REPORT OF THE WORKING GROUP ON MULTI-NATIONAL PELAGIC LONGLINE INDEX FOR WESTERN ATLANTIC BLUEFIN TUNA

J. Walter, M. Lauretta, A. Kimoto, A. Hanke, K. Ramirez, G. Melvin

SUMMARY

*This document summarizes the results of a small SCRS working group tasked to evaluate the feasibility of combining set by set data from the Japan, Canada, Mexico and United States pelagic longline fishing fleets to obtain a CPUE index for western Atlantic bluefin tuna, while maintaining data confidentiality. In 2016 the group successfully combined datasets, assigned relevant environmental and gear variables and produced a dataset of 109,063 individual longline sets over 1992-2015 from the Gulf of Mexico and the Atlantic Ocean north of 25oN latitude and west of 45oW longitude. The objective of the second meeting was to evaluate whether statistical modeling approaches could account for the very different target and non-target fishing strategies of each fleet to create a combined index. The mandate of the group was to test the null hypothesis that the fleets were different; rejection of this hypothesis would indicate that the fleets, once standardized, gave a similar relative catch rate trend. The group outlined three tests: significance of year*Flag interactions; trend, magnitude and randomness of the year*flag coefficients and strength of the correlation between an index constructed with and without year*flag interactions. The ATL index failed to reject the first two hypotheses indicating significant, non-random year*flag interactions and the GOM index failed to reject the second test. Nevertheless, indices constructed with and without year*flag interactions showed strong (>80%) correlations. The group determined that until the non-random year*flag interactions could be reconciled statistical modeling of the combined 3 fleet Atlantic or 2-fleet Gulf of Mexico indies was not warranted.*

RÉSUMÉ

*Ce document représente les résultats d'un petit groupe de travail du SCRS chargé d'évaluer la faisabilité de combiner des données recueillies opération par opération auprès des flottilles palangrières pélagiques du Japon, du Canada, du Mexique et des Etats-Unis pour obtenir un indice de CPUE pour le thon rouge de l'Atlantique Ouest, tout en maintenant la confidentialité des données. En 2016, le groupe est parvenu à combiner des jeux de données, à attribuer des variables pertinentes liées à l'environnement et à l'engin et à élaborer un jeu de données de 109.063 opérations palangrières individuelles de la période 1992-2015 du golfe du Mexique et de l'océan Atlantique au nord de 25ºN de latitude et à l'ouest de 45ºW de longitude. L'objectif de la deuxième réunion était d'évaluer si des approches de modélisation statistique pourraient expliquer les très différentes stratégies de pêche ciblées et non ciblées de chaque flottille pour créer un indice combiné. Le mandat du groupe consistait à tester l'hypothèse nulle selon laquelle les flottilles étaient différentes ; le rejet de cette hypothèse indiquerait que les flottilles, une fois standardisées, dégageaient une tendance similaire des taux de capture relative. Le groupe a décrit trois tests : signification des interactions année*pavillon ; tendance, ampleur et effet aléatoire des coefficients année*pavillon et force de la corrélation entre un indice construit avec et sans les interactions année*pavillon. L'indice ATL n'a pas réussi à réfuter les deux premières hypothèses, ce qui indique de considérables interactions non aléatoires année*pavillon et l'indice GOM n'a pas réfuté le deuxième essai. Néanmoins, des indices construits avec ou sans les interactions année*pavillon ont montré des corrélations (> 80 %). Le groupe a déterminé que tant que les interactions non aléatoires année* pavillon ne pourraient pas être réconciliées, il ne serait pas possible de garantir la modélisation statistique des indices combinés des trois flottilles de l'Atlantique ou des deux flottilles du golfe du Mexique.*

RESUMEN

Este documento resume los resultados de un pequeño grupo de trabajo encargado de evaluar la viabilidad de combinar los datos lance por lance de las flotas pesqueras de palangre pelágico de Japón, Canadá, México y Estados Unidos para obtener un índice de CPUE para el atún rojo del Atlántico occidental, manteniendo a la vez la confidencialidad de los datos. En 2016 el Grupo tuvo éxito a la hora de combinar los conjuntos de datos, asignando variables medioambientales y de arte pertinentes,

*y elaboró un conjunto de datos de 109.063 lances individuales de palangre para el periodo de 1992-2015 procedentes del golfo de México y del océano Atlántico al norte de 25ºN de latitud y al oeste de 45ºW de longitud. El objetivo de la segunda reunión era evaluar si los enfoques de modelación estadística podrían tener en cuenta las muy diferentes estrategias de pesca dirigida y no dirigida de cada flota para crear un índice combinado. El mandato del Grupo era probar la hipótesis nula de que las flotas eran diferentes, el rechazo de esta hipótesis indicaría que las flotas, una vez estandarizadas, producen una tendencia similar de la tasa de captura relativa. El Grupo describió tres pruebas: importancia de las interacciones año*pabellón, magnitud de la tendencia y aleatoriedad de los coeficientes año*pabellón y fuerza de la correlación entre un índice elaborado con interacciones año*pabellón y otro elaborado sin dichas interacciones. El índice ATL no descartó las dos primeras hipótesis indicando interacciones de año*pabellón importantes y no aleatorias, y el índice del GOM no rechazó la segunda prueba. No obstante, los índices construidos con y sin interacciones año*pabellón mostraban fuertes correlaciones (>80%). El Grupo determinó que hasta que no puedan conciliarse las interacciones año*pabellón no aleatorias, no podría justificarse la modelación estadística combinada de las tres flotas del Atlántico o de las dos flotas del golfo de México.*

KEYWORDS

Bluefin tuna, CPUE standardization, longline

Introduction

Similar to the 2016 meeting a small (1-2 representatives from each CPC involved with the western bluefin tuna assessment), met for a 5 day intercessional workshop to build upon the positive outputs of the Cercedilla workshop (Walter *et al.*, 2017). The objective of the Mexico City meeting was to investigate the possibility of statistically modeling combined longline datasets from Mexico, USA, Japan and Canada to produce one or more CPUE indices for western Atlantic bluefin tuna. In particular, the meeting primary objective was to evaluate the feasibility of statistically modeling combined datasets, focusing on whether the disparities between target and non-target fleets can be reconciled through the standardization modeling process. If statistical modeling is deemed feasible for 2 or more of the fleet/area combinations, then combined indices would be produced in advance of the Data Preparatory Working group meeting and considered for use in the 2017 stock assessment.

Prior to the meeting each participant confirmed, cleaned and evaluated all data in the combined dataset, noting irregularities. Draft diagnostics, CPUE modeling and hypothesis testing codes were developed by the working group at the meeting.

2. Methods

Although a major effort was put into updating, cleaning and standardizing each countries data prior to arriving at meeting, a number of issues were identified that had to be addressed before the data could be combined. Much of the first day was spent getting the data into a common (e.g. standardized column labels) and clean (eg., minor errors) format for each country. Once the revisions were completed, the data were combined; Canada, USA, and Japan data were combined for the western Atlantic, and USA and Mexico data were combined for the Gulf of Mexico. The group recommended the creation of the two separate datasets for exploration of combined indices, given the difference in stock structure and mixing rates between the two areas; fish caught in the Gulf of Mexico represent the western BFT spawning ground, whereas those along the coast of North America are likely of mixed and variable origin. The dataset was for the Northwest Atlantic and ranged between latitude 25^0 and 50^0N and west of 45^0W longitude. The second dataset for the Gulf of Mexico combined data from the USA fleet in the northern Gulf and Mexico fleet operating in the southern Gulf. Due to the absence of data for one or more years in the overall combined time series, the working group agreed to use only the data starting in 1993 for both datasets for the testing of hypotheses.

Temporal domain of data

As per the SCRS recommendation, data was modeled up to and including 2015. The 2016 data could be incorporated if it becomes available without major changes to the code. For the Atlantic dataset, fishing year rather than calendar year was used to be consistent with how the Japanese longline fleet operates. Fishing year is calculated as the calendar year from May to December and then retains that fishing year for the following year data from January to April. This split was determined from looking at the breaks in the fishing season for each fleet so as to be consistent with the Japanese longline fishing seasons. The group noted that for use in assessment models it would be important to adjust the timing of any potential Atlantic index to reflect this timing. As fishing year was used as the year factor, the complete 2016 fishing year data may not be available until 2017. For the Gulf of Mexico, the year was the calendar year and the months used in the analysis were December-June. Note that future treatments of the Gulf of Mexico index should shift the year for December to the next year to be consistent with the spawning season.

For the purposes of hypothesis testing for the utility of combining data, two years where zero catches of BFT for one CPC was removed from the combined dataset as it led to an inability to estimate year*flag interactions. For the Gulf of Mexico, the year 1993 was removed as zero BFT were observed in the Mexico dataset and for the Atlantic the year 1996 was removed due to zero BFT in the Canada dataset. The Group noted that if a final combined indices were deemed appropriate, these years of data could be used in the models, depending upon the nature of interactions that are estimated (i.e. if year*flag interaction was determined to be insignificant).

2.1 Preliminary analyses

Distributional assumptions

The working group discussed the need for an objective criteria using goodness of fit tests to determine which distributional modeling approach should be used: single negative binomial, two stage (delta-log normal) or zero inflated model for either the catch or the catch/effort data. The results of the goodness of fit tests for the western Atlantic and the Gulf of Mexico are presented below.

Alternative error distribution hypotheses and goodness of fit test results: Atlantic

1. Single stage discrete negative binomial on number of bluefin tuna observed. Chi square goodness of fit indicated failure to reject the null hypothesis of the observed data as being significantly different from a random sample of equal sample size generated from the best fit negative binomial distribution (**Figures 4-6**).
2. Two-stage delta model, continuous gamma distribution fit to the positive catch rates. Kolmogorov-Smirnov test failed due to ties in the data, although fitted gamma distribution comparisons to the observed data indicated a good fit to the positive catch rates.
3. Two-stage delta model, continuous Gaussian distribution fit to the log-e transformed positive catch rates. Kolmogorov-Smirnov test failed due to ties in the data, however, fitted Gaussian distribution indicated very poor fit to the log-transformed positive catch rates.

Distribution options goodness of fit results: Gulf of Mexico

1. Single stage discrete negative binomial on number of bluefin tuna observed. Chi-square goodness of fit indicated failure to reject the null hypothesis of the observed data as being significantly different from a random sample of equal sample size generated from the best fit negative binomial distribution.
2. Two-stage delta model, continuous gamma distribution fit to the positive catch rates. Kolmogorov-Smirnov test failed due to ties in the data, although fitted gamma distribution comparisons to the observed data indicated a good fit to the positive catch rates.

3. Two-stage delta model, continuous Gaussian distribution fit to the log transformed positive catch rates. Kolmogorov-Smirnov test failed due to ties in the data, however, fitted Gaussian distribution indicated very poor fit to the log-transformed positive catch rates.

In the end it was decided to use the negative binomial model for both Atlantic and Gulf of Mexico but to test whether a zero-inflated negative binomial (using GLMM.ADMB package for R) or a negative binomial gave a better fit to the data. For both the Atlantic and the Gulf datasets the negative binomial resulted in a lower bias-corrected Akaike information criterion (AICc) than the zero-inflated model, indicating a better fit.

The group also ran the model selection process for factor inclusion for the binomial component of the delta models, to examine if the hypothesis test results and overall conclusions differed from those of the single-stage negative binomial model. The group noted that the binomial model selected for the same set of fixed factors. Given that frequency of occurrence of bluefin tuna often reflects much of the overall relative abundance trend, the group did not discern a statistical advantage to using a two-stage model (such as delta lognormal) particularly as the negative binomial model more accurately reflects the count nature of catches and can account for both zeros and variable amounts of effort with an offset.

Scope of factors:

Exploratory analysis and evaluation was undertaken for each of the factors to be considered in the modeling of the CPUE indices. The factors included: Effort, sea surface temperature as either categorical or continuous, Month, year or fishing year, average depth of the gear, sea floor depth, mainline length, bait type, hook type and day vs night set. Exploratory analyses were conducted separately for the Gulf of Mexico and the Atlantic datasets. These analyses were conducted to determine appropriate factor levels for modeling categorical variables, sample sizes at each factor level and overall balance of the data at different factor levels across the flags.

Spatial areas

At the Mexico City meeting the group decided to extend the spatial domain of the Atlantic index south to 25^{o}N to accommodate additional observations from both Japan and U.S. Spatial structure for the Atlantic was evaluated by applying the glm.tree adaptive area partitioning algorithm (Ichinokawa and Brodziak 2010) and also exploring the distribution of samples over time. Considering both analyses, we agreed to use the area stratification with 5 splits (Figure 1, one in the south of 35^{o}N, 3 lines in the north of 35^{o}N). During the model exercise, we further checked the number of observations, and found there are many years with zero BFT observations in the area east of 60^{o}W in the south of 35^{o}N. Therefore, the group decided to use the agreed area stratification without that area. Nonetheless, model testing with the use of 5x5 latitude and 5x5 longitude as model factors resulted in improved fit to the models on the basis of a substantial reduction in AICc and hence further analysis of spatial structure could improve model fits, (see also **Figures 11** and **12**).

For the Gulf of Mexico, five areas were originally defined in (Walter *et al.*, 2017). During data exploration there appeared to be substantial year*area interactions within area 3 so this area was split south of 26^{o}N resulting in improved model fit. The final areas included 6 different spatial areas (**Figure 2**). Note that for area 5 (the U.S. bluefin tuna closure area enacted in 2015, this area was modeled as a separate area and included in the modeling as the closure is only for April and May.

Effort:

Comparison of catch and effort indicates a positive, but nonsignificant (r^2) relationship (**Figure 3**) thus an offset for effort is justified. Effort data were tested with effort as categorical and log(effort) offset with AICc model selection to determine how to model the dependent variable (negative binomial model). Visual evaluation of the number of bluefin per effort bin by fleet and AICc were used to determine the best sequence of effort breaks and the appropriate number of categories to best capture trends in catches. The analysis indicated the best bin breaks for categorical effort (number of hooks) was 0, 500, 1000, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, and 5000 hooks.

SST

Relationships between nBFT and sst were plotted to determined appropriate category breaks for factors (**Figures 7 and 8)**. These indicated a mostly dome shaped relationship between SST and catch rate for the Atlantic and declining relationship in the Gulf of Mexico. In the Atlantic, there was a much wider range of temperatures fished, whereas in the Gulf of Mexico most temperatures were above 18^{o}C reflecting the warmer nature of this region and the fact that this area is assumed to be primarily visited by BFT for spawning. Hence it was determined to use different SST categories in the two regions to capture the differential effects of temperature on catches.

Temporal domain

For the Atlantic, all months were included in the modeling (**Figure 9),** however the fishing year was determined, as described above. Given the monthly distribution of catch rates of BFT in the Gulf of Mexico, we decided to model months December and January to June (**Figure 10**). For the Gulf of Mexico, the months of high catch rates for the U.S. and Mexico fleets were offset indicating that a month*fleet interaction might be appropriate.

Hook type

Hook type was originally reported in the combined dataset exactly as recorded by each CPC, or in the case of Japan, assumed to be all the Japan tuna hook. To create categories for analysis all hook types were assigned to one of four categories: CIRCLE, J, weakcircle or unknown/mix hook type in the following proportions:

CIRCLE :J:UNK_MIX:weakcircle equaled 66053:47048:4601:2889.
The following assignments were made as follows: CIRCLE = "CIRCLE","Circle Hook","CIRCULAR","ATUNERO", "bigC" , "smallC", "NORUEGO","AMERICANO" , "GARRA DE AGUILA").
J= "TIBURONERO","HUACHINANGERO","J-HOOK","J Hook","RECTO" , "J","JnR", "JnR","JR" ,"JAPONES","japan"), weak circle = "L2048LM" , "39988D". When hook type was unknown or a mix of hooks were used the category UNK_MIX was assigned.

Bait type

Bait type was originally reported in the combined dataset exactly as recorded by each CPC. As bait type is not recorded in the Japan dataset it was not considered as a model factor for the Atlantic dataset, however when possible it was assigned to one of four categories: DEAD FISH, LIVE FISH, MIX_UNK, or SQUID in the following proportions:
DEAD FISH:LIVE FISH:MIX_UNK:SQUID equaled 11970:43636:51721:13264

The following assignments were made:
DEAD FISH="DEAD HERRING/MACKEREL", "DEAD HERRING", "DEAD BAIT MIX", "DEAD FISH MIX";
LIVE FISH="LIVE FISH" ,"LIVE MACKEREL","LIVE OTHER" ,"LIVE";
SQUID = "DEAD SQUID","LIVE SQUID", "DEAD SQUID/LIGHTS","DEAD SQUID MIX","UNKNOWN SQUID"
MIX_UNK="DEAD OTHER"," DEAD ","DEAD NA","NA NA", "DEAD LIGHTS","DEAD LURES",
"DEAD LIGHTS"," ", " OTHER","DEAD HERRING/MACKEREL/SQUID","DEAD OTHER",
"DEAD HERRING/SQUID", "DEAD MACKEREL/SQUID"," LIVE", "DEAD BAIT (NS)", "OTHER","ARTIFICIAL OTHER", "OTHER","UNK","DEAD LIGHT MIX", "DEAD UNK

Other model factors remained as defined in (Walter *et al.,* 2017).

fHookDepth and fSeaFloorDepth

Hook depth (*fHookDepth*) was used as an approximate proxy in the CPUE standardization to reflect whether fishing gear is set deeper or shallower. It was obtained by averaging the minimum hook depth (obtained as the sum of lengths of float (or drop) lines and branch (or gangion + leader) lines) and the maximum hook depth (obtained assuming a right triangular shape to the longline set) as outlined in (Walter *et al.*, 2017). Visual observation of the depths of fishing were used to determine categories for hook depth as below: 0,50, 75,100,115,125,130, 140,1500 meters with the Japan longline generally fishing deeper sets and Canada the shallowest. *fHookDepth* was modeled as a categorical factor.

Seafloor depth was assigned according to Lauretta *et al.*, 2016 and used also as a categorical factor with the following cuts determined by visual observation of the data: c(-6500,seq(-5000,-1000,by=500),-750,-500,-250,0) meters. There was a general tendency towards high catch rates in shallower depths.

Statistical model fitting

Models were fit with the function glm.nb in package MASS in R. Initial models were built starting with testing all single factors. Model selection was performed by systematic addition of model terms until no remaining term explained more than 1% of the model deviance. This was performed with the R function add1(), also contained in the MASS package library.

Additional model terms deemed critical for either analytical, hypothesis testing or to capture known changes in the fishery were included even if they did not account for more than 1% of the explained deviance. For example, hook type was not significant at the 1% level in the Gulf of Mexico but given the implementation of the weak hook regulation in the U.S. fishery it is important to capture and estimate this effect. Furthermore, flag was not significant but was also included due to the requirement to evaluate flag*year interactions and to obtain predictions for each year for each flag. Given the nature of fishery-dependent catch rate data and large volume of observations, traditional model selection approaches were expected to be of limited utility. Common model selection thresholds based on AICc (e.g., change in AICc by standard thresholds of 2, 7, or 10) resulted in over-parameterized models, and approaches using a certain % reduction in deviance (Ortiz and Diaz 2004) required arbitrary determinations of significance. However, the group noted that inclusion of some factors resulted in a drastically lower AICc value (change in AICc on the order of 100 to 1000s) and also met the deviance reduction criterion threshold (2%). The group also noted that change in AICc by factor inclusion of less than 20 typically did not meet the secondary selection criterion of deviance reduction.

Ultimately we determined that model terms deemed important- or empirically determined to be important through *a priori* knowledge such as weak hook effects (Walter and Cass-Calay 2014) should be included in the modeling even if they do not meet the model selection criteria. When included as model factors the ‘weak hook’ effect could be estimated by the model and resulted in an 51% reduction in catch rate relative to standard circle hooks, very similar to the 47% reduction in catch rates observed in experimental work Upcoming work from the ICCAT methods working group evaluating this idea of including important factors, even if not significant may provide insight into the merits or costs associated with this approach.

The group was presented with several different options for modeling effort, either as an offset or as a categorical factor. This was also tested by comparing models with effort modeled both ways with the model with the lowest AIC chosen. While the AIC model selection criteria was not used for model building, as it tended to favor inclusion of the most complicated model structure resulting in likely overparameterization of the model, differences in AIC were useful to determine the best treatment of a variable. The same testing process was performed to evaluate how best to model SST. For the Atlantic SST was modeled as a categorical factor. For the Atlantic and Gulf, there was a negligible difference between modeling effort as a category or as an offset and it was chosen to model it with an offset which greatly reduced the number of parameters. Similarly, for SST in the Gulf of Mexico, given that there was a generally smooth declining relationship between catch and SST (unlike in the Atlantic), SST was modeled as a continuous factor to reduce the number of parameters that needed to be estimated.

Model predictions by year were obtained using the lsmeans() package for R, except for the Atlantic where these required too much memory so the data and models were exported to SAS to obtain lsmeans.

Interactions

It was also decided to use a logical approach to deal with interactions rather than to test every possible interaction. This was done by evaluating specific hypothesis regarding interactions, which are detailed in the section hypothesis testing. Specific interactions tested were year*flag, year*area, month*area and SST*area, SST*month, and SST *area*month. The year*area and month*area interactions were tested to capture the migratory patterns of the fish (year*area was modeled to examine the hypothesis that the distribution of the fish has changed over time, e.g., poleward shifts; while month*area was modeled to account for the known change in distribution during seasonal migrations between spawning and foraging grounds). Interactions were modeled as fixed effects for the purposes of hypothesis testing. Any ultimate construction of combined indices would require a decision process for how to model year*factor interactions which make the year effect non-unique. One recommendation is to weight year*area interactions by the area in square kilometers of each spatial cell, while accounting for the fact that degrees of longitude represent smaller distances with increasing latitude.

The full initial models tested were as follows:

Atlantic
Full_Model=
glm.nb(nBFT~Flag+fYear+fMonth+fSST+fHookDepth+fSeaFloorDepth+DAY_NIGHT+ fEffort + fLon + fLat)

Gulf of Mexico
Full_Model = glm.nb(nBFT~Flag + fYear +fMonth+GOM_AREA+BAIT+ SST +fHookDepth+DAY_NIGHT +Hook_Type2 + offset (log(effort))

Note that for the Gulf of Mexico **fSeaFloorDepth** was not calculated for all of the data and **effort** was modeled with an offset of log (hooks) and SST was modeled as a linear factor. All factors were modeled as categorical factors except SST in the Gulf of Mexico models.

Hypothesis testing criteria

The mandate of the working group was to test the null hypothesis that the datasets were different and could not be combined. Rejection of the null hypothesis would be that the different fleets were tracking the same time series of relative abundance when seasonality, spatial stratification, environmental conditions (i.e. SST, seafloor depth, and seafloor gradient) and gear effects were standardized, and hence that a combined index could be produced from the 2 (Mexico-US) or 3 datasets (CAN-JPN-US). The hypothesis testing proceeded in 3 steps:

1. Test year*fleet interactions for significance according to reduction in model deviance (significant if greater than 2% reduction in deviance when compared to best model without year*fleet interaction)
2. Plot year*fleet coefficients to evaluate whether they are random (and if random, model as random effects) or whether they have trends. Rejection of null hypothesis based on visual observation of coefficients.
3. Compare predicted index with and without year*flag interactions and compare correlation. Correlation above 80% indicates rejection of null hypothesis.

Results

Final selected models after model selection and expert opinion about inclusion of model factors were as follows (**Tables 1 and 3**):

Atlantic
single=glm.nb(nBFT~fSST+fFishYear+fMonth+fLat+fLon+Flag+offset(log(Effort)),data=atl,maxit=200)

Model with year*flag interactions

yr_flag=glm.nb(nBFT~fSST+fFishYear+fMonth+fLat+fLon+Flag+offset(log(Effort))+Flag*fFishYear,data=atl,init.theta=0.4,maxit=300)

Gulf of Mexico

Single= glm.nb(nBFT~ GOM_AREA + Flag + fMonth + fYear + SST + DAY_NIGHT +Hook_Type2 + offset(log(Effort)) , data=gomNo1993)

Model with year*flag interactions
YearFlag= glm.nb(nBFT~ GOM_AREA + Flag + fMonth + fYear + SST + DAY_NIGHT +Hook_Type2 + fYear* Flag + offset(log(Effort)) , data=gomNo1993)

The results of the hypothesis testing are as follows:

1. Significance of year*flag interactions
    (a) ATL: Yes: 3.83% reduction in deviance over the model without year*flag- fail to reject H0;
    (b) GOM: 1.6% over the model without year*flag= reject H0.
2. Are year*flag interactions random?
    (a) ATL: Pattern is not random (**Figure 13**) due to shift after 1997, after this the pattern does look random and mostly without trend, though from 1997-2005 there is some difference in magnitude between USA and Japan- fail to reject H0.
    (b) GOM: Pattern is not random and there is trend (**Figure 14**) and substantial divergence between US and Mexico- fail to reject H0.
3. Compare predicted index with and without interactions and compare correlation. Correlation above 80% indicates rejection of null hypothesis.
    (a) ATL: R2 = 0.81 significant correlation between indices with and without year*flag interactions (**Figure 13**). Reject H0.
    (b) GOM: R2 = 0.83 significant correlation between indices with and without year*flag interactions (**Figure 14**). Reject H0.

Conclusions

The working group determined that combining the individual CPC datasets into combined indices was not recommended at the present time. The decision to not combine datasets for indices was based upon the failure to reject the three null hypotheses regarding the significance and trend in year*flag interactions. In both cases, the non-random year*flag interaction could not be explained, nor could they simply be modeled as random effects. Nonetheless, indices constructed with and without year*flag interactions were very similar, largely due to the relatively limited impact of these interactions on the overall model fit (% deviance reduction between 1.6-3.8 %). Further, the 'Flag' effects, or differences in the absolute catch rate for the different fleets were not significant, indicating that once gear, area, month, effort and other factors were considered, the models predicted similar catch rates even for very different fleets.

The divergence between Canada and Japan and the United States warrants further exploration, particularly because this seemed to occur in a single year and may be indicative of some change in the data. Similarly, the divergence between the US and Mexico in the Gulf of Mexico warrants further exploration and may be due to differences in the availability of the fish by month rather than true differences between the fleets over time. In the Atlantic, the remaining year*flag interactions, after the year 2000 were not as severe and may actually reflect more year*area interactions as the fleets fish in slightly different spatial areas rather than true differences between the fleets. This is particularly true as year*area interactions reduced a larger fraction of the deviance even though models with year*area interactions did not converge due to missing year*area cells or no BFT positive catch in some cells. Further investigation of the sources of the year*flag interactions and continued work to determine whether their significance can be reduced by modeling year*area or month*area or Flag*month effects may allow for the creation of joint longline indices for some or all of the available flags in the future. However, until the non-random and severe year*flag interactions can be reconciled or their significance reduced, the group recommends not creating combined indices for the 2017 data preparatory meeting.

Nonetheless there are some important lessons learned that have bearing on both how standardized indices are created as well as the upcoming stock assessment. First the explicit hypothesis testing framework for evaluating year*factor interactions is particularly informative for any CPUE standardization situation. In any CPUE standardization it is critical to examine a) how well the year*factor estimates are determined b) their magnitude and c) the trend to determine whether the patterns are random or have systematic patterns. This is extremely important for determining whether the year*factor interactions need to be considered in models, whether they can be assumed to be random or whether they have trends or patterns that would necessitate alternative treatments. One such alternative treatment recommended by Campbell (2015) would be to weight year*area interactions by area. Alternatively, in situations where year*factor interactions show non-random patterns this may be indicative of a need to examine why these exist or whether the index should be split.

A second important lesson for CPUE standardizations is that there is a need to develop further the methodology for spatial gap filling to allow for year*factor interactions (Carruthers *et al.* 2010). This is particularly true as the year*area interactions explained substantially more of the model deviance than year*flag interactions, though the year*area interaction models demonstrated poor convergence. In situations where there are strong year*area interactions, for example a shifting of the population in time and space (Schirripa *et al.* 2016) or when a fleet shifts location, vacating former fishing grounds it is essential that some means of spatial gap filling (Walters 2006, Carruthers *et al.* 2010, Carruthers *et al.* 2011) be conducted. This is necessary partly as a modeling necessity to allow models with year*factor interactions to converge and provide well determined model coefficients (Campbell 2015) but, additionally, to be able to accurately evaluate population trends.

Finally, this modeling has indicated a potentially important trend in year*area interactions that may have bearing on the assessment. For both indices year*area interactions explained substantially greater deviance than year*flag interactions indicating that there may be changes in local abundance of the population in different spatial locations over time. This may be reflective of changing spatial distribution of bluefin tuna over time due warming temperatures and a moving prey base, a hypothesis postulated by Golet *et al.* (2013) and Mackenzie *et al.* (2014) and also postulated for swordfish by Schirripa *et al.* (2016).

While the ultimate decision of the group was not to combine data to produce indices, the meeting was successful in combining the datasets to conduct a joint statistical analysis. These results do not preclude individual CPCs from bi-laterally deciding to combine datasets to create joint indices for consideration by the SCRS. The group re-iterates the previous caveat (Walter *et al.*, 2017) related to this combined dataset that it not be shared outside of the small working group and that it not be used for purposes outside of the working group terms of reference.

References

Carruthers, T. R., McAllister, M. K., Ahrens, R. (2010). Simulating spatial dynamics to evaluate methods of deriving abundance indices for tropical tunas. Canadian Journal of Fisheries and Aquatic Sciences 67: 1409-1427.

Carruthers, T. R., Ahrens, R., McAllister, M. K., Walters, C. J. (2011). Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices. Fisheries Research 109: 157-167.

Campbell, R. A. 2015. Constructing stock abundance indices from catch and effort data:Some nuts and bolts. Fisheries Research 161 (2015) 109-130.

Golet WJ, Galuardi B, Cooper AB, Lutcavage ME (2013) Changes in the Distribution of Atlantic Bluefin Tuna (*Thunnus thynnus*) in the Gulf of Maine 1979-2005. PLoS ONE 8(9): e75480.

Ichinokawa M., and J. Brodziak. 2010. Using adaptive area stratification to standardize catch rates with application to North Pacific swordfish (*Xiphias gladius*). Fisheries Research 106 (2010) 249–260.

Japan. 2012. National Report, 2010. IN Report for biennial period, 2010-11 PART II (2011) -Vol. 3.

Kimoto A., Takeuchi Y., and Itoh T. 2016. Updated standardized bluefin CPUE from the Japanese longline fishery in the Atlantic to 2015 fishing year. Col. Vol. Sci. Pap. ICCAT, 72(6): 1636-1655.

Lauretta M., Walter, J.F., Hanke A., Brown C., Andrushchenko I. and Kimoto A. 2016a. A method for combining indices of abundance across fleets that allow for precision in the assignment of environmental covariates while maintaining confidentiality of spatial and temporal information provided by CPCs. Col. Vol. Sci. Pap. ICCAT, 78 (8): 2318-2327.

Lauretta M., Hanke A., and Andrushchenko I. 2016b. An index of abundance of bluefin tuna in the Northwest Atlantic Ocean from combined Canada-U.S. pelagic longline data. Col. Vol. Sci. Pap. ICCAT, 72 (7): 1729-1747.

MacKenzie, B. R., Payne, M. R., Boje, J., Høyer, J. L. and Siegstad, H. (2014). A cascade of warming impacts brings bluefin tuna to Greenland waters. Glob Change Biol, 20: 2484–2491. doi:10.1111/gcb.12597

Ortiz, M and G. A. Diaz. 2004. Standardized catch rates for yellowfin tuna (*Thunnus albacares*) from the U.S. pelagic longline fleet. Col. Vol. Sci. Pap. ICCAT, 56(2):660-675.

Walter, J.F, Lauretta M, Kimoto A, Hanke A., K Ramirez and G. Melvin. 2017. Progress report of the working group on multi-national pelagic longline index for western Atlantic bluefin tuna. Col. Vol. Sci. Pap. ICCAT. 73(7): 2299-2323.

Schirripa, M. J. *et al.* 2016. A hypothesis of a redistribution of North Atlantic swordfish based on changing ocean conditions. Deep Sea Research.

Walters, C. 2003. *Canadian Journal of Fisheries and Aquatic Sciences*, 2003, 60(12): 1433-1436, 10.1139/f03-152

Walter, J. and S. Cass-Calay. 2014. Preliminary update of catch rates of large bluefin tuna (*Thunnus Thynnus*) from the U.S. pelagic longline fishery (1987-2013) accounting for weak hook implementation. Col. Vol. Sci. Pap. ICCAT, 70(2): 646-653.

Table 1. Deviance table for Gulf of Mexico data describing the percent reduction in deviance obtained from testing all single factors, then entering this in the model and iteratively testing additional factors until no factors meet the 1% inclusion criterion.

	Df	Deviance	AIC	Percent reduction in deviance	
GOM_AREA	5	6257.7	19882	25.9743	
fMonth	6	8084.3	18490	10.3915	
fYear	22	8356.7	18233	3.2867	
DAY_NIGHT	1	8516.4	17972	1.5576	
SST	1	8565.3	18105	1.4228	
fHookDepth	7	8589.8	17931	0.6222	*
Hook_Type2	3	8604.2	17937	0.4562	*
BAIT	3	8627	17960	0.1924	*
Flag	1	8643.1	17972	0.0057	*
<none>		8643.6	17971	0	

* Note that successive addition of model factors was stopped when none explained greater than 1% of deviance so estimates that are less than 1% do not account for inclusion of the factor above it.

Table 2. Gulf of Mexico deviance table for Atlantic model test 1: Are year*flag interactions significant?

	Df	Deviance	AIC	Percent reduction in deviance
<none>	8701.5	17914	0	
fYear:GOM_AREA	104	8258.3	17679	5.093
Flag:fYear	21	8560.2	17815	1.623
fMonth:GOM_AREA	30	8121.2	17394	6.668

Table 3. Deviance table for Atlantic data describing the percent reduction in deviance obtained from testing all single factors, then entering this in the model and iteratively testing additional factors until no factors meet the 1% inclusion criterion.

	Df	Deviance	AIC	Percent reduction in deviance	
fSST	11	21519	106718	15.7637	
fMonth	11	22529	100594	11.167	
fFishYear	22	22719	103590	10.8071	
fLat	4	24519	99594	3.15	
fLon	6	24731	99158	1.7388	
fSeaDepth	12	24926	98987	0.7594	*
fHookDepth	7	25044	99096	0.2867	*
Flag	2	25079	99120	0.1498	*
fNight	1	25116	99155	0.0021	*
<none>	25116	99154	0		

* Note that successive addition of model factors was stopped when none explained greater than 1% of deviance so estimates that are less than 1% do not account for inclusion of the factor above it.

Table 4. Atlantic deviance table for Atlantic model test 1: Are year*flag interactions significant?

	Df	Deviance	AIC	Percent reduction in deviance
<none>		25133	99120	0
Flag:fFishYear	44	24170	98244	3.8342
fFishYear:fLat	85	22910	97067	8.844
fMonth:fLat	38	24267	98330	3.4457
fFishYear:fLon	132	23207	97458	7.6629
fMonth:fLon	65	23668	97785	5.8314

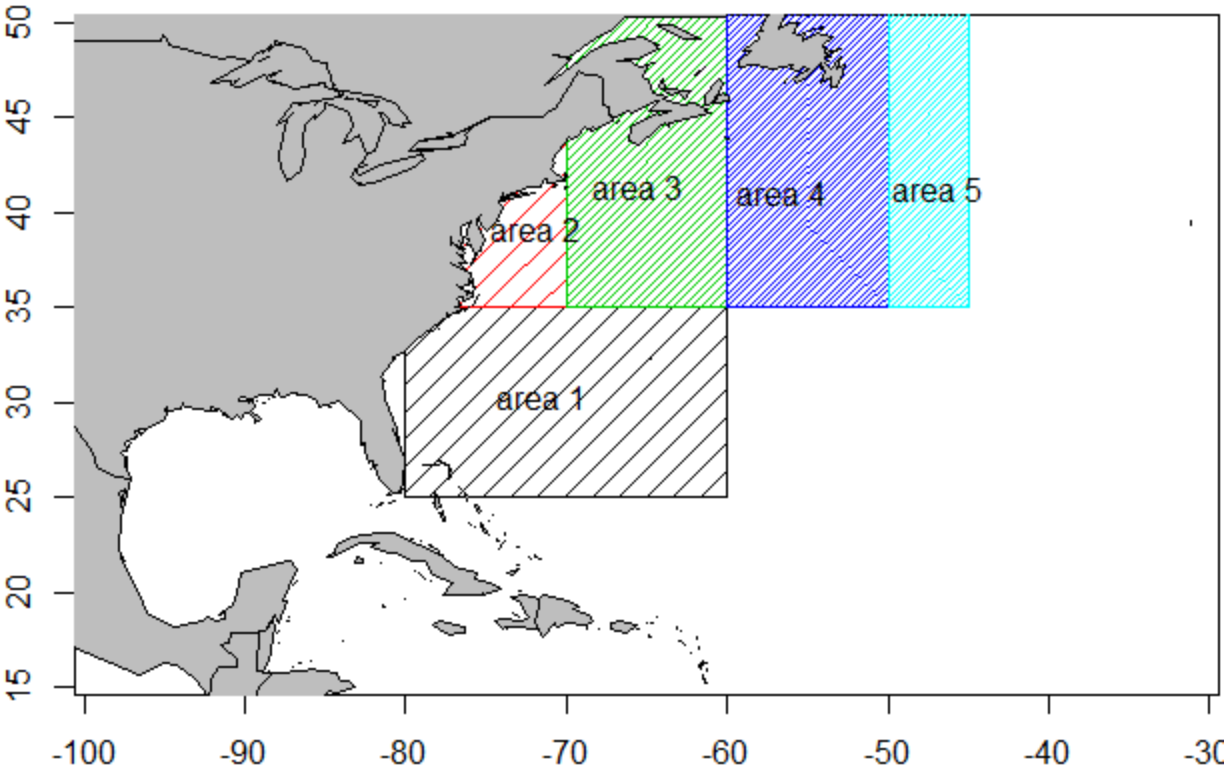


Figure 1. Western Atlantic area stratification.

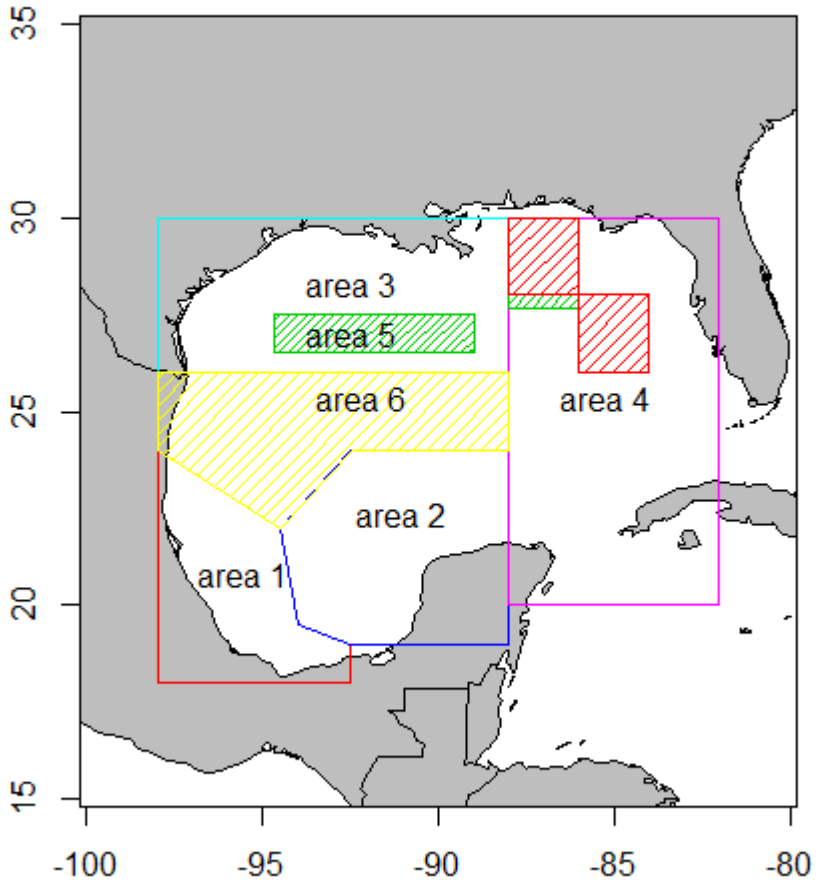


Figure 2. Spatial partitioning for Gulf of Mexico.

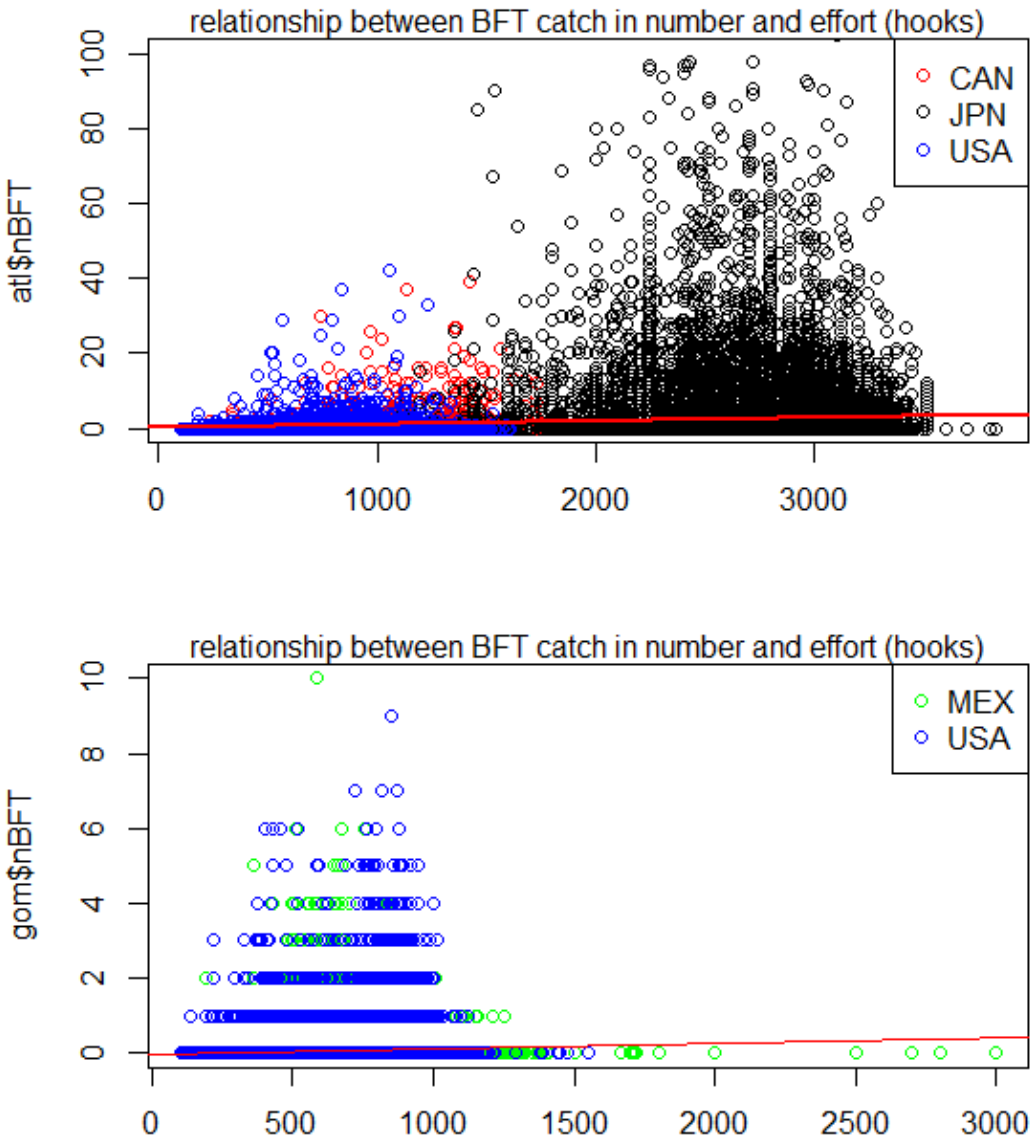
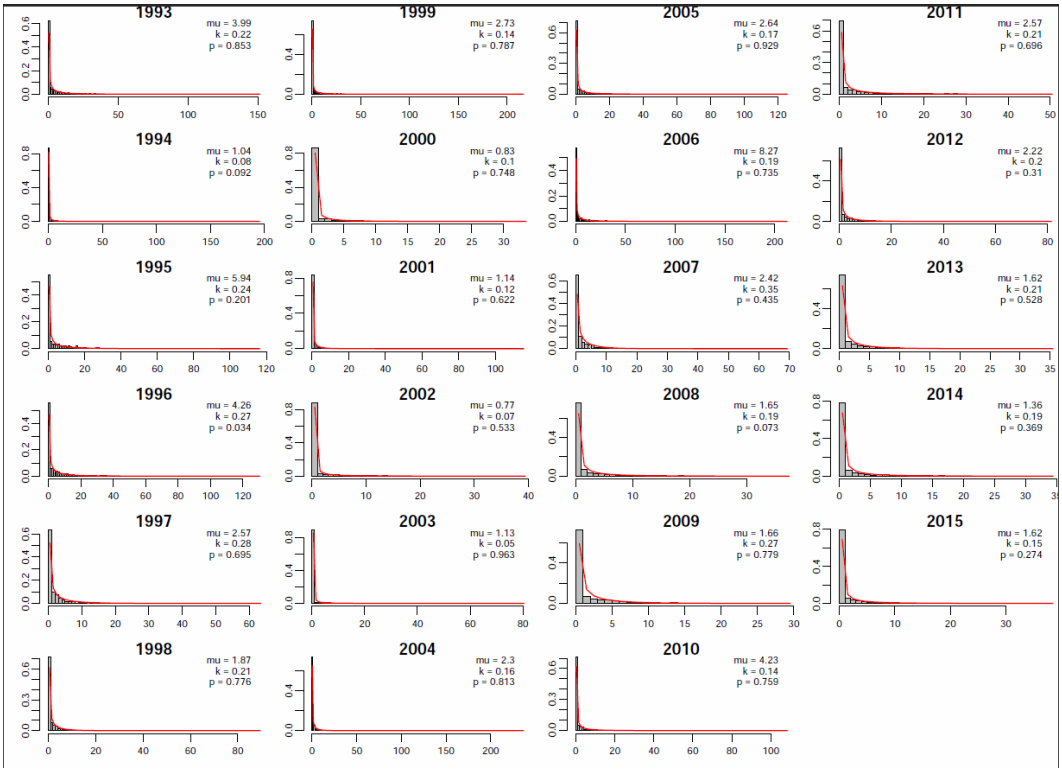


Figure 3. Relationship between BFT catch in number and effort (hooks).

A) ATLANTIC



B) GULF OF MEXICO

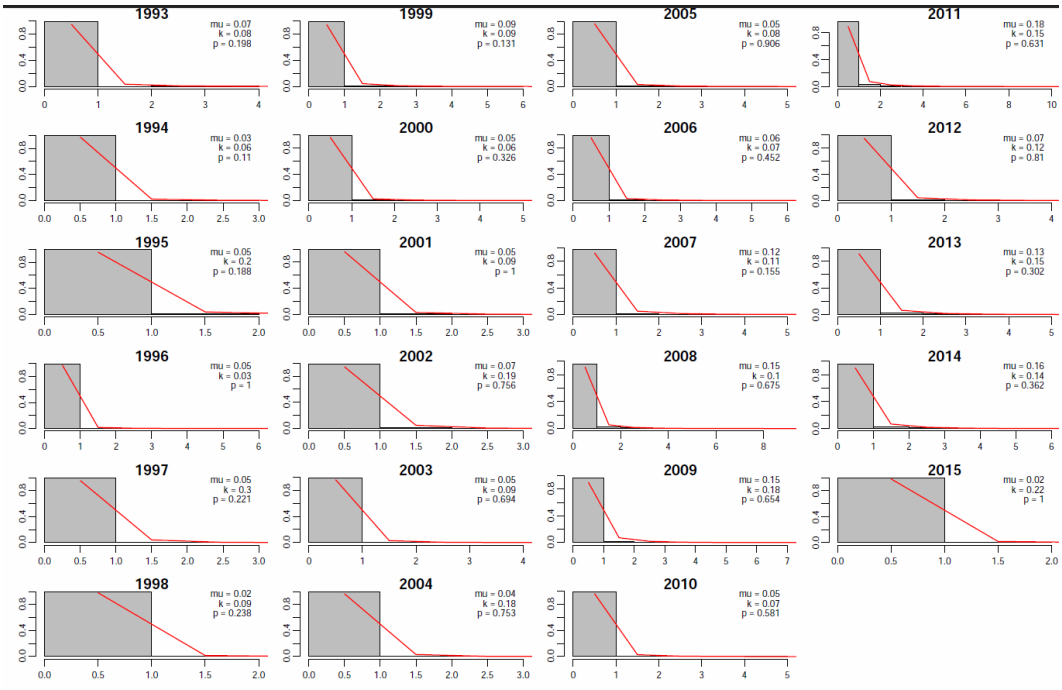
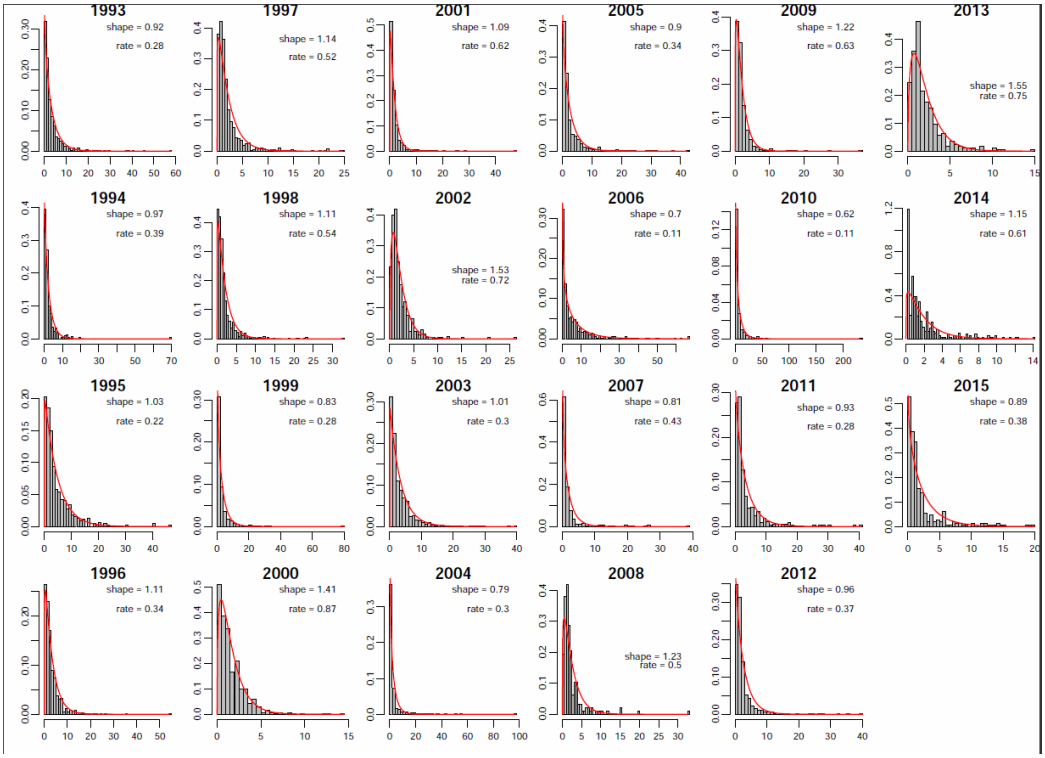


Figure 4. Negative binomial model fits to annual catches of BFT.

A) ATLANTIC



B) GULF OF MEXICO

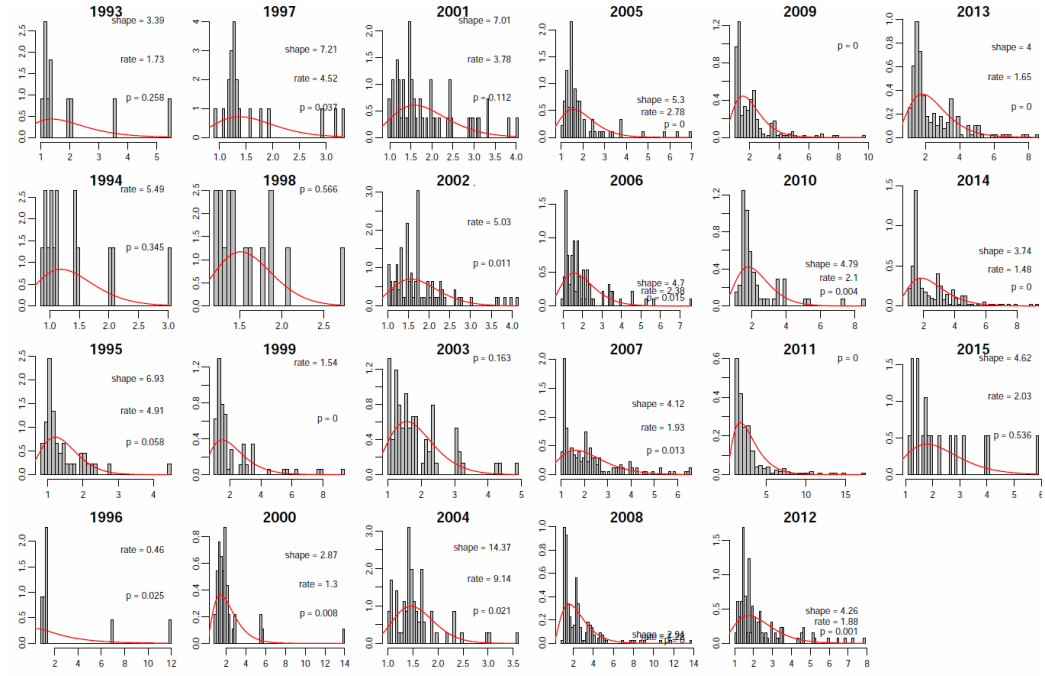
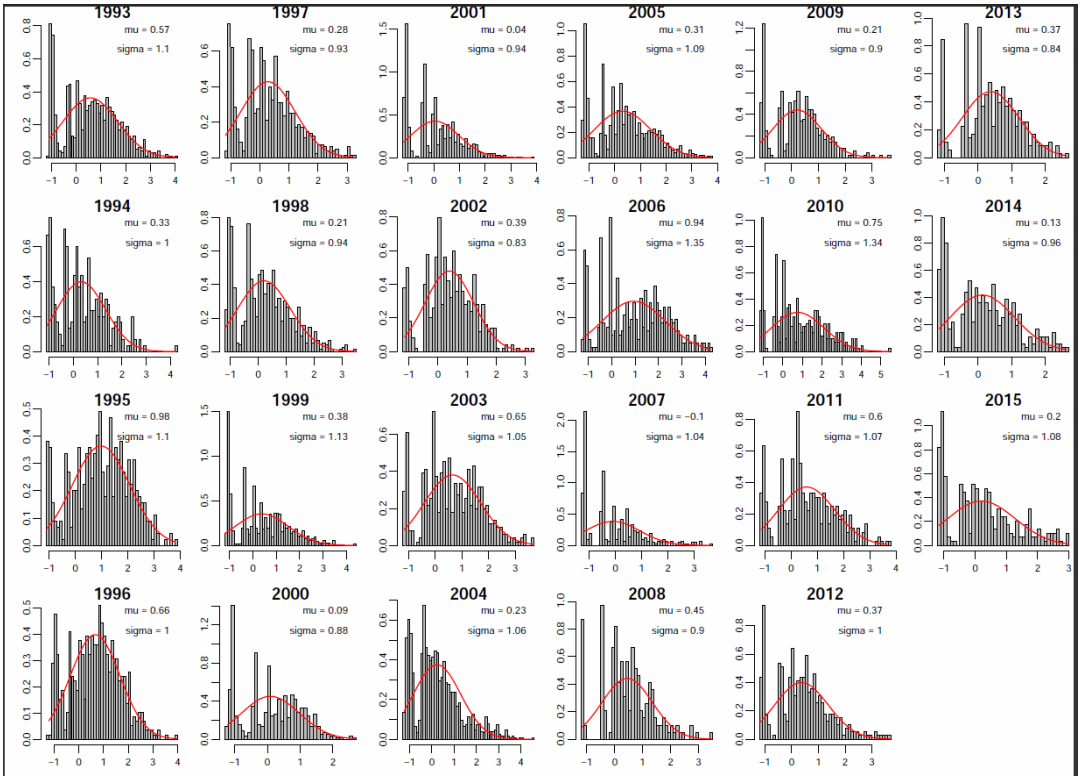


Figure 5. Gamma model fits to annual positive CPUE of BFT.

B) ATLANTIC



B) GULF OF MEXICO

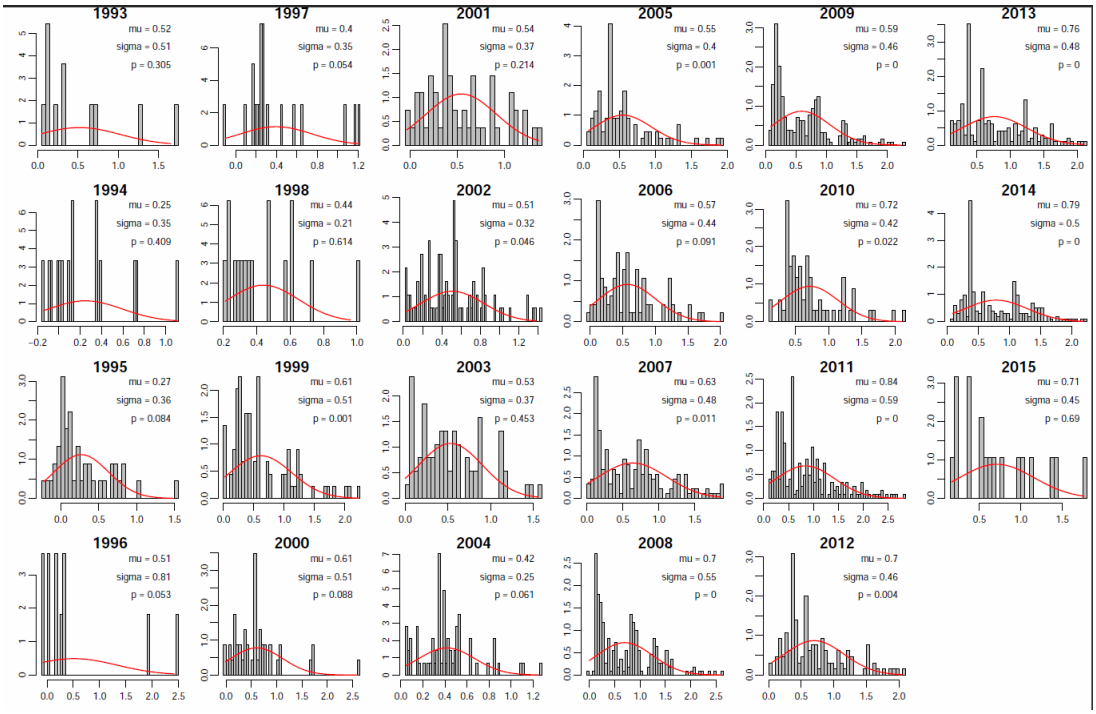


Figure 6. Gaussian model fits to annual $\log_e$-transformed positive CPUE of BFT.

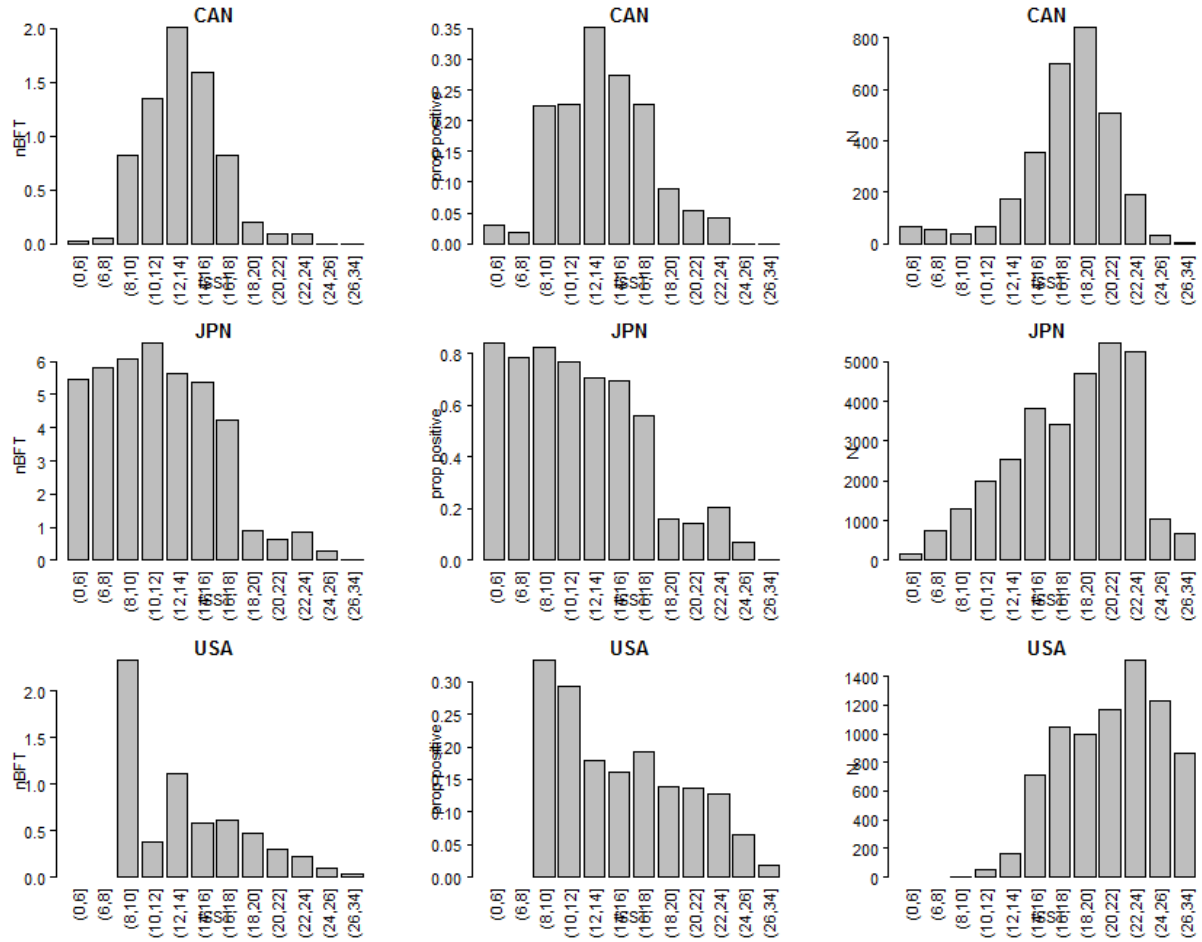


Figure 7. Plot of number of BFT vs SST, percent positive by SST and number of operations by SST for the Atlantic dataset.

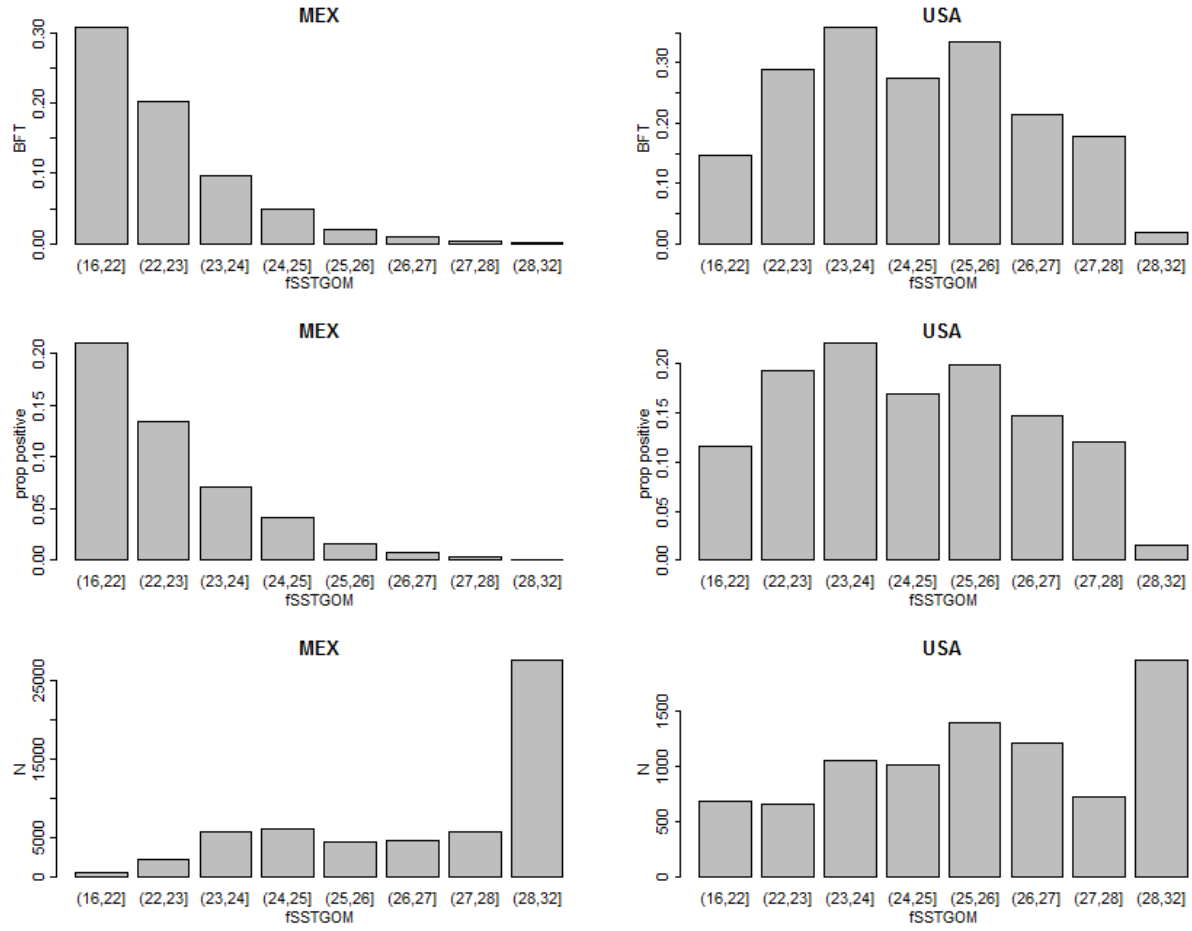


Figure 8. Plot of number of BFT vs SST, percent positive by SST and number of operations by SST for the Atlantic dataset.

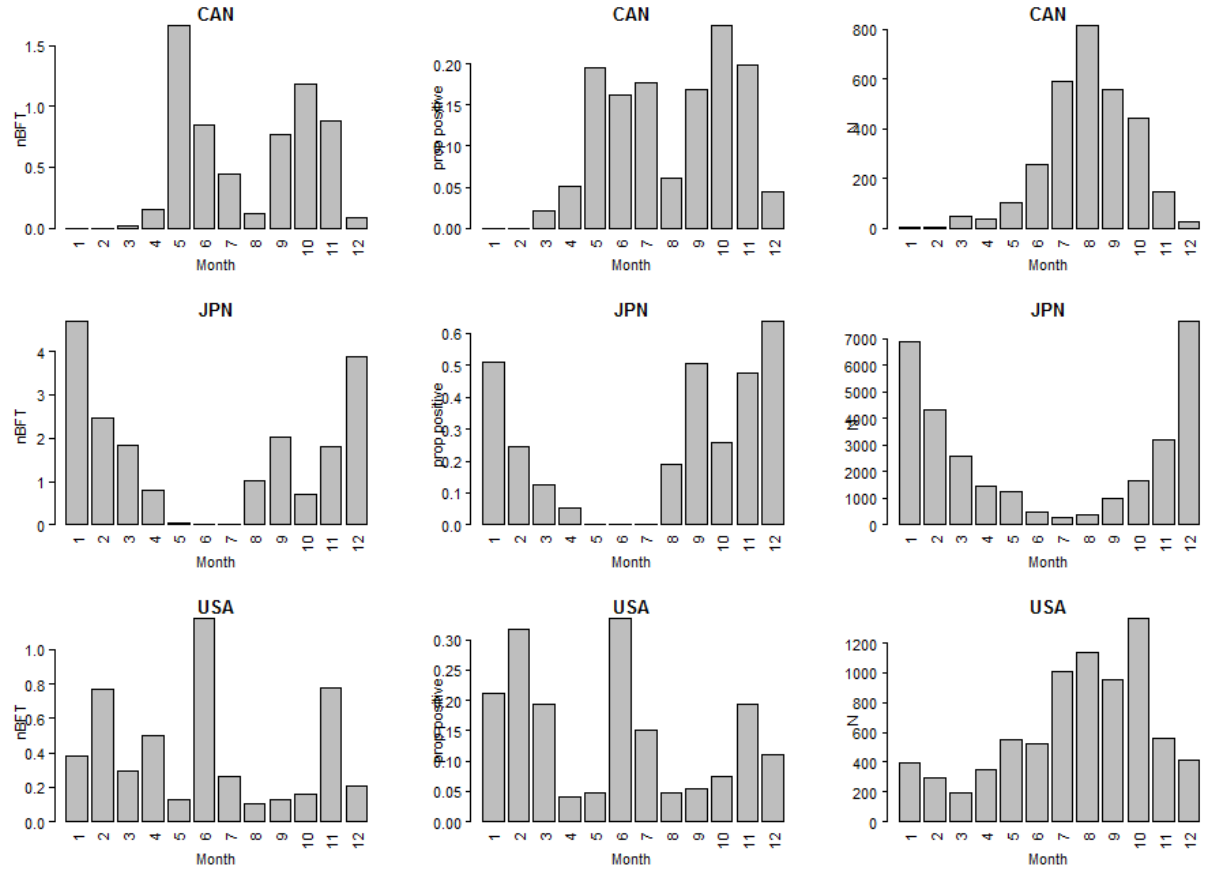


Figure 9. Plot of number of BFT vs month, percent positive by month and number of operations by month for the Atlantic dataset.

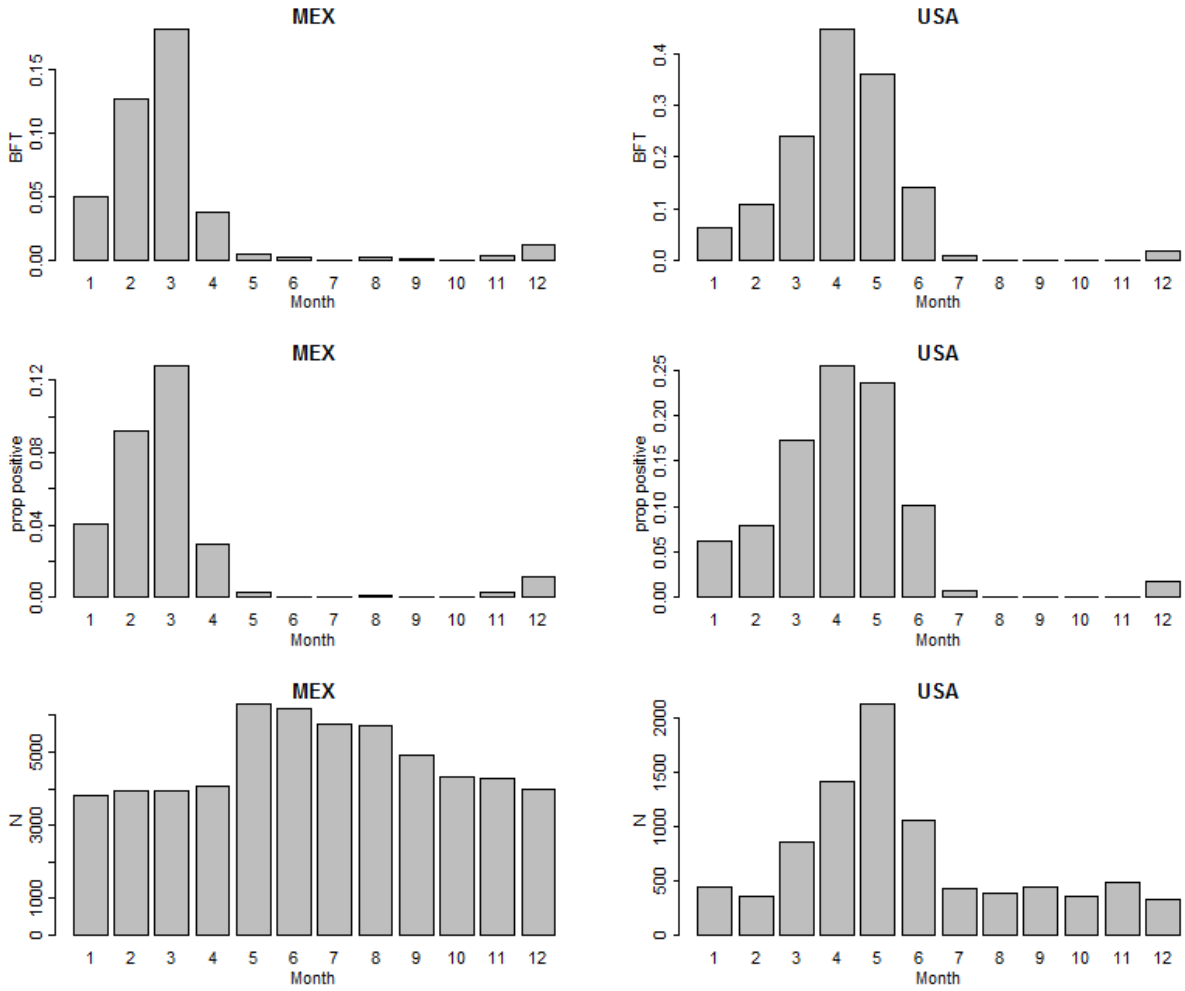


Figure 10. Plot of number of BFT vs month, percent positive by month and number of operations by month for the GOM dataset.

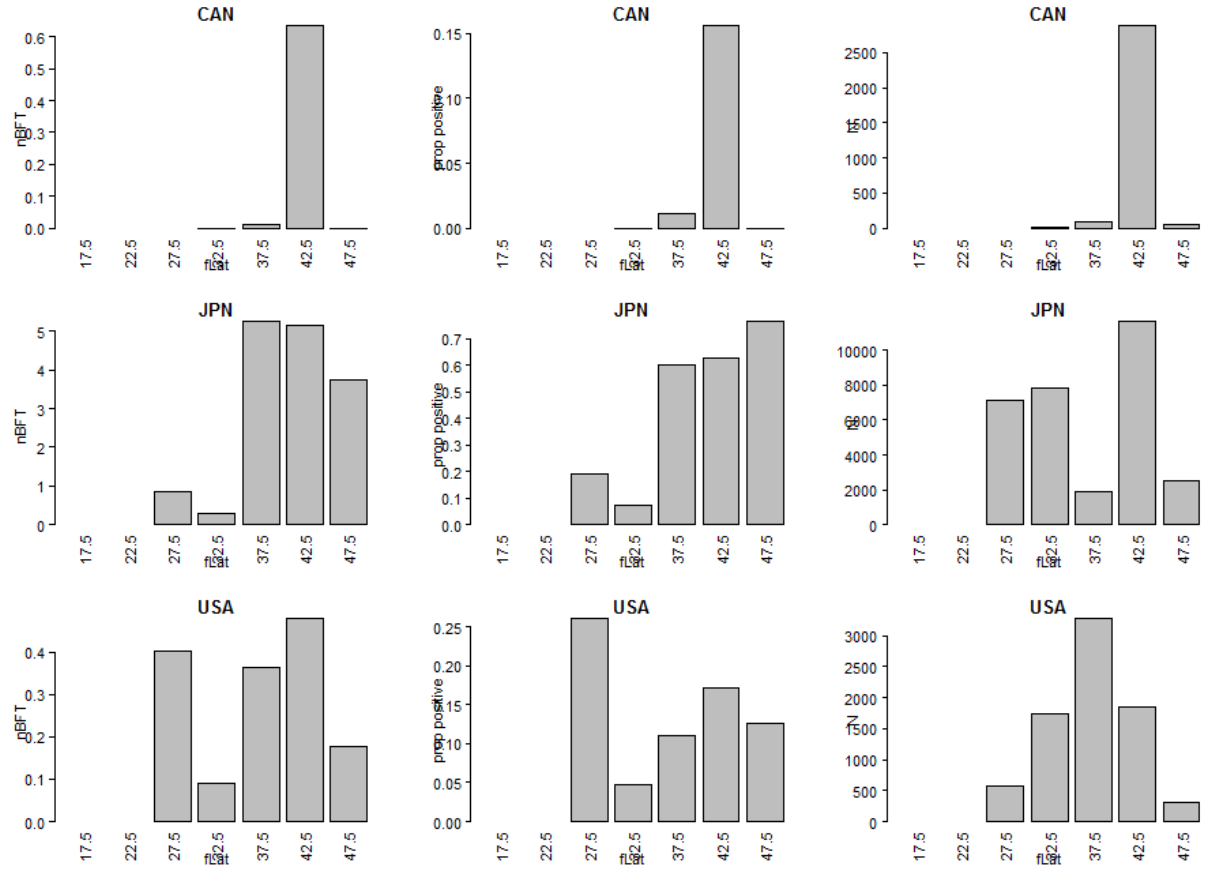


Figure 11. Plot of number of BFT vs latitude, percent positive by latitude and number of operations by latitude for the Atlantic dataset.

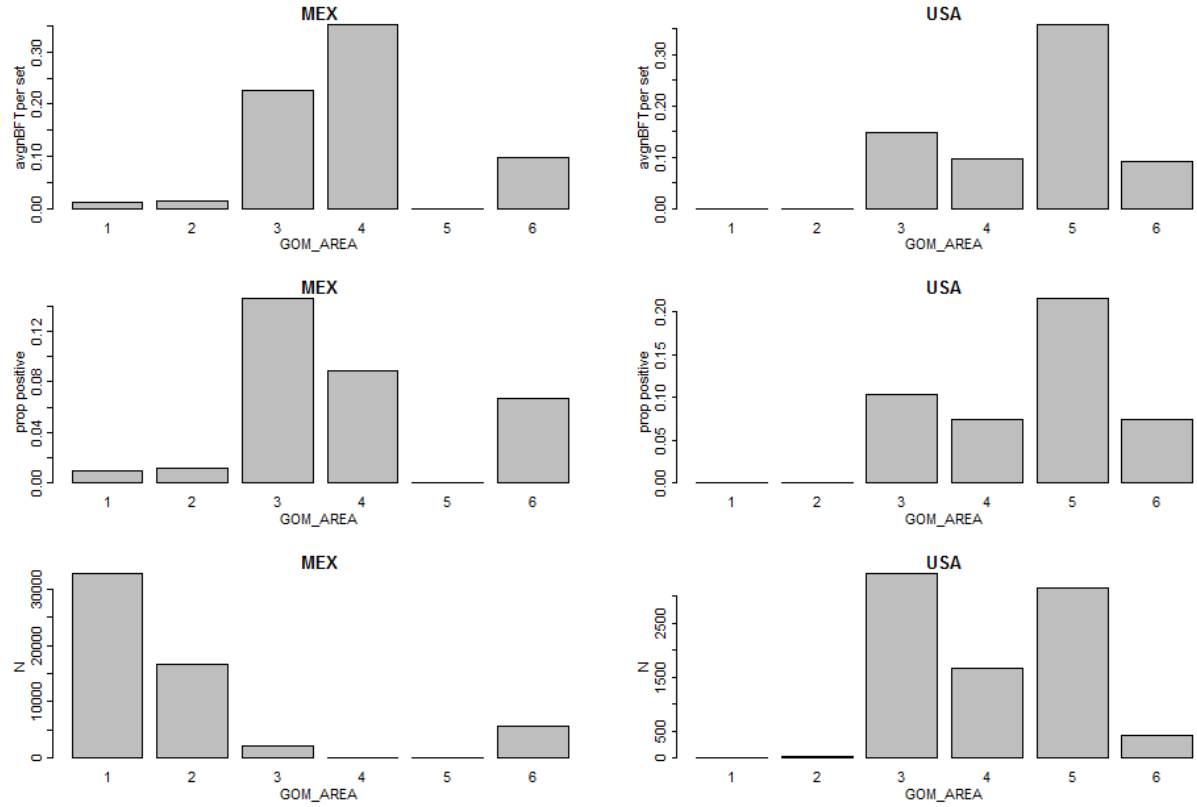


Figure 12. Plot of number of BFT, percent positive and number of operations by area for the GOM dataset.

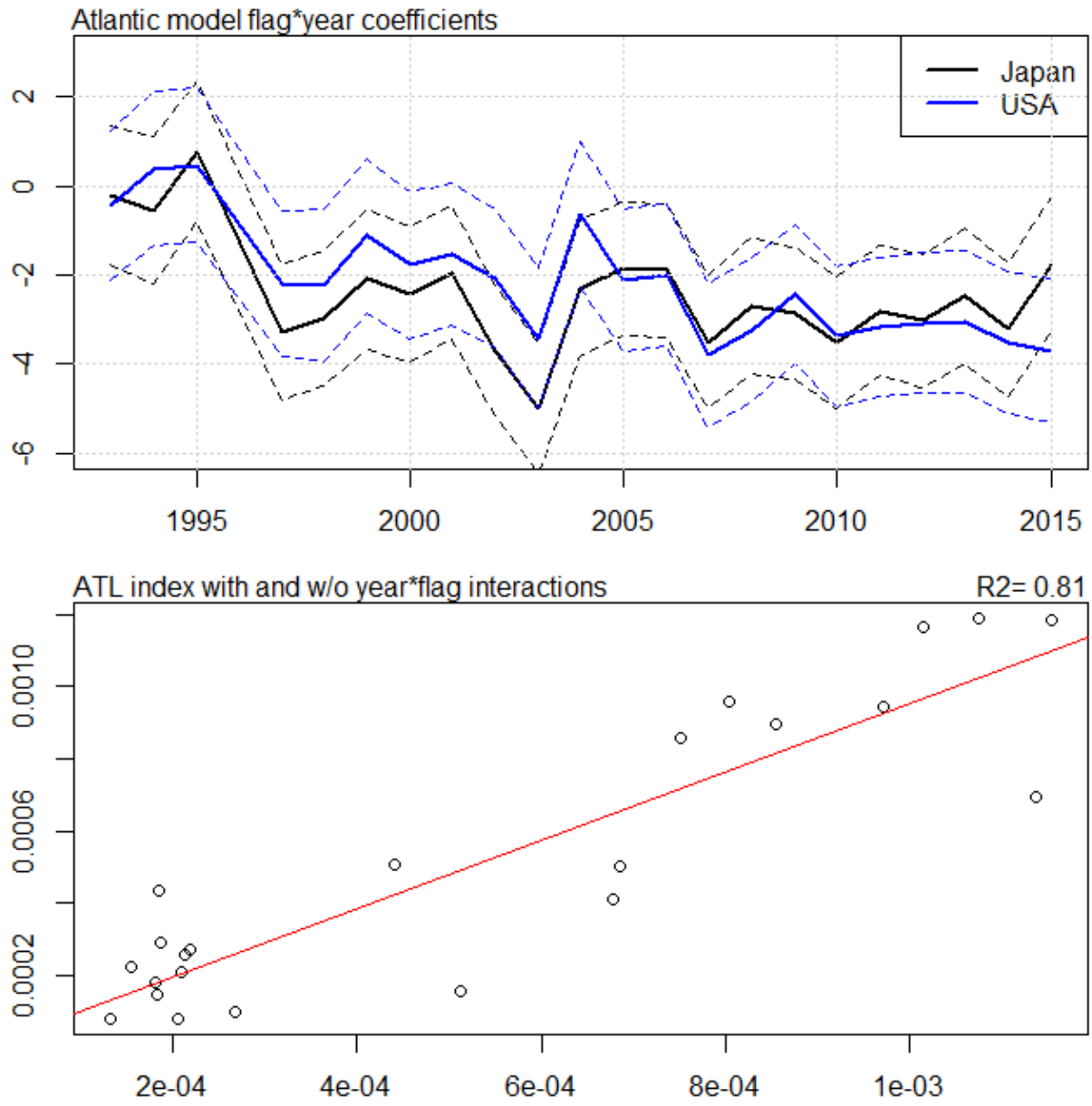


Figure 13. A. Year*flag interaction coefficients for the Atlantic index, noting that Canada is modeled as a baseline of 0, so Japan and USA coefficients are offsets from Canada. B. Correlation between indices constructed with and without year*flag interactions.

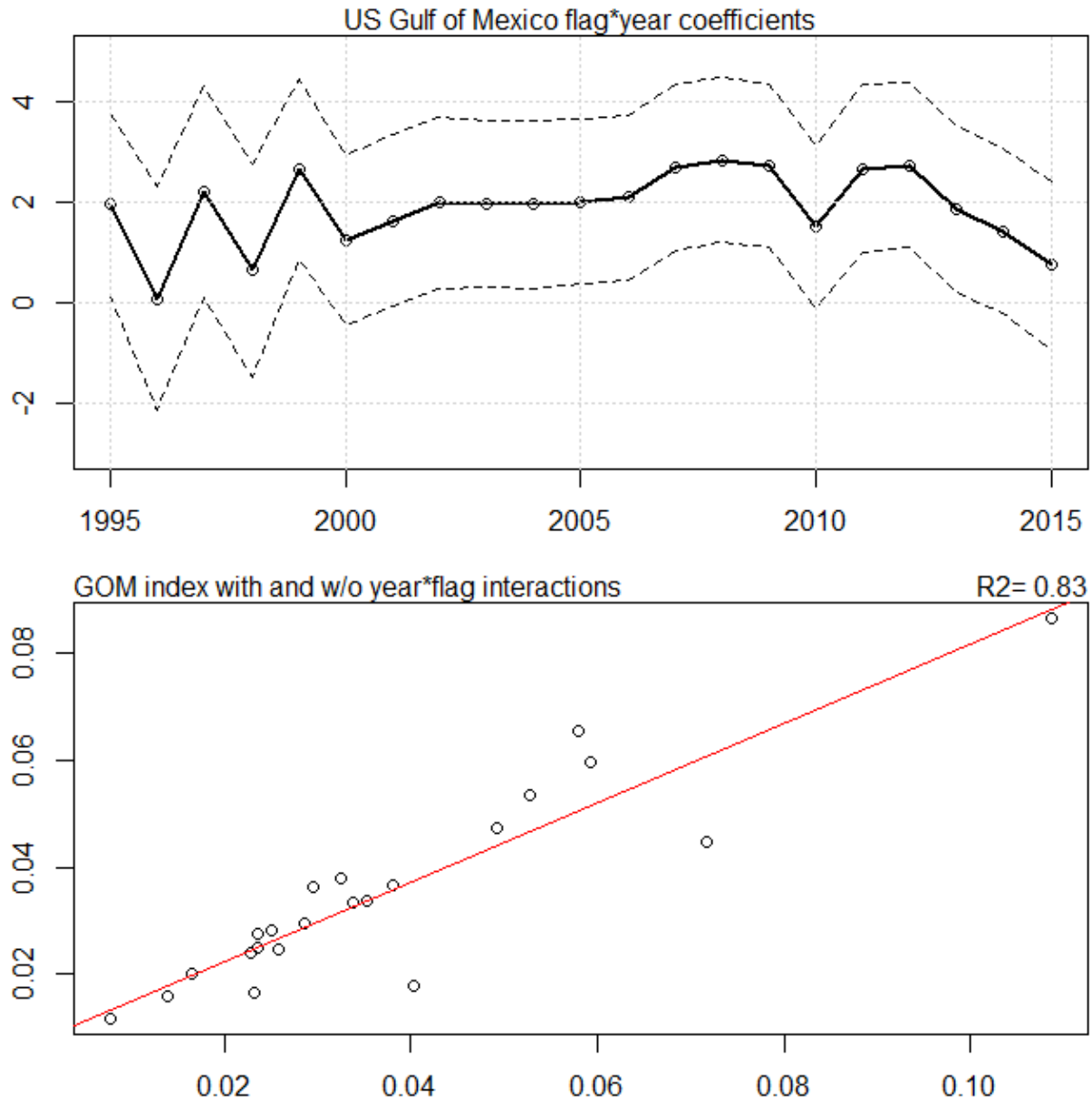


Figure 14. A. Year*flag interaction coefficients for the Gulf of Mexico index noting that Mexico is modeled as a baseline of 0, so USA coefficients are offsets from Mexico. B. Correlation between indices constructed with and without year*flag interactions.