritizing monitoring and conservation efforts for fish spawning aggregations in the US Gulf of Mexico. Sci. Rep. 8, 8473.
Grüss, A., Chagaris, D.D., Babcock, E.A., Tarnecki, J.H., 2018b. Assisting Ecosystem-Based Fisheries Management Efforts Using a Comprehensive Survey Database, a Large Environmental Database, and Generalized Additive Models. Mar. Coast. Fish. 10, 40–70.
Grüss, A., Drexler, M., Ainsworth, C.H., 2014. Using delta generalized additive models to produce distribution maps for spatially explicit ecosystem models. Fish. Res. 159, 11–24.
Grüss, A., Drexler, M.D., Ainsworth, C.H., Babcock, E.A., Tarnecki, J.H., Love, M.S., 2018c. Producing Distribution Maps for a Spatially-Explicit Ecosystem Model Using Large Monitoring and Environmental Databases and a Combination of Interpolation and Extrapolation. Front. Mar. Sci. 5, 16.
Grüss, A., Drexler, M.D., Chancellor, E., Ainsworth, C.H., Gleason, J.S., Tirpak, J.M., Love, M.S., Babcock, E.A., 2019. Representing species distributions in spatially-explicit ecosystem models from presence-only data. Fish. Res. 210, 89–105.
Grüss, A., Perryman, H.A., Babcock, E.A., Sagarese, S.R., Thorson, J.T., Ainsworth, C.H., Anderson, E.J., Brennan, K., Campbell, M.D., Christman, M.C., et al., 2018d. Monitoring programs of the US Gulf of Mexico: inventory, development and use of a large monitoring database to map fish and invertebrate spatial distributions. Rev. Fish Biol. Fisher., doi: 10.1007/s11160-018-9525-2.
Grüss, A., Thorson, J.T., Sagarese, S.R., Babcock, E.A., Karnauskas, M., Walter III, J.F., Drexler, M., 2017. Ontogenetic spatial distributions of red grouper (*Epinephelus morio*) and gag grouper (*Mycteroperca microlepis*) in the US Gulf of Mexico. Fish. Res. 193, 129–142.
Grüss, A., Yemane, D., Fairweather, T.P., 2016. Exploring the spatial distribution patterns of South African Cape hakes using generalised additive models. Afr. J. Mar. Sci. 38, 395–409.
Ichinokawa, M., Brodziak, J., 2010. Using adaptive area stratification to standardize catch rates with application to North Pacific swordfish (*Xiphias gladius*). Fish. Res. 106, 249–260.
Johnson, K.F., Councill, E., Thorson, J.T., Brooks, E., Methot, R.D., Punt, A.E., 2016. Can autocorrelated recruitment be estimated using integrated assessment models and how does it affect population forecasts? Fish. Res. 183, 222–232.
Li, Z., Ye, Z., Wan, R., Zhang, C., 2015. Model selection between traditional and popular methods for standardizing catch rates of target species: a case study of Japanese Spanish mackerel in the gillnet fishery. Fish. Res. 161, 312–319.
Lo, N.C., Jacobson, L.D., Squire, J.L., 1992. Indices of relative abundance from fish spotter data based on delta-lognornial models. Can. J. Fish. Aquat. Sci. 49, 2515–2526.
Lynch, P.D., Shertzer, K.W., Latour, R.J., 2012. Performance of methods used to estimate indices of abundance for highly migratory species. Fish. Res. 125, 27–39.

Marchal, P., Andersen, B., Bromley, D., Iriondo, A., Mahévas, S., Quirijns, F., Rackham, B., Santurtún, M., Tien, N., Ulrich, C., 2006. Improving the definition of fishing effort for important European fleets by accounting for the skipper effect. Can. J. Fish. Aquat. Sci. 63, 510–533.

Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scand. J. Stat. 39, 53–74.

Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70, 141–159.

Maunder, M.N., Sibert, J.R., Fonteneau, A., Hampton, J., Kleiber, P., Harley, S.J., 2006. Interpreting catch per unit effort data to assess the status of individual stocks and communities. ICES J. Mar. Sci. 63, 1373–1385.

Maunder, M.N., Starr, P.J., 2003. Fitting fisheries models to standardised CPUE abundance indices. Fish. Res. 63, 43–50.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. Chapman and Hall, London, UK.

Miyabe, N., Takeuchi, Y., 2003. Standardized bluefin CPUE from the Japanese longline fishery in the Atlantic including those for mixing studies. Col. Vol. Sci. Pap. ICCAT 55, 1190–1207.

Nakano, H., 1989. Stock status of Pacific swordfish, *Xiphias gladius*, inferred from CPUE of the Japanese longline fleet standardized using general linear models. NOAA Technical Memorandum NMFS-SEFSC-142 142, pp. 195–209.

National Marine Fisheries Service, 2016. U.S. National Bycatch Report First Edition Update 2. U.S. Department of Commerce (90 p).

Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat. Med. 17, 857–872.

Ono, K., Punt, A.E., Hilborn, R., 2015. Think outside the grids: An objective approach to define spatial strata for catch and effort analysis. Fish. Res. 170, 89–101.

Pereira, J.C., Leandro, R.A., Petrere Jr, M., Nishida, T., 2012. Comparison between univariate and bivariate geostatistical models for estimating catch per unit of effort (cpue): A simulation study. Fish. Res. 121, 115–125.

Punt, A.E., Walker, T.I., Taylor, B.L., Pribac, F., 2000. Standardization of catch and effort data in a spatially-structured shark fishery. Fish. Res. 45, 129–145.

Sharma, R., Pons, M., Martin, S., Kell, L., Walter, J., Lauretta, M., Schirripa, M., Anderson, H. editor: E., 2017. Factors related to the decline and rebuilding of billfish stocks in the Atlantic and Indian oceans. ICES J. Mar. Sci. 75, 880–891.

Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53, 577–588.

Stow, C.A., Jolliff, J., McGillicuddy Jr, D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. J. Mar. Syst. 76, 4–15.

Stramma, L., Prince, E.D., Schmidtko, S., Luo, J., Hoolihan, J.P., Visbeck, M., Wallace, D.W., Brandt, P., Körtzinger, A., 2012. Expansion of oxygen minimum zones may reduce available habitat for tropical pelagic fishes. Nat. Climate Change 2, 33.

Su, N.-J., Sun, C.-L., Punt, A.E., Yeh, S.-Z., DiNardo, G., 2011. Modelling the impacts of environmental variation on the distribution of blue marlin, *Makaira nigricans*, in the Pacific Ocean. ICES J. Mar. Sci. 68, 1072–1080.

Thompson, S., 2002. Sampling. Wiley, New York, NY.

Thorson, J.T., 2019. Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fish. Res. 210, 143–161.

Thorson, J.T., 2017. Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative. Can. J. Fish. Aquat. Sci. 1369–1392.
Thorson, J.T., Barnett, L.A., 2017. Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat. ICES J. Mar. Sci. 74, 1311–1321.
Thorson, J.T., Fonner, R., Haltuch, M.A., Ono, K., Winker, H., 2016. Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. Can. J. Fish. Aquat. Sci. 74, 1794–1807.
Thorson, J.T., Shelton, A.O., Ward, E.J., 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, 1297–1310.
Thorson, J.T., Ward, E.J., 2013. Accounting for space–time interactions in index standardization models. Fish. Res. 147, 426–433.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth edition. Springer-Verlag, New York, NY.
Walter, J.F., Hoenig, J.M., Christman, M.C., 2014a. Reducing bias and filling in spatial gaps in fishery-dependent catch-per-unit-effort data by geostatistical prediction, I. methodology and simulation. North Am. J. Fish. Manag. 34, 1095–1107.
Walter, J.F., Hoenig, J.M., Christman, M.C., 2014b. Reducing bias and filling in spatial gaps in fishery-dependent catch-per-unit-effort data by geostatistical prediction, II. application to a scallop fishery. North Am. J. Fish. Manag. 34, 1108–1118.
Walters, C., 2003. Folly and fantasy in the analysis of spatial catch rate data. Can. J. Fish. Aquat. Sci. 60, 1433–1436.
Wilberg, M.J., Thorson, J.T., Linton, B.C., Berkson, J., 2010. Incorporating time-varying catchability into population dynamic stock assessment models. Rev. Fish. Sci. 18, 7–24.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.
Wood, S.N., 2006. Generalized additive models: an introduction with R. Chapman and Hall, London, UK.
Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157, 157–177.
Ye, Y., Dennis, D., 2009. How reliable are the abundance indices derived from commercial catch–effort standardization? Can. J. Fish. Aquat. Sci. 66, 1169–1178.
https://github.com/James-Thorson/VAST

Figure captions

Fig. 1. Study region, located in the North Atlantic, which encompasses the ten National Marine Fisheries Service (NMFS) areas defined for the stock assessments of the International Commission for the Conservation of Atlantic Tunas (ICCAT): (1) the Gulf of Mexico (GOM); (2) the Mid Atlantic Bight (MAB); (3) the South Atlantic Bight (SAB); (4) Florida East Coast (FEC); (5) the Caribbean (CAR); (6) the Northeast Coastal area (NEC); (7) the Sargasso area (SAR); (8) the Northeast Distant area (NED); (9) the North Central Atlantic (NCA); and (10) the Offshore South area (OFS).

Fig. 2. (a-c) Relative abundance, **(d, g, j)** eastward center of gravity (COG; in km), **(e, h, k)** northward COG (in km) and **(f, i, l)** effective area occupied (in ln(km²)) of the three virtual populations of blue marlin (*Makaira nigricans*) over the period 1987-2015. **(a, d-f)** are for Population 1, **(b, g-i)** are for Population 2, and **(c, j-l)** are for Population 3. **(a-c)** show the true annual abundances of the virtual populations of blue marlin relative to their mean over the period 1987-2015, while **(d-l)** were estimated by the spatio-temporal modeling platform VAST from all the blue marlin catch-per-unit-effort data provided by the longline catch-per-unit-effort data simulator LLSIM. For **(d-l)**, the shaded area represents 95% confidence intervals.

Fig. 3. Evolution over the period 1987-2015 of the sample size of the levels of the factors considered in the analyses conducted with data from the longline catch-per-unit-effort (CPUE) data simulator LLSIM. Here, all the data provided by LLSIM are considered.

Fig. 4. Annual time series of nominal and estimated catch-per-unit-effort (CPUE) relative to mean CPUE for the virtual population of blue marlin (*Makaira nigricans*) #1, under the

10%BIAS scenario (see legend for color code). Replicate #2 is considered here. Nine methods were employed to estimate CPUEs (Table 1). The annual time series of the true simulated abundance of the virtual population of blue marlin #1 divided by its mean simulated abundance is also given here. The dashed lines represent the 95% confidence intervals of estimated CPUEs.

Fig. 5. Annual time series of nominal and estimated catch-per-unit-effort (CPUE) relative to mean CPUE for the virtual population of blue marlin (*Makaira nigricans*) #3, under the 10%BIAS scenario (see legend for color code). Replicate #1 is considered here. Nine methods were employed to estimate CPUEs (Table 1). The annual time series of the true simulated abundance of the virtual population of blue marlin #3 divided by its mean simulated abundance is also given here. The dashed lines represent the 95% confidence intervals of estimated CPUEs.

Fig. 6. Bias of estimated annual catches-per-unit-effort (CPUEs) for the simulated populations of blue marlin (*Makaira nigricans*), under three scenarios (ALL, 10%, and 10%BIAS). For the 10% and 10%BIAS scenarios, barplots represent mean biases over five replicates, while the black bars overlaid on barplots represent minimum and maximum biases over the five replicates. See the main text for details on the scenarios. Nine methods were employed to estimate CPUEs (Table 1).

Fig. 7. Mean absolute error (MAE) of estimated annual catches-per-unit-effort (CPUEs) for the simulated populations of blue marlin (*Makaira nigricans*), under three scenarios (ALL, 10%, and 10%BIAS). For the 10% and 10%BIAS scenarios, barplots represent mean MAEs over five replicates, while the black bars overlaid on barplots represent minimum and

maximum MAEs over the five replicates. See the main text for details on the scenarios. Nine methods were employed to estimate CPUEs (Table 1).

Fig. 8. Coverage (in %) for the simulated populations of blue marlin (*Makaira nigricans*), under three scenarios (ALL, 10%, and 10%BIAS). Coverage is the percentage of years over the period 1987-2015 the 50% confidence interval for a normalized estimated catch-per-unit-effort (CPUE) contains the normalized true abundance. For the 10% and 10%BIAS scenarios, barplots represent mean coverages over five replicates, while the black bars overlaid on barplots represent minimum and maximum coverages over the five replicates. See the main text for details on the scenarios. Nine methods were employed to estimate CPUEs (Table 1).

Fig. 9. Annual time series of nominal and estimated catch-per-unit-effort (CPUE) relative to mean CPUE for the Atlantic blue marlin (*Makaira nigricans*) population, computed from the data collected within the National Marine Fisheries Service Pelagic Observer Program over the period 1992-2017. All the methods listed in Table 1 except the GLMint method were employed to estimate CPUEs.

Fig. 10. Annual time series of estimated catch-per-unit-effort (CPUE) relative to mean CPUE for the Atlantic blue marlin (*Makaira nigricans*) population, computed from the data collected within the National Marine Fisheries Service Pelagic Observer Program over the period 1992-2017. All the methods listed in Table 1 except the GLMint method were employed to estimate CPUEs.

Fig. 11. (a) Eastward center of gravity (COG; in km), **(b)** northward COG (in km) and **(c)** effective area occupied (in ln(km²)) of the Atlantic blue marlin (*Makaira nigricans*)

population, estimated by the spatio-temporal modeling platform VAST from the data collected within the National Marine Fisheries Service Pelagic Observer Program over the period 1992-2017. For all panels, the shaded areas represent 95% confidence intervals.

1268 **Tables**

1269 **Table 1.** Overview of the nine catch-per-unit-effort (CPUE) standardization methods used in

1270 this study.

Method	Overview
GLM	Method using generalized linear models (GLMs) that integrate fixed year and area effects.
GLMint	Method using GLMs that integrate fixed year and area effects and a fixed year-area interaction term.
GLMMint	Method using generalized linear mixed models (GLMMs) that integrate fixed year and area effects and a random year-area interaction term.
GLMwt	Method that (1) uses GLMs integrating fixed year and area effects; and (2) takes into account the surface area of the areas making up the study region to weight CPUE observations.
GLMwt.int	Method that (1) uses GLMs integrating fixed year and area effects and a fixed year-area interaction term; and (2) takes into account the surface area of the areas making up the study region to weight CPUE observations.
GLMprwt	Method that (1) uses GLMs integrating fixed year and area effects; (2) takes into account the surface area of the areas making up the study region to weight CPUE observations; and (3) assigns prior weights to raw CPUE data based on the year-area stratum to which the raw CPUE data belong.
GLMprwt.int	Method that (1) uses GLMs integrating fixed year and area effects and a fixed year-area interaction term; (2) takes into account the surface area of the areas making up the study region to weight CPUE observations; and (3) assigns prior weights to raw CPUE data based on the year-area stratum to which the raw CPUE data belong.
GAM	Method using generalized additive models (GAMs) that integrate an interaction term between eastings and northings accounting for spatial autocorrelation at a broad spatial scale.
VAST	Method using spatio-temporal models that account for both spatial and spatio-temporal autocorrelations at a fine spatial scale.

1271 **Table 2.** Factors considered in the analyses conducted with data from the longline catch-per-

1272 unit-effort (CPUE) data simulator LLSIM.

Factor	**Levels**	**Level with the largest sample size in the full dataset**	**Level with the largest sample size in the dataset containing non-zero CPUE data**
Year	1987-2015	1989	2013
Season	Winter (January-March), spring (April-June), summer (July-September), fall (October-December)	Summer	Winter
National Marine Fisheries Service (NMFS) area	Gulf of Mexico (GOM), Mid Atlantic Bight (MAB), South Atlantic Bight (SAB), Florida East Coast (FEC), Caribbean (CAR), Northeast Coastal (NEC), Sargasso (SAR), Northeast Distant (NED), North Central Atlantic (NCA), Offshore South (OFS)	NEC	FEC
Type of hook used ("hook")	Circle hook, J and circle hooks, unknown	Unknown	Unknown
Type of bait used ("bait")	Artificial, dead, live, unknown	Dead	Dead
Number of light sticks used ("light")	0, 1-500, 501-1500, unknown	3	3
Number of hooks between floats ("hbf")	2, 3, 4, 5, 6	4	4

Table 3. "Walters' table" (Campbell, 2015) constructed from the raw data from the longline catch-per-unit-effort (CPUE) data simulator LLSIM, showing the number of data points in each year-area stratum. Here, the ALL scenario is considered; see the main text for details on scenarios.

Year\NMFS area	Caribbean (CAR)	Florida East Coast (FEC)	Gulf of Mexico (GOM)	Mid Atlantic Bight (MAB)	North Central Atlantic (NCA)	Northeast Coastal (NEC)	Northeast Distant (NED)	Offshore South (OFS)	South Atlantic Bight (SAB)	Sargasso (SAR)
1987	568	2619	3208	9	2169	436	942	43	258	252
1988	700	3247	2720	37	478	2339	1510	204	760	275
1989	556	3795	2286	33	493	3454	1822	131	883	331
1990	610	3003	1780	29	534	3921	1210	273	1350	246
1991	552	2777	2055	23	408	4172	1198	78	1078	261
1992	431	2718	2113	17	448	3792	1251	151	1051	313
1993	644	2439	1677	67	459	3767	1167	65	1361	392
1994	715	2310	1595	72	945	3908	984	81	1619	578
1995	507	2340	2070	83	1426	4478	916	280	1341	280
1996	589	2510	2422	105	891	3125	766	645	2261	468
1997	527	2764	2534	57	379	3228	793	678	1689	265
1998	416	2231	2426	21	337	2911	634	342	1241	305
1999	188	2407	2953	16	170	2456	447	258	1240	139
2000	323	2640	2837	16	105	2258	639	77	969	173
2001	242	1501	3281	38	185	2711	351	75	949	215
2002	205	1710	2920	17	100	1992	524	102	683	288
2003	152	1513	3358	26	181	1508	583	34	710	235
2004	307	1633	3202	31	79	1667	466	30	715	207
2005	156	1370	2397	12	127	1701	483	35	630	153
2006	66	1292	1977	26	166	2098	427	158	579	157
2007	27	1704	1870	30	72	2245	348	229	978	180
2008	85	2140	1529	35	76	2505	345	172	943	357
2009	33	1842	2442	20	6	2079	330	202	900	549
2010	52	1980	622	21	41	2507	323	223	966	462
2011	12	1887	866	33	50	2542	296	164	995	745
2012	7	2265	1907	71	37	3061	444	206	1016	756
2013	17	2290	1427	64	35	3086	408	215	1008	992
2014	11	2121	1583	67	71	2788	385	158	978	1012
2015	20	1171	1042	40	58	2530	302	145	777	1100

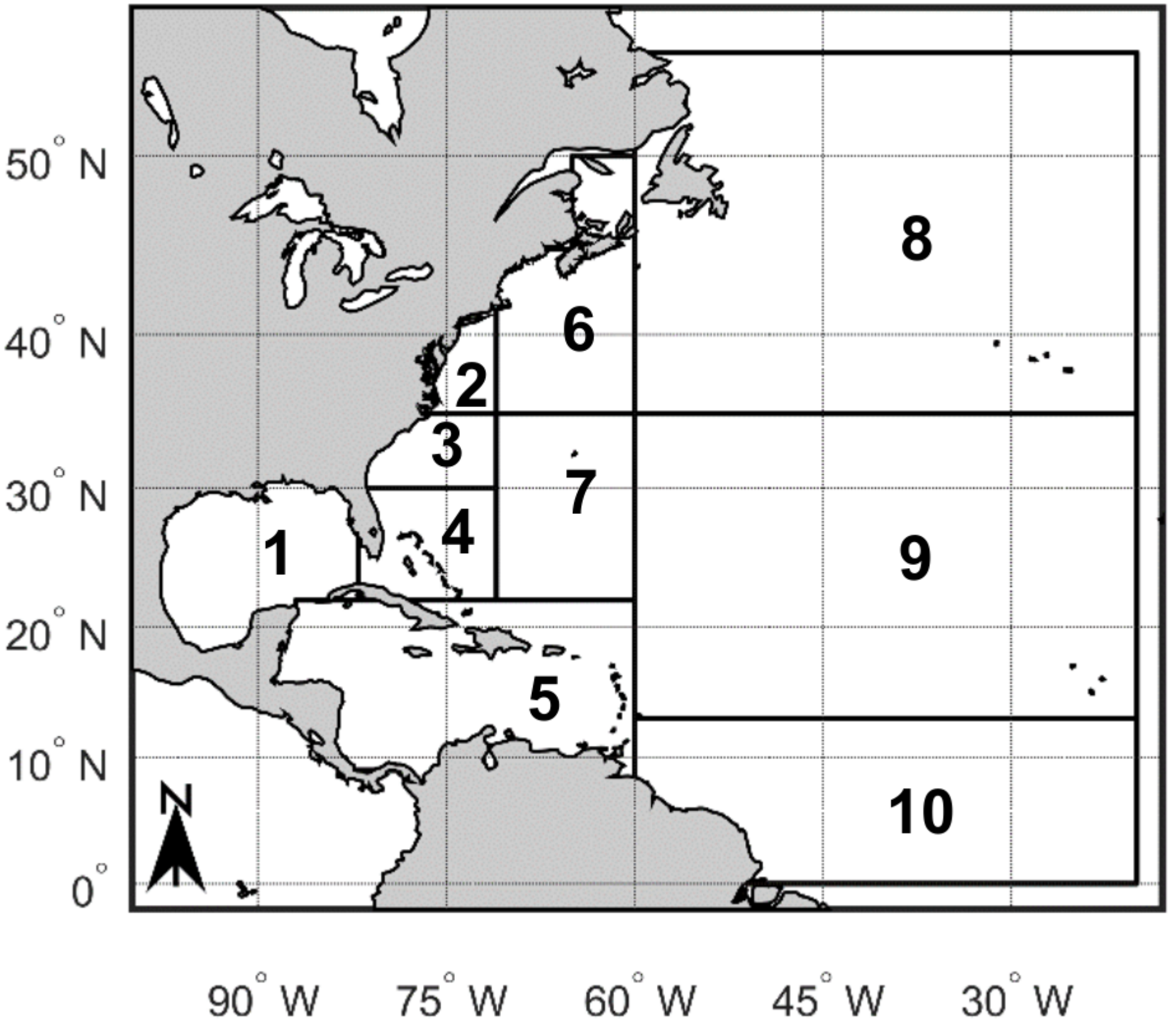
1276 **Table 4.** Factors considered in the analyses conducted with data collected within the National

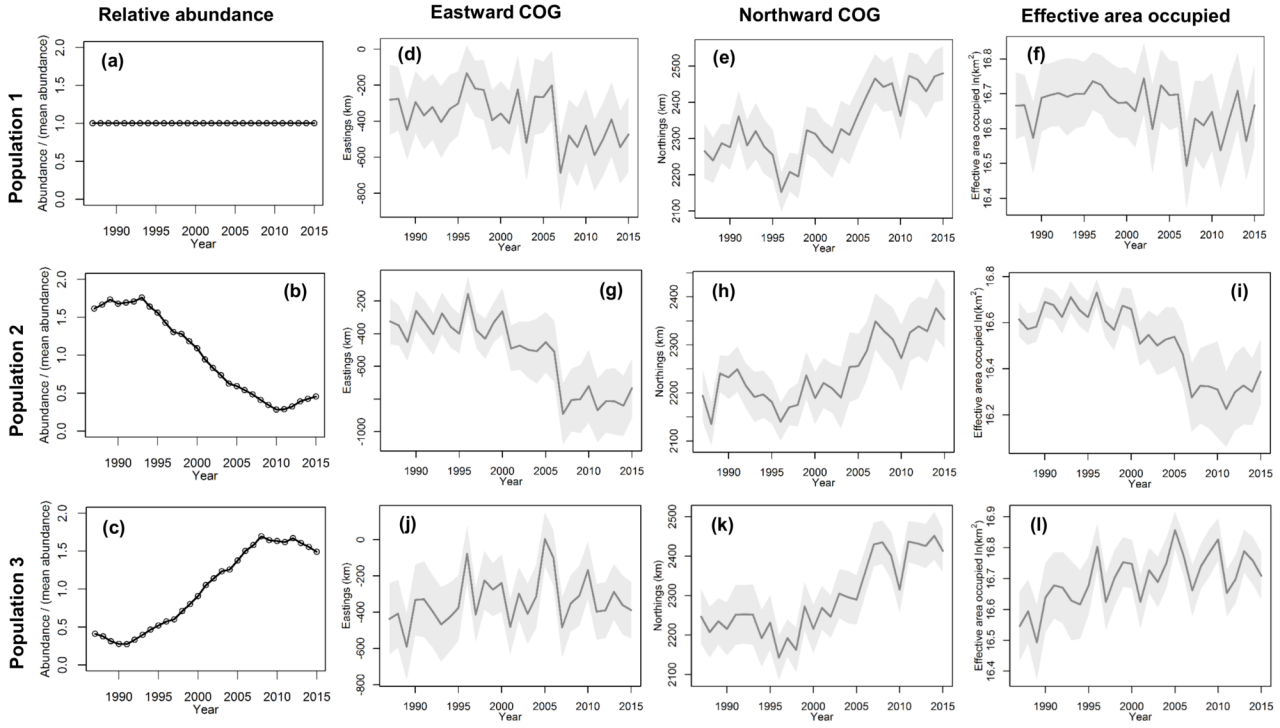
1277 Marine Fisheries Service (NMFS) Pelagic Observer Program.

Factor	**Levels**	**Level with the largest sample size in the full dataset**	**Level with the largest sample size in the dataset containing non-zero CPUE data**
Year	1992-2017	2013	2009
Season	Winter (January-March), spring (April-June), summer (July-September), fall (October-December)	Spring	Spring
NMFS area	Gulf of Mexico (GOM), Mid Atlantic Bight (MAB), South Atlantic Bight (SAB), Florida East Coast (FEC), Caribbean (CAR), Northeast Coastal (NEC), Sargasso (SAR), Northeast Distant (NED), North Central Atlantic (NCA), Offshore South (OFS)	GOM	GOM
Type of hook used ("hook")	Circle hook, J hook, unknown	Circle hook	Circle hook
Number of light sticks used ("light")	0, 1-500, 501-1500, unknown	3	3
Number of hooks between floats ("hbf")	(0-4.02], (4.02-4.15], (4.15-5.19], (5.19-318]	2	2

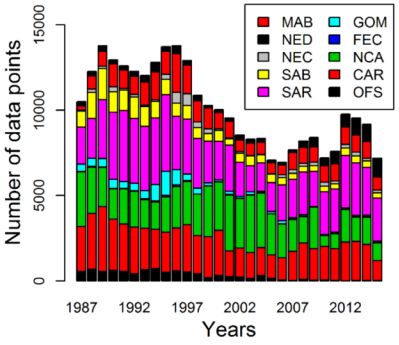
Table 5. "Walters' table" (Campbell, 2015) constructed from the raw data collected within the National Marine Fisheries Service (NMFS) Pelagic Observer Program, showing the number of data points in each year-area stratum.

Year\NMFS area	Caribbean (CAR)	Florida East Coast (FEC)	Gulf of Mexico (GOM)	Mid Atlantic Bight (MAB)	North Central Atlantic (NCA)	Northeast Coastal (NEC)	Northeast Distant (NED)	Offshore South (OFS)	South Atlantic Bight (SAB)	Sargasso (SAR)
1992	11	10	35	65	0	33	70	0	29	0
1993	41	18	203	181	52	68	75	0	65	0
1994	35	19	113	151	19	77	61	0	40	0
1995	47	14	193	136	83	51	65	0	29	0
1996	6	7	115	12	41	11	0	27	61	9
1997	9	13	150	36	19	64	42	25	29	1
1998	10	31	73	53	8	23	0	4	49	0
1999	17	22	160	38	2	23	40	8	31	3
2000	0	29	167	61	14	48	47	0	43	0
2001	10	13	198	64	15	21	1	0	61	0
2002	21	63	158	58	12	16	0	0	19	1
2003	4	55	269	69	46	36	0	0	51	17
2004	39	51	264	88	3	23	76	0	61	32
2005	10	30	303	92	16	3	14	0	64	22
2006	0	31	273	89	0	49	48	10	54	17
2007	19	64	615	110	0	13	44	13	50	10
2008	0	87	828	117	0	83	28	21	38	0
2009	0	113	862	143	0	63	39	7	117	29
2010	16	98	375	142	0	66	34	10	96	42
2011	0	129	341	111	0	95	32	17	119	42
2012	0	153	451	128	0	76	0	19	81	40
2013	6	144	828	201	3	94	29	13	106	48
2014	0	142	565	176	0	60	22	29	195	45
2015	14	128	415	234	0	68	45	30	153	43
2016	10	95	528	249	0	65	36	41	181	7
2017	7	70	295	263	1	28	25	11	174	15

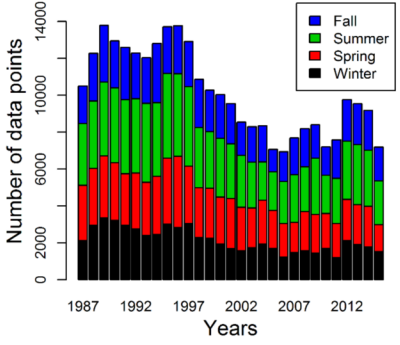

50° N
40° N
30° N
20° N
10° N
0°
90° W
75° W
60° W
45° W
30° W
N
1
2
3
4
5
6
7
8
9
10


Relative abundance
Eastward COG
Northward COG
Effective area occupied
Population 1
Population 2
Population 3
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
Abundance / (mean abundance)
Eastings (km)
Northings (km)
Effective area occupied ln(km^2)
Year

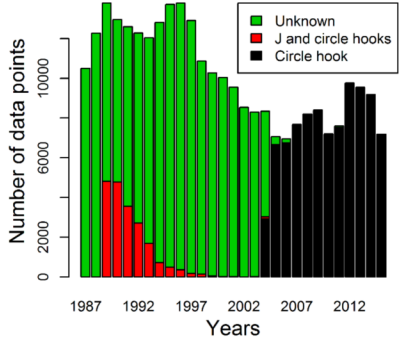
National Marine Fisheries Service area



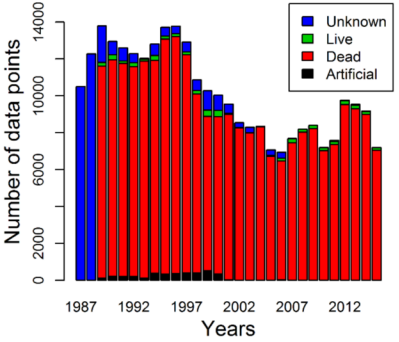
Season



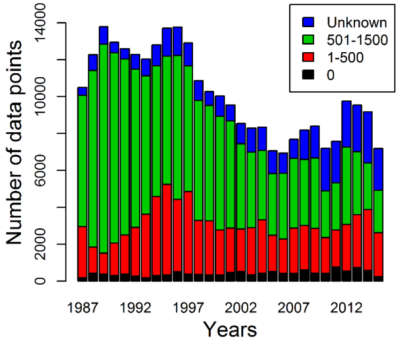
Type of hook used ("hook")



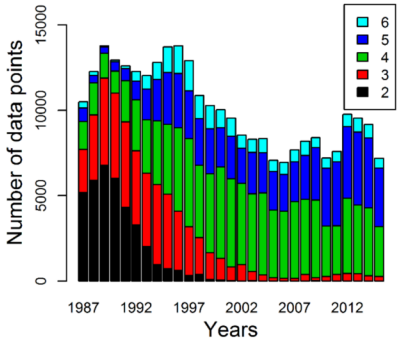
Type of bait used ("bait")



Number of light sticks used ("light")

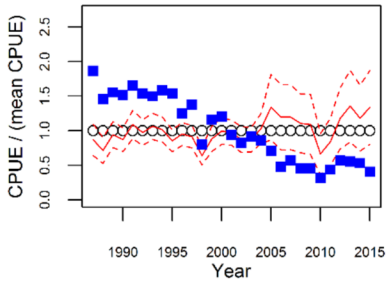


Number of hooks between floats ("hbf")

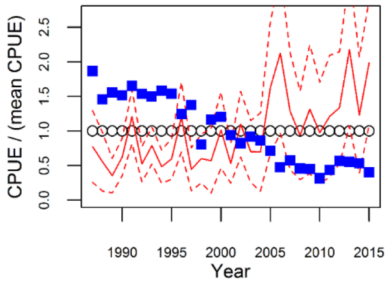


Population 1 – Scenario 10%BIAS - Replicate 2

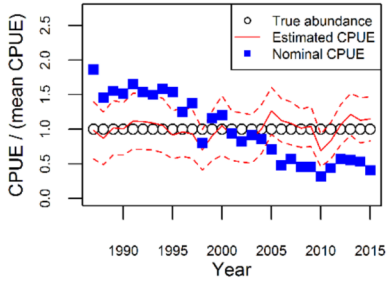
GLM



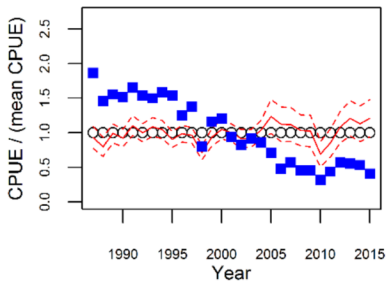
GLMint



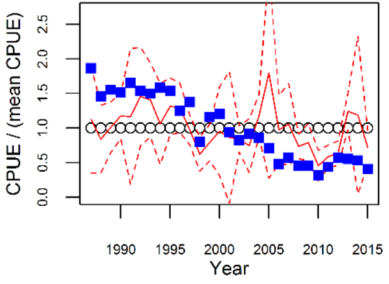
GLMMint



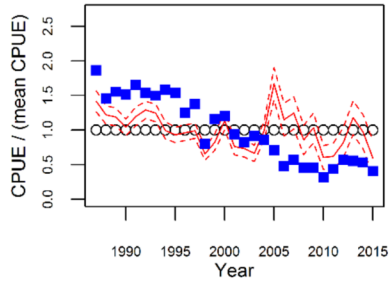
GLMwt



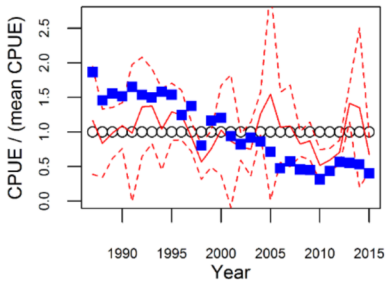
GLMwt.int



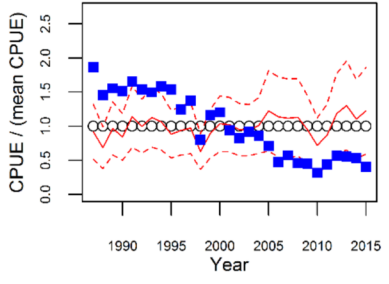
GLMprwt



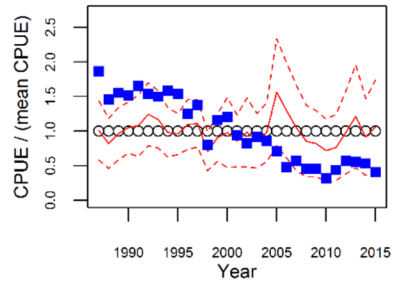
GLMprwt.int



GAM

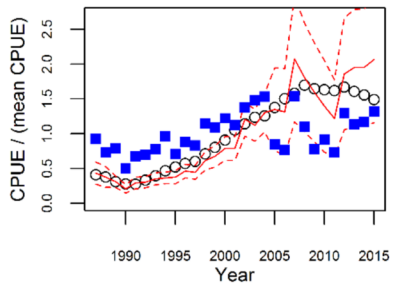


VAST

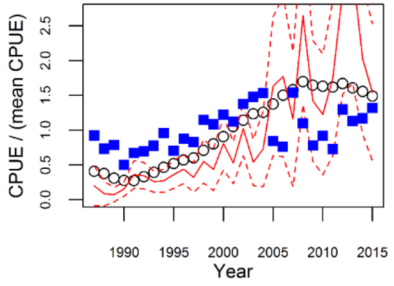


Population 3 – Scenario 10%BIAS - Replicate 1

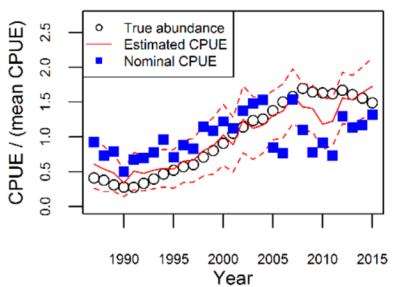
GLM



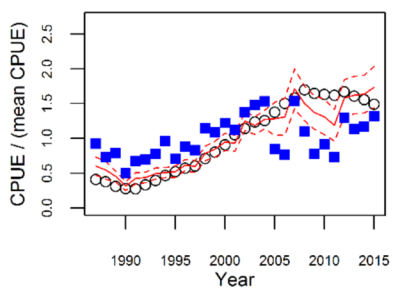
GLMint



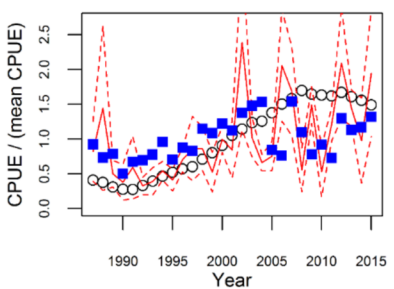
GLMMint



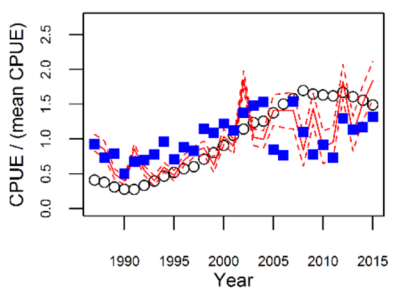
GLMwt



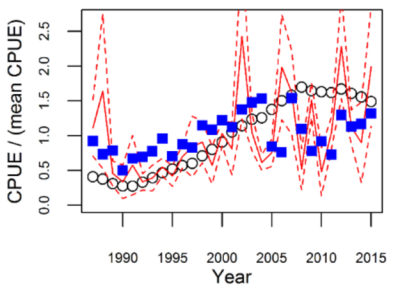
GLMwt.int



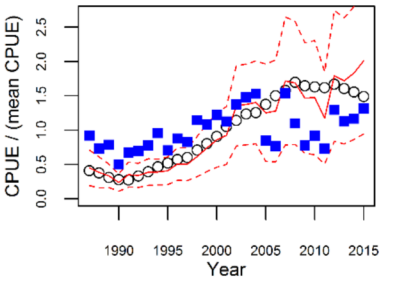
GLMprwt



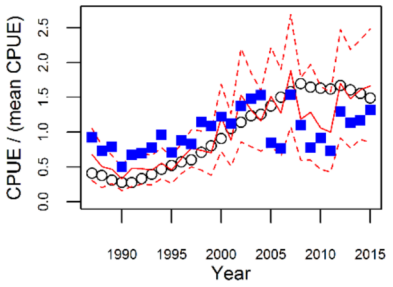
GLMprwt.int

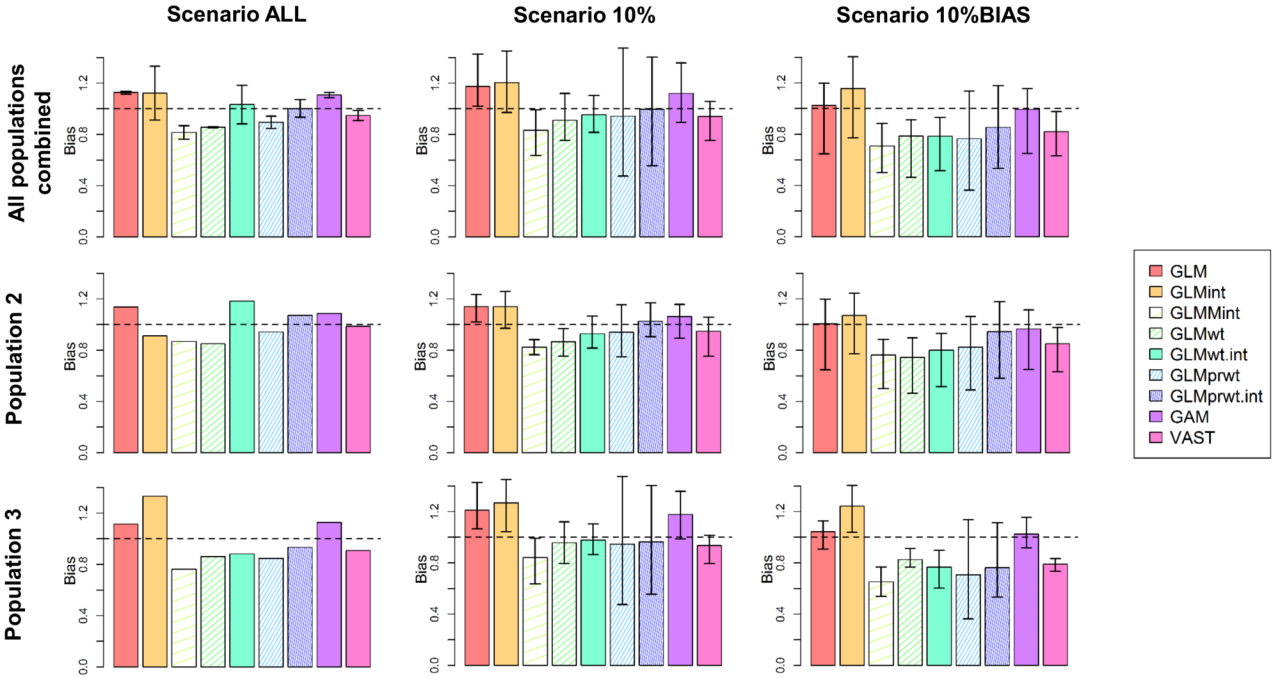


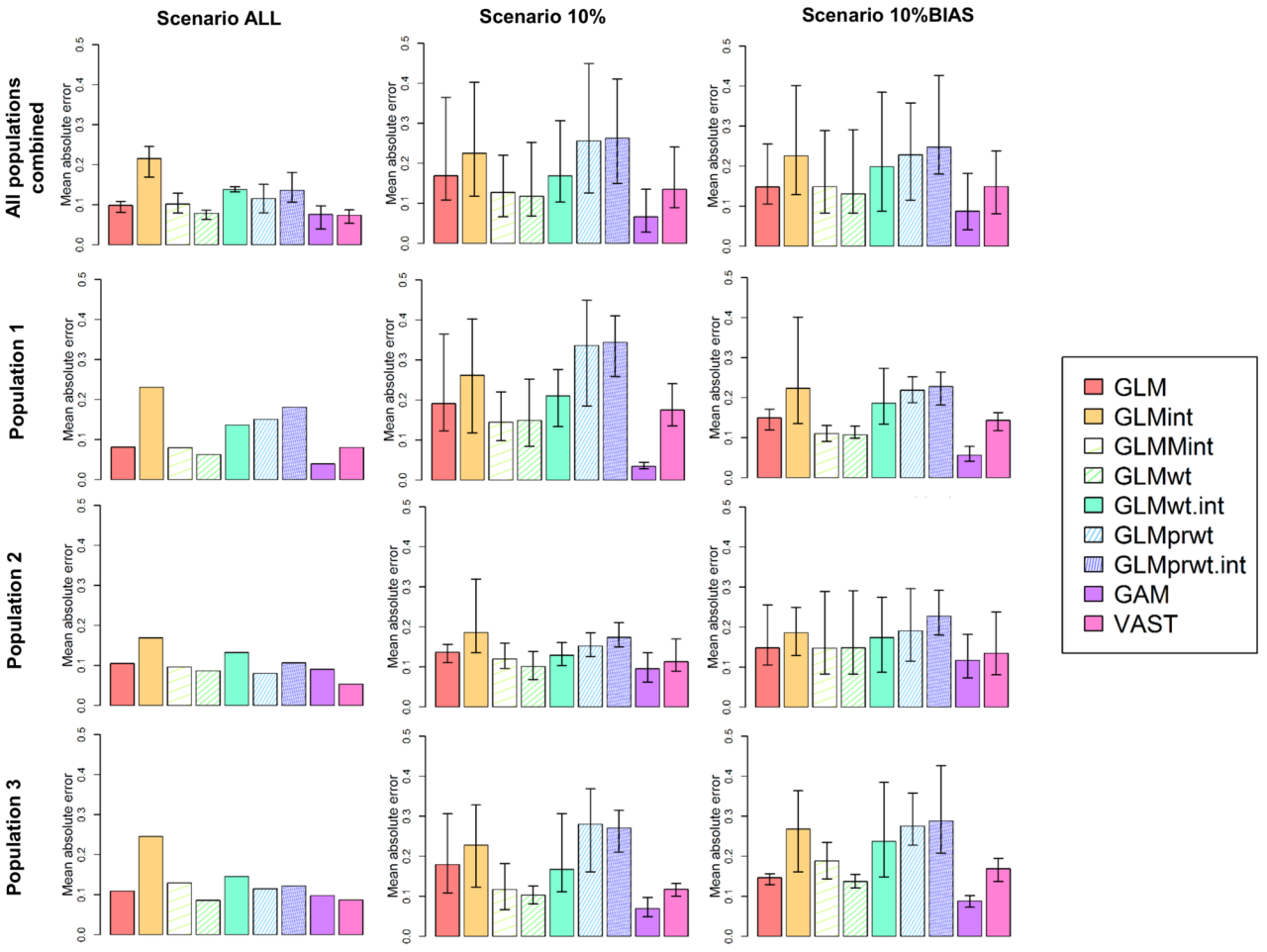
GAM

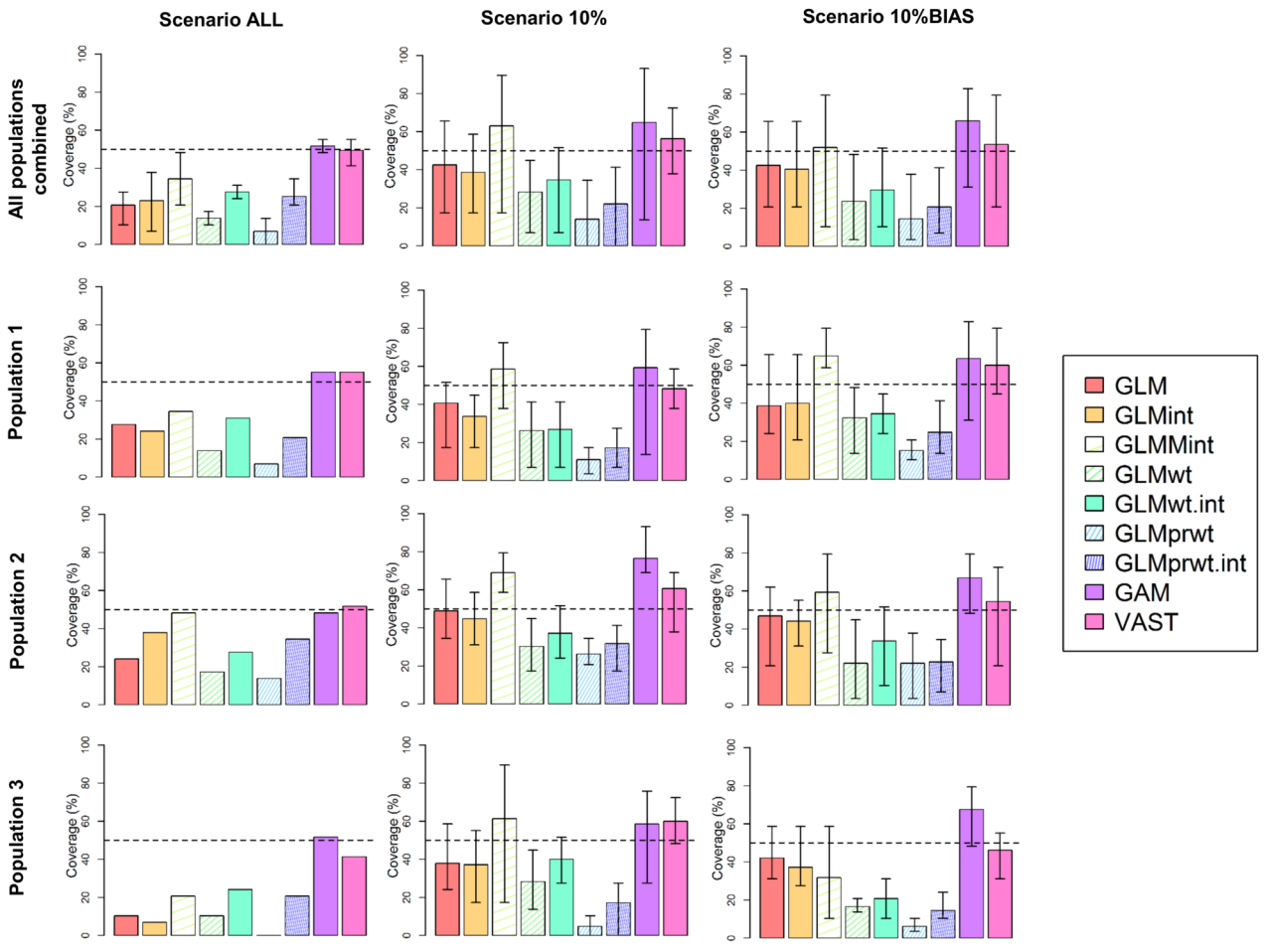


VAST

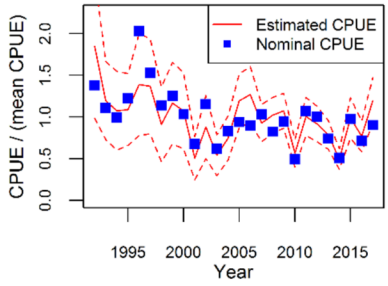



Scenario ALL
Scenario 10%
Scenario 10%BIAS
All populations combined
Population 2
Population 3
Bias
0.0
0.4
0.8
1.2
GLM
GLMint
GLMMint
GLMwt
GLMwt.int
GLMprwt
GLMprwt.int
GAM
VAST

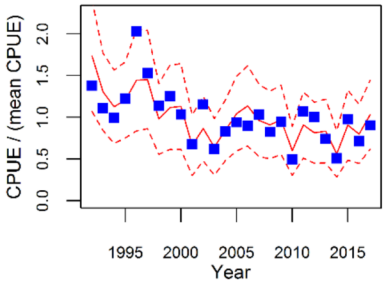

Scenario ALL
Scenario 10%
Scenario 10%BIAS
All populations combined
Population 1
Population 2
Population 3
Mean absolute error
0.0
0.1
0.2
0.3
0.4
0.5
GLM
GLMint
GLMMint
GLMwt
GLMwt.int
GLMprwt
GLMprwt.int
GAM
VAST


Scenario ALL
Scenario 10%
Scenario 10%BIAS
All populations combined
Population 1
Population 2
Population 3
Coverage (%)
0
20
40
60
80
100
GLM
GLMint
GLMMint
GLMwt
GLMwt.int
GLMprwt
GLMprwt.int
GAM
VAST

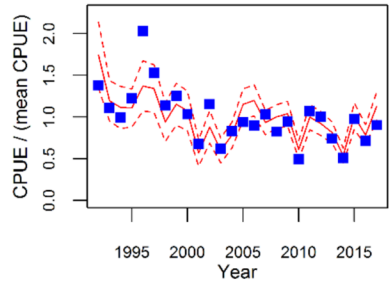
GLM



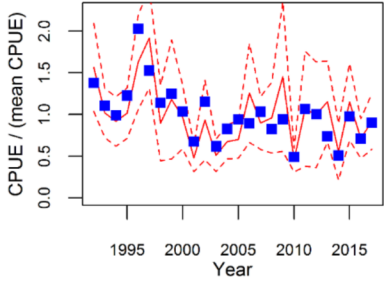
GLMMint



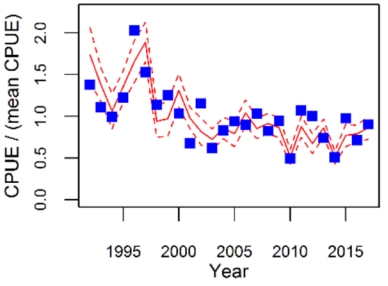
GLMwt



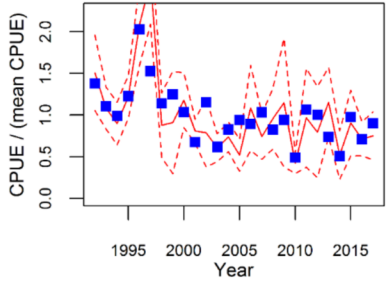
GLMwt.int



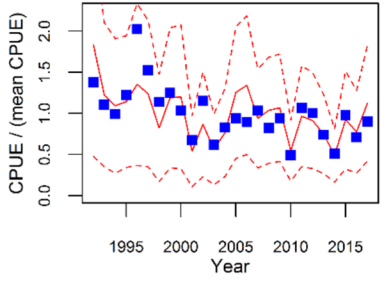
GLMprwt



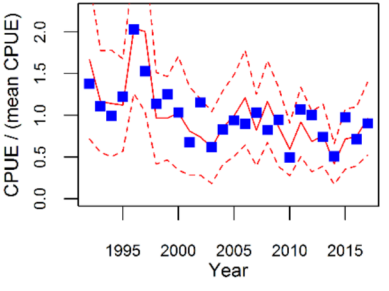
GLMprwt.int

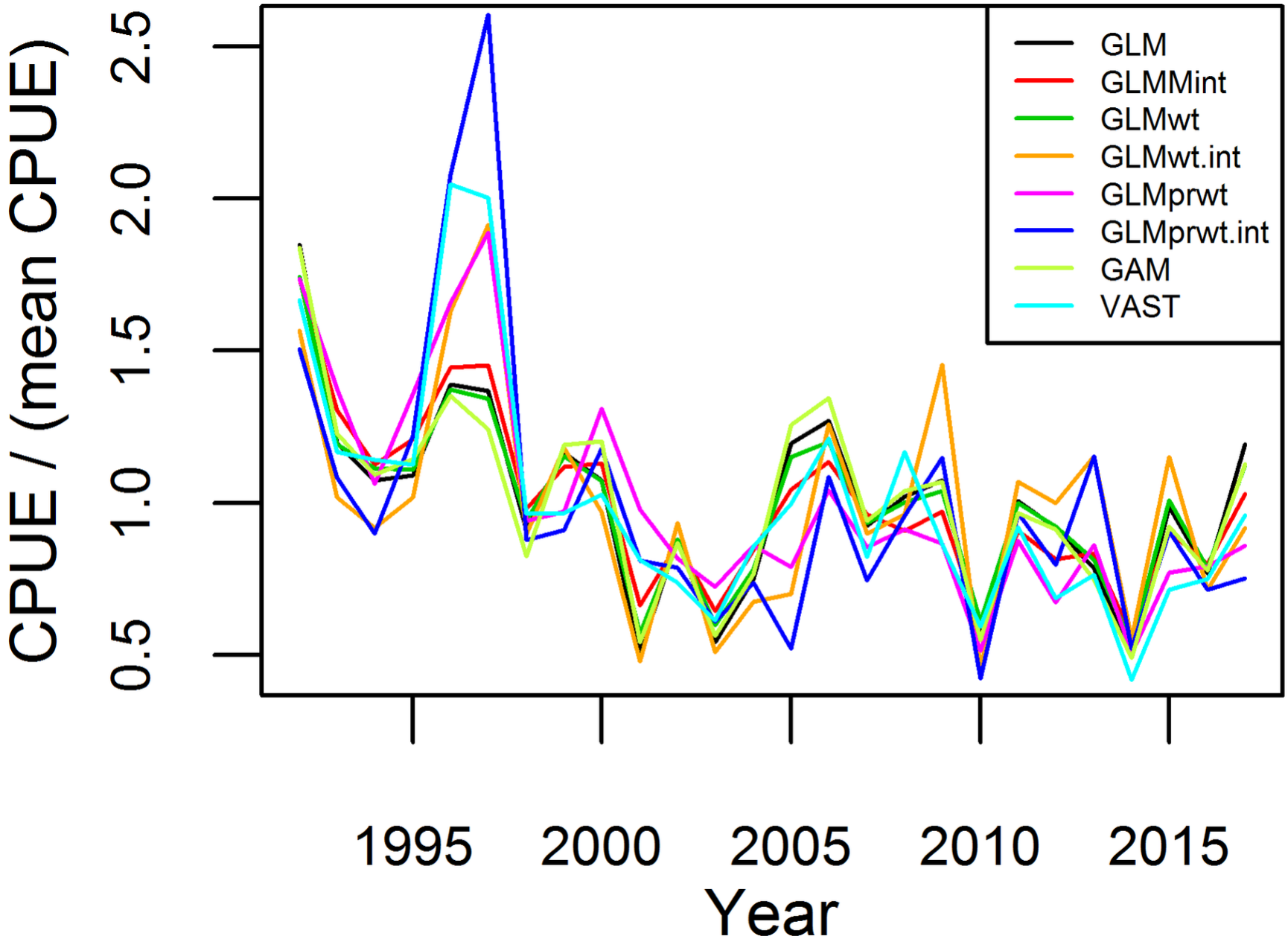


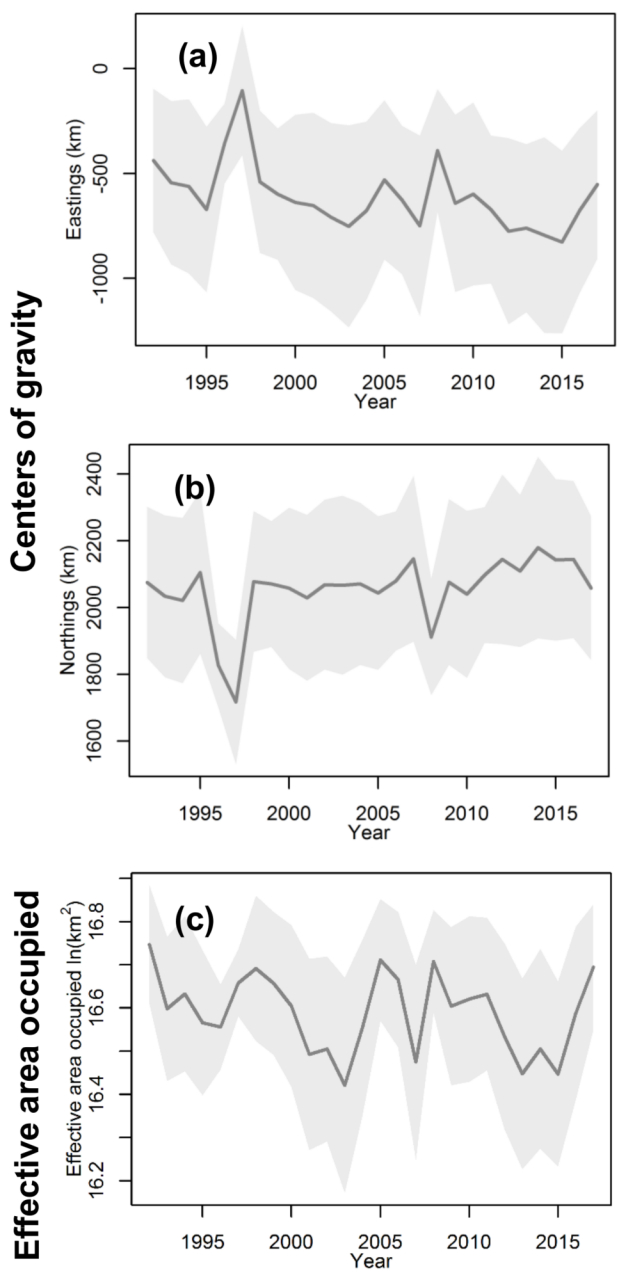
GAM



VAST




CPUE / (mean CPUE)
0.5
1.0
1.5
2.0
2.5
1995
2000
2005
2010
2015
Year
GLM
GLMMint
GLMwt
GLMwt.int
GLMprwt
GLMprwt.int
GAM
VAST


Centers of gravity
(a)
Eastings (km)
0
-500
-1000
1995
2000
2005
2010
2015
Year
(b)
Northings (km)
2400
2200
2000
1800
1600
1995
2000
2005
2010
2015
Year
Effective area occupied
(c)
Effective area occupied ln(km^2)
16.8
16.6
16.4
16.2
1995
2000
2005
2010
2015
Year