Southwest Fisheries Science Center
Administrative Report H-00-06

ESTIMATION OF SEA TURTLE TAKE AND MORTALITY IN THE HAWAIIAN LONGLINE FISHERIES

Marti L. McCracken

Honolulu Laboratory
Southwest Fisheries Science Center
National Marine Fisheries Service, NOAA
2570 Dole Street, Honolulu, Hawaii 96822-2396

August 2000

NOT FOR PUBLICATION

1. INTRODUCTION

The objective of this report is to explain the statistical methodology used to estimate total turtle take and mortality for the Hawaiian longline fishery. Our target population, the population about which information is wanted, consists of all trips by the Hawaiian longline fisheries recorded in their logbooks. Although turtle takes are recorded in the logbooks, this information is considered unreliable. On the basis of biological opinions issued under the Endangered Species Act (ESA), an observer program was established in 1994. The observer program places trained observers on randomly selected trips of the Hawaiian longline fishery. Because a list of trips does not exist until the end of the year, trips are sampled such that the sample population, the population from which samples are drawn, coincides with the target population as closely as possible. During observed trips, the number of turtle takes by species, the condition of the turtles, other species of concern, and possible explanatory variables are recorded by the observer for each longline set. A turtle take is defined as an interaction between a turtle and the fishing vessel or gear, and usually implies that the turtle became entangled in the line or was caught on a hook. In this report, the observer and logbook records from 1994 through 1999 were used to estimate total turtle take and mortality by species by year. Since sampling did not begin until late February 1994, we extrapolated from our sampled population to predict take for the first couple of months in 1994.

Model-based predictors, instead of sample-based estimators, have been used to estimate total turtle takes by species since 1996. Whereas sample-based estimators assume sampling probabilities to raise observed total takes to the fleet level, model-based predictors assume a statistical model of turtle takes. Sample-based estimators have the advantage of being basically free of assumptions concerning the structure of the target population and the characteristics being estimated, but they are typically less efficient than model-based predictors. Because less than 5% of trips are sampled a year, model-based predictors have been used to estimate total take with the objective of gaining precision while maintaining acceptable accuracy.

This paper presents the methods used for modeling and prediction and discusses the results. The next section describes the data structure, and the methodology used for developing the prediction model is explained in Section 3. Section 4 describes how total takes and mortalities were estimated and prediction intervals approximated. The final prediction models and estimated total takes and mortalities are presented in Section 5.

2. Data structure

Data from the observer program are hierarchical, with trip as our independent observational unit and sets within a trip defined as subunits. Because of this structure, two types of stochastic dependence among sets from the same trip may exist. (1) Takes from sets within a trip may be more closely related than takes across trips. (2) Takes from sets close together in time and space within a trip may be more closely related. If take is modeled at the trip level, we could assume independent observations, but information is lost in the explanatory variables. For example, latitude and longitude are recorded for each set and these would need to be summarized if modeling at the trip level. Hence, modeling at the set level is preferred, but results can be misleading if the hierarchical structure of the data is ignored. Therefore, if explanatory variables do not explain the dependence among sets,

the dependence should be modeled. Trips by the same vessel may also be dependent, but without sufficient replication of vessels, we lack the information to model this dependence.

The 10 takes of green turtles observed during 1994-99 came from different trips; thus, there was no evidence within the data that a green turtle take in one set implies a higher probability of a green turtle take in another set from the same trip. Only four trips had more than one set where positive leatherback takes were observed, and only three trips had more than one set where positive olive ridley takes were observed. Out of 279 observed trips, 245 trips had zero leatherback takes and 254 trips had zero olive ridley takes; hence, there was little information or evidence in the data concerning dependency, if it exists. Instead of assuming a dependence structure that could not be checked, sets were treated as independent, but model diagnostics were used extensively to verify results.

For loggerheads turtles, 224 trips had zero observed takes, but out of the 55 trips with positive takes, 29 had positive takes in more than one set. If this dependence structure could be successfully modeled parametrically, we would expect a better predictive model, but further work is needed to determine if there is an 'appropriate' model. Therefore, loggerhead takes were modeled at the set level, but nonparametric resampling was used to model the dependence structure when estimating uncertainty.

Prediction models for turtle takes were constructed using observer data and corresponding logbook data. The response variable was the take recorded by the observer at the set level. This type of data is frequently referred to as count data; i.e., we counted the number of times an event, turtle take, occurred. Turtle take is a discrete variable, typically 0, 1, 2, or 3, with the event of a take being rare. Figures 1-4 show approximate positions of observed sets with positive takes represented by red circles. For all four species, the set-level data contained a very high percentage of zero takes: 99.7% for green turtles, 98.8% for leatherbacks, 99.1% for olive ridleys, and 96.2% for loggerheads. If the average count is sufficiently large so that counts, or a transformation of them, are approximately normal, robust modeling methods based on the normal distribution can be used. These methods are generally more straightforward theoretically. For average counts significantly below ten, results can be misleading if normality is assumed. Our average take per set was less than one and there was a high frequency of zeros; therefore, the distribution of takes should be treated as discrete.

Furthermore, since the take rate over the levels of our explanatory variables is commonly less than one, the calculation of degrees of freedom (d.f.) needs to be adjusted to reflect the effective degrees of freedom (McCullagh and Nelder, 1989) when relying on maximum likelihood asymptotic distributions for model fitting, hypothesis testing, and interval estimation. Degrees of freedom in this case refers to the parameter in the Student's t distribution, the asymptotic distribution of our maximum likelihood estimate. Degrees of freedom is usually calculated as the number of independent observations n minus the number of parameters being estimated p, but when an event is rare, this value is misleading and needs to be modified to improve the correspondence between the distribution of the test statistic and the asymptotic distribution. Assuming a Poisson distribution with constant mean, the effective degrees of freedom is given by $f = (n-p)/(1+.5\bar{\gamma}_2)$, where $\bar{\gamma}_2$ is the standardized fourth cumulant. Approximate confidence limits and hypothesis tests for maximum likelihood estimates should be based on the t_f distribution not t_{n-p}. To illustrate this concept, the effective

degrees of freedom for a sample of 1000 independent observations assuming a Poisson distribution with a constant rate of take is given in Figure 5. Because of the structure of the data being analyzed, we do not have a good estimate of effective degrees of freedom.

3. Developing models for turtle take

When developing a prediction model, only explanatory variables well represented in logbooks were considered. Table 1 lists the variables that were considered. Basically we want to use the simplest possible model that will estimate takes accurately and precisely. This means reaching a balance between including as few regressors as possible, so as to end up with the simplest working model and to control the variance of the predictions; and including as many regressors as possible, so as not to miss anything, to gain the best predictive power, and to avoid bias. Classification trees, generalized additive models (GAM), and generalized linear models (GLM) were used as exploratory tools in developing the final predictive models.

Many of the explanatory variables considered were related to each other, and if one of these variables was included in the model, including the other variables was redundant. To understand the relationships between explanatory variables, they were considered individually, in groups of related variables, and in subsets that showed possibilities for prediction. If a group of variables was associated with take, but there was dependency among them, careful consideration was taken not to be redundant and to select the best predictors. If a subset of these were not clearly superior but all were comparable, the decision was based on (1) minimizing measurement error, (2) the distribution of each variable around its mean since prediction error will increase as we move away from the average predictor values, and (3) the degree that the range of the logbook variable was covered in the observed sets since extrapolation is often less reliable than interpolation.

Because the units of measurement of different quantitative variables varied considerably, all quantitative variables were scaled according to the equation

$$x_i' = \frac{2x_i - max(\mathbf{x}) - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})},$$

where x_i represents the i^{th} observation of variable x and $\mathbf{x}$ represents the vector of observed values for x. This transformation should also improve the fit of the GAMs since the reliability of the GAM fit is always reduced at the endpoints of the range of the predictor (Hastie and Tibshirani 1990), and the transformation will bring in the tails of the distribution.

3.1. **Classification trees.**

Classification trees were used to explore variables associated with take and help determine possible categorical splits for continuous variables and pooling of categories for categorical variables. For example, for loggerhead takes, the classification tree suggested transforming the continuous variable sea surface temperature (sst) into a categorical variable with two categories, $sst < 23.77°C$ and $sst \geq 23.77°C$. Additionally, the trees suggested that months, numbered 1-12, be grouped into three categories: (1,2), (5,6), and (3,4,7-12).

Classification trees split the sample space based on the multinomial distribution. Turtle takes are not multinomial, but if the response variable is defined as 'take' and 'no take,' we have a binomial response, a special case of the multinomial distribution. Since the vast majority of observed takes were 0 or 1, little information was lost in using classification trees and the ability to detect associations was increased.

The tree functions in S-PLUS (Statistical Sciences, Inc., 1993) were used to grow and prune trees, see Venables and Ripley (1999) for details. At each step, the next split of a tree was chosen to gain the maximal amount of reduction in the deviance, a function of the conditional likelihood. A tree continued to grow until the number of cases reaching each leaf was small, $n_i < 10$, or further splits did not reduce the deviance significantly. The full tree overfitted the data, thus it was necessary to reduce the number of splits so that the tree described the important features of the data. To prune the tree, cross-validation was used to suggest the level of pruning. The basic idea of cross-validation was to divide the original data into ten mutually exclusive subsets. For each subset, a tree was grown using the data in the remaining subsets and pruned to various sizes. Each tree's fit was evaluated using the removed subset. The level of pruning was determined by comparing the average deviances measured at various tree sizes.

The cross-validation algorithm treated set as the independent unit, not trip. Because there was little evidence that sets within the same trip were correlated, we would not expect much of a difference in the results, but as a safety measure, trees were pruned to a range of sizes near the size suggested by cross-validation. Variables and subsets of variables that the classification trees suggested were associated with take were recorded and explored further using generalized additive models.

One of the major benefits of using trees for prediction is that explanatory variables with missing values can easily be included. If one of the variables with a reasonable number of missing values in the logbooks had appeared to be superior as a predictor, using trees for predicting total takes would have been given further consideration. This was not the case.

3.2. **Generalized Linear Models.** The splits of the continuous variables suggested by the trees were unsmooth predictors. For example, if one drew the probability of a loggerhead take for the two categories of *sst* suggested by the classification tree with sea surface temperature on the x-axis and the probability on the y-axis, there would be a break in the line at $sst = 23.77°C$. The next step was to investigate smooth forms of predictors using GLMs and GAMs and compare these to the unsmooth predictors.

A natural distribution to assume for counts of rare events is the Poisson distribution. For Poisson counts, the generalized linear model known as the log-linear model is applicable. The log-linear Poisson model assumes that all counts are independent and is formulated by assuming that the count Y_i is a Poisson random variable with mean μ_i and observed value $y_i = \mu_i + \epsilon_i$, where ϵ_i is the residual or 'error'. The mean is modeled as

$$log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

where $(x_{i1}, \ldots, x_{ip})$ are the explanatory variables and the βs are (usually) unknown parameters to be estimated. The right side of the equation is referred to as the linear predictor. The function that connects the linear predictor to the mean μ of Y, in this case the log function, is known as the link function. Typically maximum likelihood estimates (MLE) are used as parameter estimates and these are derived using iterative weighted least squares or the Newton-Raphson method.

Unlike the normal distribution that is completely determined by its two parameters, the mean μ and the variance σ^2, the relationship between the mean and the variance for a Poisson variate is a fixed relationship. Given a Poisson distribution with parameter λ, this

relationship is $\mu = \sigma^2 = \lambda$. Hence, determining the mean of the distribution implicitly determines its variance and there is no flexibility. This is in contrast to regression models where the response is assumed normally distributed and the variance parameter and mean can be estimated separately; thus, any constant degree of variability about the mean is accommodated in the fitted model. For these reasons, a Poisson distribution may seem suitable for a response, but when we inspect the model, the variance does not appear to equal the mean as assumed. If the variance is larger, we say that the data exhibit overdispersion; if it is smaller, we say the data exhibit underdispersion. Overdispersion is the more common and typically arises as a result of some sort of clustering or clumping process. With animals counts, we would expect the data to exhibit overdispersion if the animals tended to aggregate and underdispersion if the animals were territorial and spread very evenly over their range. Another situation where data will exhibit overdispersion is when subunits within a primary unit are correlated but primary units are independent. Failure to acknowledge overdispersion can lead to serious underestimation of standard errors and hence to misleading inferences about the form of the linear predictor. Because we are modeling at the set level, but our independent unit is the trip, it is important to consider overdispersion.

Various methods could be adopted for dealing with the problem, but one straightforward approach is quasi-likelihood estimation. Quasi-likelihood assumes a functional relationship between the mean and the variance; that is, it assumes that the variance has the form $Var(Y) = V(\mu)$ for some chosen function V. One commonly assumed form for Poisson data is $Var(Y) = \phi\mu$, where $\phi > 1$ represents overdispersion and $\phi < 1$ underdispersion. Even relatively substantial errors in the assumed functional form of $Var(Y)$ generally have only a small effect on the conclusion (McCullagh and Nelder, 1989). An advantage of quasi-likelihood estimation is that point estimates do not depend on the value of σ^2. Using quasi-likelihood methods, ϕ cancels out in our estimating equations, so that regression parameters remain the same as if ϕ were equal to 1. The parameter ϕ is typically called the dispersion parameter, and an estimate of ϕ can be used in estimating the standard errors for the $\hat{\beta}$s. However, quasi-likelihood estimators are not maximum likelihood estimators without the additional assumption that the responses are Poisson distributed. Nevertheless, McCullagh (1983) showed they have similar properties as MLE. Under quite general conditions, they are consistent and asymptotically normal. When quasi-likelihood estimators are not MLE, Cox (1983) and Firth (1987) showed that they still retain relatively high efficiency as long as the degree of overdispersion is moderate. However, for the estimation of ϕ, quasi-likelihood does not behave like a log-likelihood.

For Poisson data, a conventional estimate of ϕ is

$$\hat{\phi} = \frac{1}{n-p}\sum_i \frac{(y_i - \mu_i)^2}{\mu_i} = \frac{X^2}{(n-p)},$$

where X^2 is the generalized Pearson statistic. However, for the turtle data this estimator was unsatisfactory because of the number of μ_i very near 0. For example, there have been no observed loggerhead takes south of $22^\circ N$; all 1,263 sets observed below this latitude had zero takes. One could argue that the probability of a positive take was zero, and thus the number of takes observed in a set was not a Poisson variate but a fixed number, or that the probability was very small and the sample size was too small to expect a positive observed take. Because all observations are zero, the data provides us with no measure of variance, and if these numbers were included when modeling, the estimated dispersion parameter is

0.666. However, if these values were excluded, the estimated dispersion parameter is 1.015, and this still may be underestimating the dispersion. Although loggerhead takes provide the most dramatic example, the point estimates for ϕ showed similar problems for all species. Therefore, this estimate was not used for the dispersion parameter.

Also because several $\mu_i < 1$, the effective degrees of freedom were reduced. Thus, asymptotic distributions frequently assumed for MLE could not be automatically assumed. This affects interval estimation and sequential model fitting procedures when the methodology depends on the asymptotic distribution of MLE.

3.3. **Generalized additive models.** GAMs are a flexible extension of GLMs. Whereas GLMs restrict the parameters relating the regressors to the response to be of linear form, GAMs allow any shape ranging from a straight line to nonparametric curves of increasing complexity. Basically GAMs replace the GLM linear predictor, $\eta = \sum_{j=1}^{p} \beta_j x_j$, with a flexible additive function $\eta = \sum_{j=1}^{p} s_j(x_j, d_j)$, where $s_j(x_j)$ is an unspecified smooth function of x_j and d_j is the degrees of freedom of the smoother. GAMs still assume additivity of the effects of the x_js on the linear predictor scale, but allow the effect of x_j to follow any smooth curve. Categorical variables are permitted in GAMs and are expressed in the same manner as they are in GLMs. Similar to GLMs, the model specification is completed by a variance function, $Var(Y) = V(\mu)$. An advantage of GAMs is that they can be used to compare nonlinear smooth, linear smooth, and nonsmooth predictors.

The GAM is fitted by estimating the smoothing functions, the s_js, just as GLM is fitted by estimating the parameters, the β_js; however, GAM estimates are not maximum likelihood estimates and have not been shown to have their properties. Before the function s can be estimated, the required level of smoothing must be specified. The level of smoothing is determined by the value of the smoothing parameter what is commonly determined by the degrees of freedom specified for the smoother. The minimal amount of smoothing is just a straight line $s(x) = \beta x$ and the maximal is when $s(x)$ fits the data perfectly; the smoother goes through all points by allowing a separate gradient for each successive pair of points. At these two extremes, the GAM is equivalent to the GLM since it is possible to specify s in parametric form. Between these extremes, the function s is usually not specified but is estimated nonparametrically from the data by means of a scatterplot smoother. The shape of the function is therefore determined by the data rather than being restricted to a parametric form. As the degrees of freedom is increased, the function s gains flexibility and becomes 'rougher,' displaying more hills and valleys and more complex shapes. As with any model building, we want the simplest possible model that will achieve our required purpose. In the context of GAMS, we want the fewest degrees of freedom in the smoothers that will achieve our modeling objectives.

Several options are available for the scatterplot smoother. Smoothing splines fit the data using piecewise cubic polynomials. Models are fitted by satisfying a penalized least squares criterion. Compared to another common smoother, locally weighted regression (loess), the theoretical and numerical behavior of smoothing splines are cleaner. When modeling turtle take there was little difference between the fits when using the different smoothers so smoothing splines were used.

3.4. **Fitting and selecting the models.** All models in this analysis were fitted using the statistical software S-PLUS (Statistical Sciences, Inc., 1993). For GLMs and GAMs, a log-link and quasi-likelihood estimation were specified. Within the GAM framework and for each

continuous explanatory variable, expressing the relationship between take and the variable as a parametric curve (straight line or a polynomial of degree two) or a nonparametric curve was compared. If the classification tree suggested transforming the continuous variable into a categorical variable, this transformation was compared to the smooth curves.

To develop a subset of plausible linear predictors, stepwise selection based on the generalized information criterion was used. The generalized information criterion in this context is defined as

$$GIC = D - \alpha p\hat{\phi},$$

where D is the deviance, p is the number of parameters in the model, and α is either constant or a function of n. Two of the most commonly used forms of the GIC are Akaike's information criterion (AIC) where $\alpha = 2$ and Bayesian information criterion (BIC) where $\alpha = log(n)$. S-PLUS stepwise selection within the GAM framework is based on AIC. The GIC, including AIC, assumes MLE and independent observations and is a function of the dispersion parameter. However: (1) GAMs are not MLEs, (2) Assuming that sets within trips are independent is questionable, especially for loggerheads, (3) We do not have a good estimate of the dispersion parameter, and (4) AIC may not be the optimal form of the GIC for our data. The GIC with $\alpha = 2$ (AIC) leads to a ranking of the model in order of the estimated mean square error of prediction, and under many situations, 2 is the optimum value for α. But as n increases there will eventually be a point where larger values of α are optimum (Atkinson 1980). Also, larger values of α are indicated if the prediction problem is ill conditioned, the matrix X^TX nearly singular (Atkinson 1980). Along these same lines, as $\mu \to 0$ there is likely a point where larger values for α are optimum since $\mu = 0$ is not estimable. When initially fitting the models and using AIC ($\alpha = 2$) and assuming a Poisson family ($\phi = 1$), model diagnostics indicated that models were being overfitted. Because we have a large n, over 3,000, and because a large proportion of these are zero, a larger value of α is likely required for optimum model selection. Additionally, with such a high level of collinearity among explanatory variables, X^TX can quickly become ill conditioned. BIC adjusts for the sample size, and for the analyses in this paper, BIC is the GIC with $\alpha = log(3107) \cong 8$. However, even BIC in this circumstance may not be optimum. Atkinson (1981) suggests that the range α =2 to 6 may provide a set of plausible initial models for further analysis, but since $\alpha = 8$ is BIC, we extended this range.

For these four reasons, stepwise selection within S-PLUS was used only as an aid in determining the final predictive model. Since we do not have a good estimate of ϕ and a higher value for α than 2 is likely optimum, the dispersion parameter was specified over a range of values from 1 to 10. The higher values for the dispersion parameter place a higher penalty on adding a regressor, so we were adjusting for α as well as overdispersion. When using stepwise selection, we recorded the final model selected and models with similar GIC values. Under this framework, stepwise selection provided a useful tool.

For $\phi = 4$ to $\phi = 10$, the variables selected by stepwise selection were similar to those selected using classification trees and cross-validation. The similarities between the two are an indication that by adjusting the dispersion parameter within the S-PLUS environment, GIC can be used as a tool for model selection. Furthermore, model diagnostics indicated that variables being selected were associated with take. When $\phi < 4$, model diagnostics indicated that many of the variables selected showed marginal if any association with take.

Variables were introduced into the GAMs in the same manner as with the classification trees, first in similar groups and then combined. Also, since selection was dependent

on the order regressors were introduced, the order was varied. For continuous variables, smoothers with different degrees of freedom as well as polynomials of order 2 were considered. Polynomials place parametric constraints on the shape, and because of the problems with collinearity in using higher order polynomials, only polynomials of order 2 were considered. Under the GAM environment, polynomials are fitted as if under the GLM framework; consequently, issues such as residual degrees of freedom, standard-error bands, and tests of significance are straightforward under a Poisson family. For nonparametric GAMs, we rely on approximations and heuristics (Chambers and Hastie, 1993).

Once the GAMs or GLMs were fitted, informal verification of the goodness of fit was obtained through plots of residuals and estimated standard errors. Figure 6 is an example of a diagnostic plot supplied in S-PLUS and used extensively in this analysis. The solid line on the figure is the estimated scaled latitude curve $s(lat, 4)$ with 4 d.f. fitted to the loggerhead take data; only latitudes greater than $22°N$ were included in the data. The dashed lines lie approximately two standard errors away from the central curve on either side and give a rough indication of the level of variability around the fitted curve. Even under Poisson assumptions and independence, the calculation of the standard errors in S-PLUS involves some crude approximations with unknown properties (Chambers and Hastie, 1993); hence, it is recommended that these bands only be used for diagnostic purposes. Also, these bands should not be interpreted as confidence bands as there has been no adjustment for bias and the normality assumption is questionable. In Figure 6, the standard error band displays a clear curvature that follows the curvature of the fitted curve but flares out near the endpoints of the latitude range. The fact that we cannot draw a horizontal line across the plot without going outside the band provides evidence that latitude was associated with take.

The black circles on Figure 6 represent partial deviance residuals; these are simply the fitted term plus the deviance residual (Chambers and Hastie, 1993). The deviance is a function of the differences in the log-likelihoods and provides a measure of discrepancy between the fitted model and the saturated model; i.e., between the model and the data. The deviance residual for observation y_i is the square root of its contribution to the overall deviance multiplied by $sign(y_i - \hat{\mu}_i)$; i.e., it is positive if y_i is greater than its fitted value $\hat{\mu}_i$ and negative if y_i is less than $\hat{\mu}_i$. For our data, zeros dominated the responses and $\hat{\mu}_i$ was consistently near zero; therefore, responses with 0 takes have negative deviance residuals and responses of 1 or greater have positive deviance residuals and greater absolute values. A satisfactory diagnostic plot typically has residuals distributed evenly and randomly above and below the fitted curve. Again, due to the nature of our data, the curve will be closest to the zero values, and symmetry is not expected since zero is a lower boundary for counts. A healthy fit is indicated by the positive residuals following the pattern of the curve; i.e., the distance between the curve and the positive residuals is not distinctively greater in one section of the curve than in another.

Parts of the plot where the standard error band is particularly wide suggest that there are problems with the fit. Poor fits are most likely to be caused by sparse data or by a choice of degrees of freedom of the smoother that is too low to give an adequate representation of the relationship. Sparse data can be detected using the rug-plot along the bottom of Figure 6. The rug-plot gives the frequency of the x-values (Chambers and Hastie, 1993); where the frequency is high the rug-plot is a solid block. In Figure 6, the rug-plot indicates there were few observations at the higher latitudes; hence, the standard error band there is very wide. The band widens at the lower range of latitudes because the reliability of a GAM fit is

always reduced at the endpoints of the range (Hastie and Tibshirani, 1990). If the diagnostic plot indicates that the data are too sparse for the fit to be satisfactory, an alternative model formulation based on other regressors might be considered.

For categorical variables an appropriate step function is produced. Figure 7 is the diagnostic plot for leatherback takes fitted to a model with four categories of latitude. The jittering of residuals at the base of the plot and around the horizontal lines result in solid bars because there are so many observations of the value zero; the width of the bars is proportional to the number of observations in the category. In this example, the fit of the second category is distinct from the others.

Interactions with nonparametric smooth terms are not fully supported in S-PLUS. One can model an interaction term two dimensionally, $s(x_2, x_2)$, using the locally weighted regression smoother, but, as mentioned in Section 3.3, the theoretical and numerical behavior of loess are less clear. In our case there was no advantage to using GAMs, and once this was determined, GLMs were used to fit the models. Likely two-way interactions were introduced into the models and the stepwise procedure based on the GIC criterion was used in the same manner as for GAMs. None of the interactions tested appeared to be associated with take.

4. Prediction

Once the predictive model was selected, the take for all unobserved sets was predicted by substituting in the values recorded in the logbook for all regressors in the model and the expected take was used as the point estimate,

$$\hat{Y}_i = \hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}),$$

where x_{ij} was the logbook value for regressor j on set i. The point estimate for total predicted take $\hat{Y}_+$ was the sum of predicted takes for unobserved sets added to the sum of takes for observed sets,

$$\hat{Y}_+ = \sum_{i=1}^{n} Y_i + \sum_{i=1}^{N-n} \hat{Y}_i,$$

where n was the number of observed sets, and N was the number of longline sets recorded in the logbooks.

4.1. **Approximating prediction intervals.** To measure the uncertainty in the point estimates, prediction intervals were approximated. These are similar to confidence intervals but wider. Confidence intervals measure the uncertainty in the parameter estimates and are used when interval estimates for the mean response $\beta\mathbf{x}$ are required. When model-based estimators are used to predict a total, we assume that the total is a random variable. Prediction intervals take into account both the inaccuracy of parameter estimates, and the random fluctuations in the unobserved takes around the mean, $\mu_i = \beta\mathbf{x}$. For example, if the average take of turtles per set is estimated as $\hat{\mu} = 1$. We do not really expect $\hat{\mu}$ to equal the true average, so we calculate an interval, known as a confidence interval, that is expected to contain the true value of μ. The probability that a confidence interval will enclose μ is called the confidence coefficient. Now suppose that the true value of $\mu = 1$. We do not expect the take to equal 1 for every set but expect some takes to equal 0, others 1, and some greater than 1: this is what is meant by the variation around the mean.

To be independent from questionable asymptotic distributions, prediction intervals were approximated using a bootstrapping algorithm suggested in Davison and Hinkley (1997). This resampling algorithm produces response variation in addition to variation in parameter estimates. For the green, leatherback, and olive ridley turtles a parametric bootstrap was used. There was no evidence that the counts were overdispersed or underdispersed for these species and therefore a Poisson model was assumed. The basic algorithm was as follows.

Algorithm:
For r=1,...,R,

1. created a bootstrap response y_i^* at $\mathbf{x_i}$ by

$$y_i^* = \hat{\mu}_i + \epsilon_i^* \sqrt{\hat{\mu}_i} (i = 1, ..., n),$$

   where ϵ_i^* were generated Poisson variates with mean $\hat{\mu}_i$.
2. refitted the model with the y_i^*s and computed predicted totals $\hat{\mu}_{yr}^*$ for $yr = 1994, \cdots, 1999$.
3. for each year, calculated the sum of generated observed y_{yr}^* and calculated the statistic

$$d_{yr}^* = \frac{y_{yr}^* - \hat{\mu}_{yr}^*}{\sqrt{\hat{\mu}_{yr}^*}}$$

   .

Finally, the R values of d_{yr}^* were ordered to give $d_{yr,(1)}^* \leq \cdots \leq d_{yr,(R)}^*$. The prediction limits were calculated as

$$(\hat{\mu}_{yr} + d_{yr,((R+1)*.025)} \sqrt{\hat{\mu}_{yr}}, \hat{\mu}_{yr} + d_{yr,((R+1)*.975)} \sqrt{\hat{\mu}_{yr}}).$$

For approximating intervals, it is recommended that $R \geq 999$; R=999 was used in all bootstrap approximations reported here.

For loggerheads, the data's error structure was modeled instead of assuming a Poisson distribution. The algorithm follows the one given above except for three changes.

1. ϵ_i was generated using Pearson residuals,

$$\frac{y - \mu}{\sqrt{\phi Var(\mu)}}.$$

   For each trip, the mean of the Pearson residuals was calculated; let ϵ_t denote this mean for trip t. Then, for all sets within a trip, the difference between the mean and the residual for that set was calculated; let ϵ_h denote this difference. Bootstrap samples ϵ_i^* were generated by drawing a random ϵ_t for each trip and adding a random ϵ_h for each set,

$$\epsilon_i^* = \epsilon_t^* + \epsilon_h^*.$$

2. $\sqrt{\hat{\mu}_i}$ was replaced with $\sqrt{\hat{\mu}_i \hat{\phi}}$ where $\hat{\phi} = 1.015$, as discussed in Section 3.2.
3. As suggested by Davison and Hinkley (1997), stratification of residuals was necessary to create homogenous groups. Through trial and error, only stratification of ϵ_t was found necessary.

One drawback of this algorithm was that y^* could take on negative and non-integer values. To fix this, y^* was rounded to the nearest appropriate value. The appropriateness of this algorithm was confirmed using the diagnostics suggested in Davison and Hinkley (1997).

We did not adjust the point estimates for bias, but the effect is implicitly adjusted for in the bootstrap distribution used above. This algorithm was tried on the data for the other turtle species, but was obviously performing poorly due to the extent of the zero takes present. Since there was no evidence of overdispersion, it was felt that the approximated prediction intervals calculated using the parametric algorithm were more accurate.

4.2. **Estimating mortality.** Mortality point estimates were calculated as described in Kleiber (1998). The information used to estimate the probability of a kill given a take is provided in Table 2. Total kill was estimated by multiplying this probability by the estimated number of takes. To calculate prediction intervals, the probability of dying p was treated as a binomial probability estimated from a sample size equal to the number of turtles n observed for that species. The bootstrapping algorithms for takes were used for mortality prediction intervals, but the y_i^*s and μ_i^*s were multiplied by $p_i^* = x_i^*/n$, where x_i^* was generated as a binomial variate with parameters n and $\hat{p}_i$. Due to the lack of information, these approximated prediction intervals do not take into account uncertainty in the estimates of the probability that given its condition a turtle will die as a result of the interaction.

4.3. **Adjusting for hardshells.** Observed takes of unidentified hardshell turtles, only 10 during 1994-1999, were allocated to loggerhead, olive ridley, or green turtle takes in the same proportion as observed takes of these species. Except for green turtles, the species they were assigned to was based on what species they were most likely to be. This was determined using the prediction models for each species. This was easily done by examining the sea surface temperature for the observation: olive ridley's takes were higher at warmer temperatures, greater than $23.77°C$, and loggerhead takes were higher at cooler temperatures, less than $24.22°C$. If the sea surface temperature was not a clear indicator, the observed latitude was used to determine the species since loggerhead takes were higher in the northern latitudes. The two observations where the choice between olive ridley or loggerhead was most ambiguous were split fractionally between the three species so that the desired proportions were acquired.

5. Results

Some of the variables considered and found to be associated with take were poorly represented in the logbooks from 1994 and 1999, and thus were not considered for prediction. The prediction models used are given in Table 3. The maximal and minimal latitudes and sea surface temperatures recorded in the observed sets were close to the extremes in the logbooks, and between these two extremes, the range of values were well represented in the observed sets. For loggerhead takes only sets with latitudes north of $22°N$ were included when fitting the final model. Observations below this latitude were all observations of zero take; $24.4°N$ was the southernmost location at which a positive take was recorded in the observered sets. Including observations below $22°N$ resulted in an unstable model and negative bias: the predicted take totals were 480, 390, 415, 350, 358, and 348 for years 1994-1999, respectively. The decision to truncate the data at $22°N$ was based on the stability of the model and where the fitted curves tended to 'flatten out' at $\hat{\mu} \approx 0$; with the Poisson model, $\hat{\mu}$ cannot equal zero.

For the green turtle, none of the possible predictors appeared to have a strong association with take, but this could be a result of the small sample size in relation to the rarity of a green turtle take and does not mean that no relationships exist. For the olive ridley,

loggerhead, and leatherback turtles, sea surface temperature, latitude, and the distance to the approximate $17°C$ and $19°C$ isotherms were associated with take, but there was a high degree of collinearity between these variables. The plots in Figures 8-11 relate take to the predictor variables. For the olive ridley turtle there was a clear distinction between the proportion of takes between the two categories of sea surface temperature, but over latitude, the pattern was less clear. For the loggerhead turtle where both latitude and sea surface temperature were in the model, there was a clump of positive observations at the higher latitudes and at these latitudes, the clump was located in the colder temperatures. When comparing loggerhead take with latitude versus the three classifications for month, there were fewer observed trips at the higher latitudes in months 5 and 6. Latitude was split into four categories for the prediction of leatherback take. The categories most southern and most northern had fewer observations but a higher proportion of positive takes. In the middle category representing the middle lower latitudes there was a high proportion of the observed sets but few positive takes. The middle category representing the middle higher latitudes also had a high proportion of the observed sets, but it clearly had a greater proportion of positive takes, but not as high as the two end categories.

Plots of annual take and kill for each species are given in Figures 12 to 15. The shape of each plot represents the shape of the smoothed 'de-studentized' bootstrap distribution for estimated take. The shape is mirrored around a line representing the prediction interval with the point estimate represented as a square. The shaded red area represents the estimated proportion of the distribution above the 'trigger level' (authorized level of take or kill for a given year), the proportion is given above each year's graph. These results are tabulated in Tables 4 to 7. The probabilities that total take for loggerhead and leatherback turtles exceeded the authorized incidental take in 1999 are very small. For olive ridley and green turtles, the probabilities are larger but below 50%. The probabilities that total mortality exceeded the authorized mortality are very small for the green and loggerhead turtles and well below 50% for the leatherback turtle. For the olive ridley turtle, the point estimate is above the trigger level, and the estimated probability is 66%. From 1995 to 1999, the estimates for kill and take of olive ridley turtles appear rather stable with only the slightest indication of an upward trend. The increased number of estimated leatherback takes in 1998 and 1999 can be explained numerically by the increased percentage of trips below $14.95°N$ latitude. From 1994 to 1997 there were from 409 to 610 sets in this area, in 1998 there was 2,186 sets, and in 1999 there were 1,758 sets. In almost all cases, point estimates of take or kill fell clearly within the prediction intervals for earlier and later years; thus, there is no evidence of a trend.

6. Discussion

The point estimates of olive ridley mortality for 1995-1999 were above the authorized take level for this species stipulated in the current Biological Opinion. Out of 32 observed olive ridley takes, 6 have been reported as dead with 3 of them in 1998 and 1 in 1999. Take estimates for olive ridleys are below the authorized level for 1998 and 1999, although the estimated probability that the take exceeded the level was 46% for 1999. All estimates for leatherback, loggerhead, and green turtles were below the authorized take level.

It cannot be stressed strongly enough that since the observer program was an observational study, only associations with take can be identified. It is not possible to make causality statements from this type of study. This fact must be kept in mind when interpreting the

models used to predict take. Furthermore, these models were created with the objective of predicting take; under different objectives, a different model may be preferred.

REFERENCES

Atkinson, A. C.
1980. A note of the generalized information criterion for choice of a model. Biometrika 67, 413-418.

Atkinson, A. C.
1981. Likelihood ratios, posterior odds and information criteria. J. Econometrics 16, 15-20.

Chambers, J. M., and T. J. Hastie.
1993. Statistical models in S. Chapman and Hall, New York.

Cox, D. R.
1983. Some remarks on over-dispersion. Biometrika 70, 269-74.

Davison, A. C. and Hinkley, D. V.
1997. Bootstrap methods and their application. Cambridge University Press, New York.

Firth, D.
1987. On the efficiency of quasi-likelihood estimation. Biometrika 74, 233-45.

Hastie, T. J., and R. J. Tibshirani.
1990. Generalized additive models. Chapman and Hall, New York.

Kleiber, P.
1998. Estimating annual takes and kills of sea turtles by the Hawaiian longline fishery, 1991-97, from observer program and logbook data. Honolulu Lab., Southwest Fish. Sci. Cent., Natl. Mar. Fish. Serv., NOAA, Honolulu, HI 96822-2396. Southwest Fish. Sci. Cent. Admin. Rep. H-98-08. 21pg.

McCullagh, P.
1983. Quasi-likelihood functions. Ann. Statist. 11: 59-67.

McCullagh, P., and J. A. Nelder.
1989. Generalized linear models, 2nd edition. Chapman and Hall, New York.

Statistical Sciences Inc.
1993. S-PLUS reference manual. StatSci, a division of Mathsoft, Inc., Seattle.

Venables, W. N. and Ripley, B. D.,
1999. Modern Applied Statistics with S-PLUS, 3rd Ed., Springer, New York.

TABLE 1. Explanatory variables considered for predicting total take

	Variable	Notes
Location in time and space	latitude (lat)	degrees north
	longitude (lon)	degrees east
	distance to $17°C$ isotherm	calculated from lat, lon, and sst
	distance to $19°C$ isotherm	calculated from lat, lon, and sst
	year	94-99, looked at pooling categories
	month	January-December, looked at pooling categories
	day	represented as a circular variable for a year using the cosine and sine functions
Condition of gear	hooks	
	hooks/float	
Environment	temperature	sea surface temperature (sst)
Catch of other species		total and proportion of total
	yellowfin	
	skipjack	
	albacore	
	swordfish	
	blue shark	
	mahimahi	
	striped marlin	
	blue marlin	
	wahoo	
	spearfish	
	opah	
	albatross	
Other	vessel length	registered length
	trip type	3 categories (swordfish,tuna,mixed)

TABLE 2. Observed turtle take and estimated probability (in parentheses) that a turtle was killed by the interaction with the fishery. Takes and estimates are given by condition and were calculated from the 1994-1999 observer data. The probability in the Total row was the probability used to estimate the total number of kills.

Condition	Loggerhead	Leatherback	Olive Ridley	Green
Hooked				
Released				
Internal	83	1	16	1
	(.29)	(.29)	(.29)	(.29)
External	56	23	10	8
	(0)	(0)	(0)	(0)
Unknown	3	4	0	0
	(.17)	(.01)	(.18)	(.03)
Dead	1	1	6	1
	(1)	(1)	(1)	(1)
Entangled				
Released				
Okay	3	3	0	0
	(0)	(0)	(0)	(0)
Injured	0	3	0	0
	(0)	(0)	(0)	(0)
Dead	0	2	0	0
	(1)	(1)	(1)	(1)
No record				
Released	1	3	0	0
	(.17)	(.01)	(.18)	(.03)
Total	147	40	32	10
	(.17)	(.08)	(.33)	(.13)

TABLE 3. Explanatory variables included in the prediction models

Species	Explanatory variables
Loggerhead	month in three categories: [1,2],[5,6],[3,4,7-12] latitude as a polynomial: $lat + lat^2$ sea surface temperature in two categories: $[sst \leq 23.77°C]$,$[sst > 23.77°C]$
Olive Ridley	sea surface temperature in two categories: $[sst \leq 24.22°C]$,$[sst > 24.22°C]$
Leatherback	latitude in four categories: $[lat \leq 14.95°N]$, $[14.95°N < lat \leq 24.84°N]$, $[24.84°N < lat \leq 33.82°N]$,$[lat > 33.82°N]$
Green	none

TABLE 4. Loggerhead take and kill estimates with 95% prediction intervals and probabilities that the real takes or kills exceeded the authorized levels. Authorized levels for take and kill for 1994-1997 were 305 and 46 and for 1998-1999 were 489 and 103.

	Takes			Kills		
Year	Est.	95 %PI	Prob.	Est.	95 %PI	Prob.
1994	501	[315-669]	.98	88	[36-141]	.95
1995	412	[244-543]	.90	72	[31-115]	.89
1996	445	[290-594]	.96	78	[34-127]	.92
1997	371	[236-482]	.82	65	[28-102]	.84
1998	407	[259-527]	.09	71	[32-112]	.06
1999	369	[234-466]	.01	64	[28-102]	.03

TABLE 5. Leatherback take and kill estimates with 95% prediction intervals and probabilities that the real takes or kills exceeded the authorized levels. Authorized levels for take and kill for 1994-1997 were 271 and 23 and for 1998-1999 were 244 and 19.

	Takes			Kills		
Year	Est.	95 %PI	Prob.	Est.	95 %PI	Prob.
1994	109	[68-153]	.00	9	[0-22]	.03
1995	99	[62-141]	.00	8	[0-21]	.02
1996	106	[69-148]	.00	9	[1-21]	.02
1997	88	[55-124]	.00	7	[0-18]	.00
1998	139	[79-209]	.00	12	[1-28]	.23
1999	132	[76-193]	.00	11	[1-27]	.20

TABLE 6. Olive ridley take and kill estimates with 95% prediction intervals and probabilities that the real takes or kills exceeded the authorized levels. Authorized levels for take and kill for 1994-1997 were 152 and 41 and for 1998-1999 were 168 and 46.

	Takes			Kills		
Year	Est.	95 %PI	Prob.	Est.	95 %PI	Prob.
1994	107	[70-156]	.04	36	[8-64]	.36
1995	143	[90-205]	.39	47	[7-84]	.63
1996	153	[103-210]	.53	51	[11-90]	.70
1997	154	[103-216]	.54	51	[8-92]	.70
1998	157	[102-221]	.38	52	[11-92]	.62
1999	164	[111-231]	.46	55	[11-96]	.66

TABLE 7. Green turtle take and kill estimates with 95% prediction intervals and probabilities that the real takes or kills exceeded the authorized levels. Authorized levels for take and kill for 1994-1997 were 119 and 18 and for 1998-1999 were 52 and 15.

	Takes			Kills		
Year	Est.	95 %PI	Prob.	Est.	95 %PI	Prob.
1994	37	[15-65]	.00	5	[0-16]	.01
1995	38	[15-70]	.00	5	[0-17]	.02
1996	40	[19-70]	.00	5	[1-17]	.02
1997	38	[14-73]	.00	5	[0-17]	.00
1998	42	[18-76]	.28	5	[1-19]	.06
1999	45	[18-82]	.34	6	[1-19]	.06

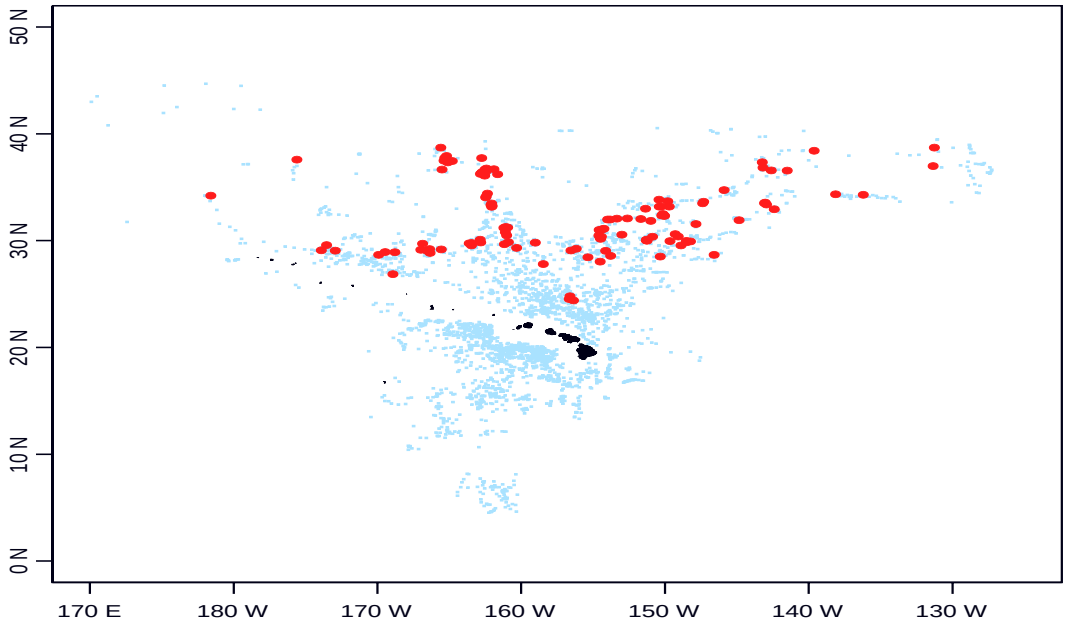

Figure 1. Approximate positions of observed sets with positive loggerhead takes represented as red dots and takes of zero represented as blue dots.

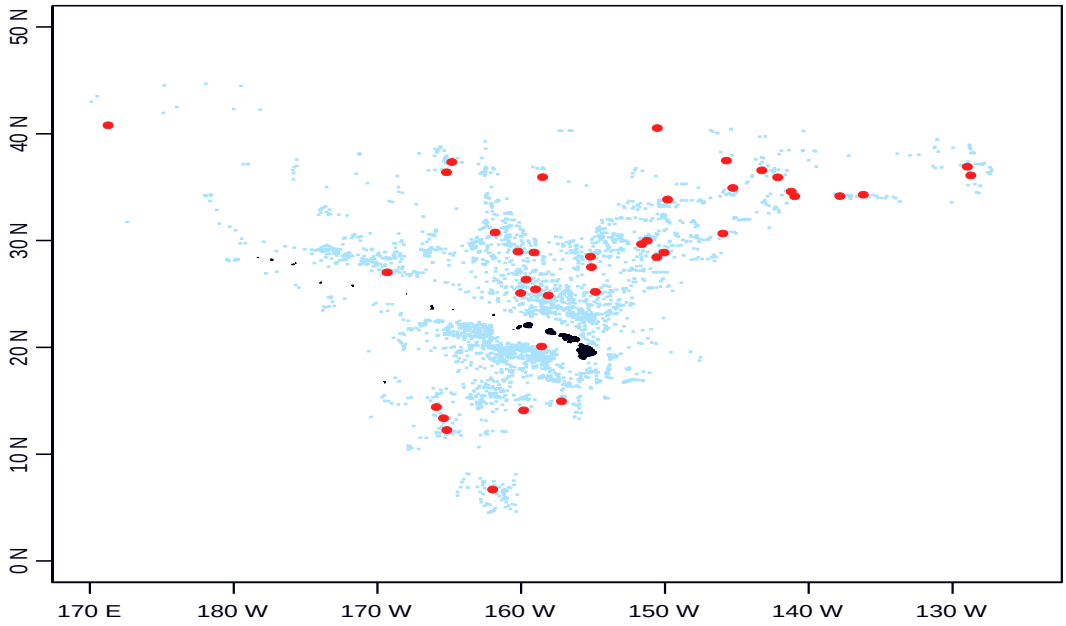

Figure 2. Approximate positions of observed sets with positive leatherback takes represented as red dots and takes of zero represented as blue dots.

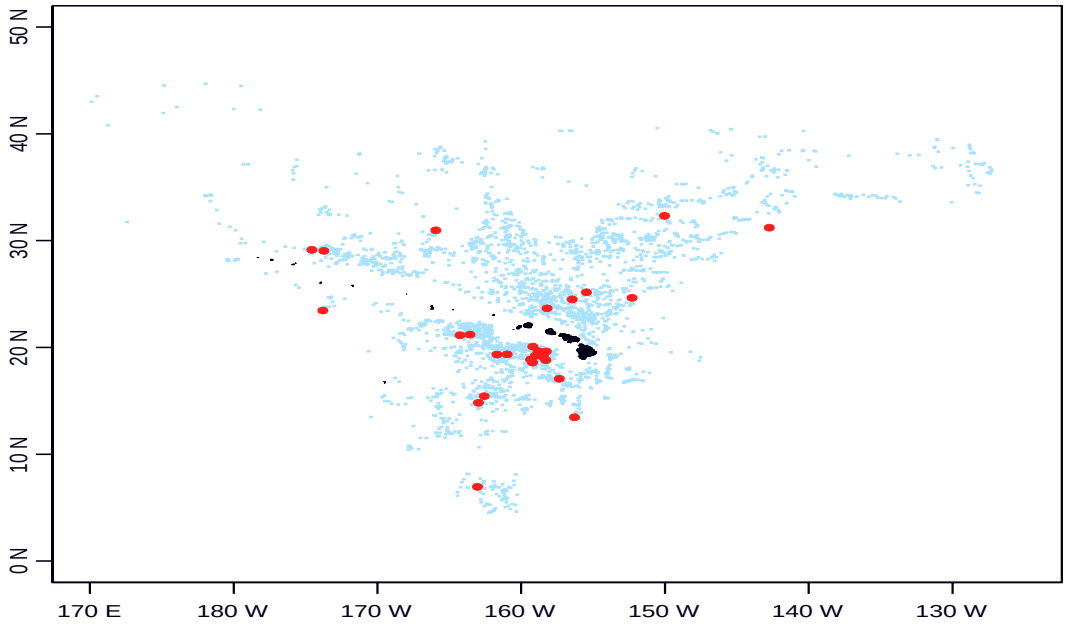

Figure 3. Approximate positions of observed sets with positive olive ridley takes represented as red dots and takes of zero represented as blue dots.

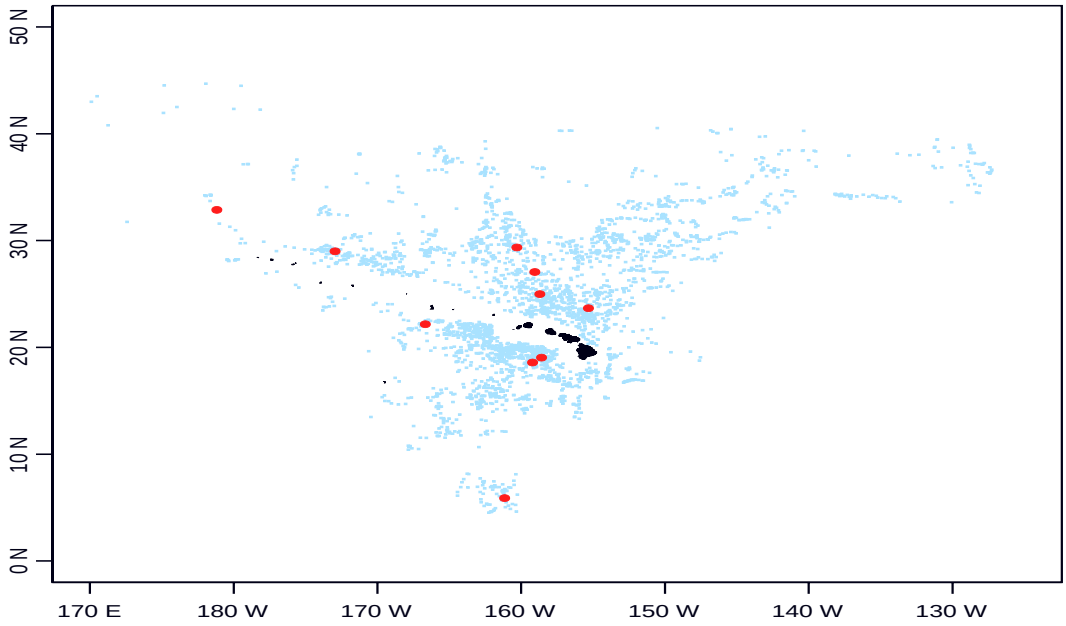

Figure 4. Approximate positions of observed sets with positive green turtle takes represented as red dots and takes of zero represented as blue dots.

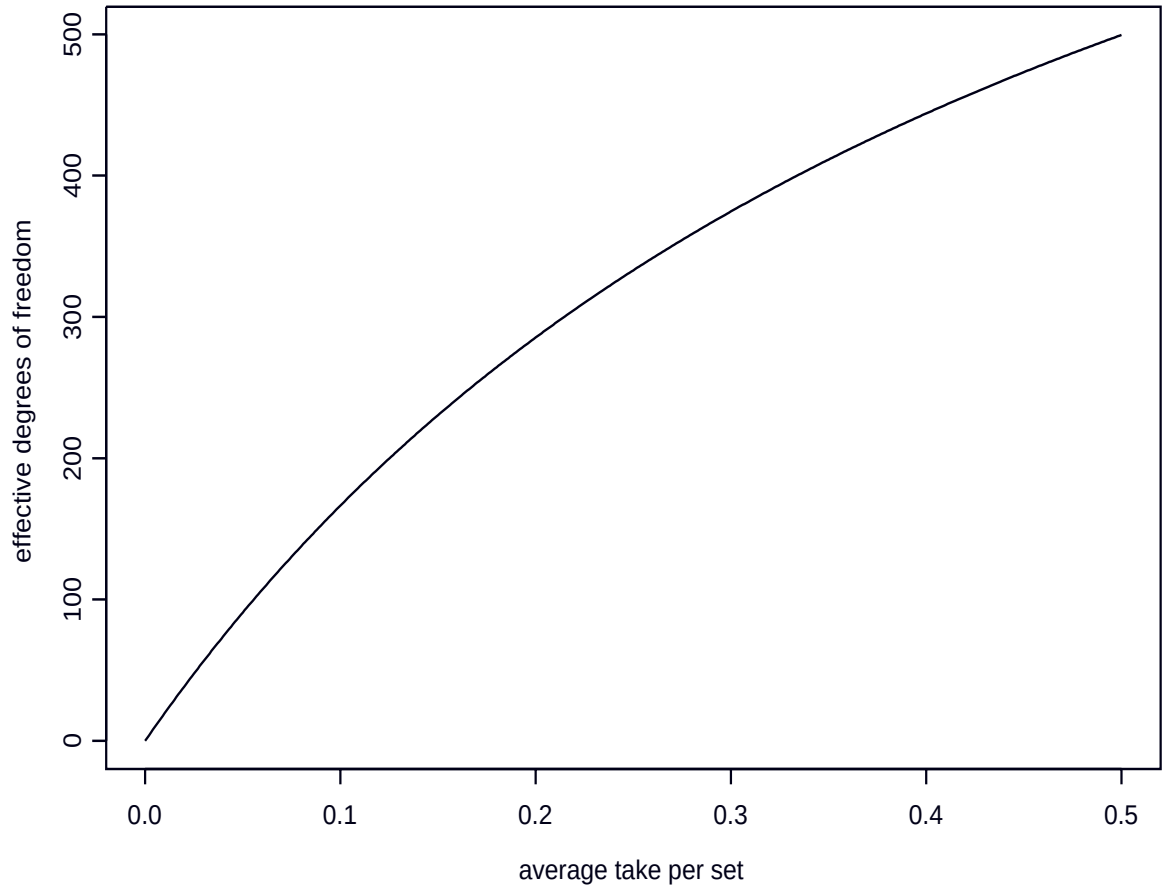

Figure 5. Approximate effective degrees of freedom assuming a Poisson distribution and 1000 independent observations.

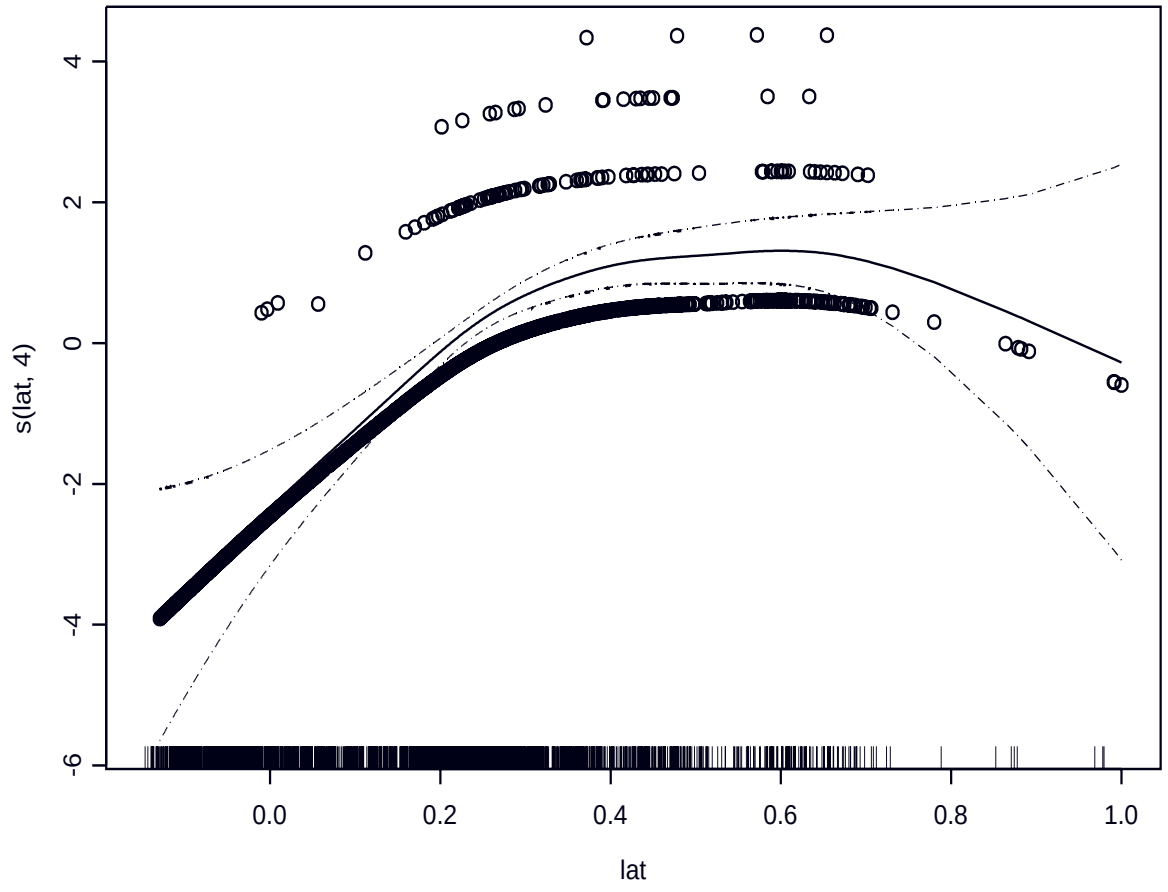

Figure 6. The fit of loggerhead take to the scaled latitude, for latitudes greater than or equal to $22°N$. The solid line represents the fitted smooth curve with 4 degrees of freedom, the dashed lines denote the fitted smooth plus or minus 2 standard errors (approximate) and demarcate a "standard error band", the black circles represent partial deviance residuals, and the bars on the x-axis are the rug-plot. The residuals are well distributed above and below the curve and follow the basic line of the curve. The standard error band shows a definitive curve and is narrow in the center of the curve, but wider at the right endpoints where there are few observations.

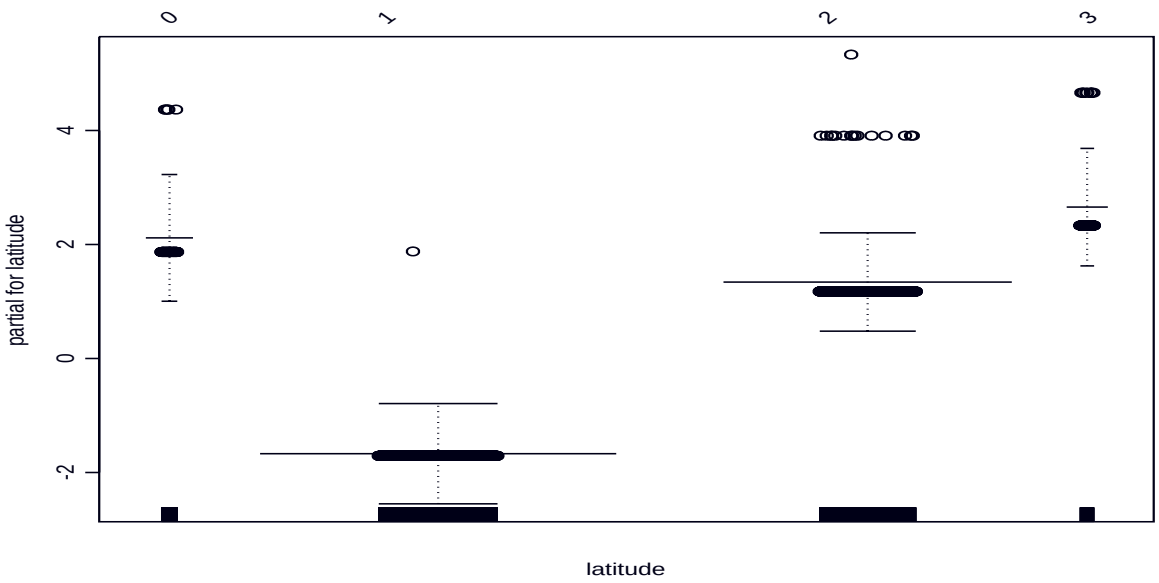

Figure 7. The fit of leatherback take to latitude expressed in four categories: 0=[$lat \leq 14.95°N$], 1=[$14.95°N < lat \leq 24.84°N$], 2=[$24.84°N < lat \leq 33.82°N$],3=[$lat > 33.82°N$]. The long horizontal solid line represents the fit. The length of this line is a function of the number of observations in the category. The vertical lines are twice standard error bands, the jittered black circles represent partial deviance residuals, and the solid bars on the x-axis are the rug-plot. The fitted take in the second category is distinct from the others.

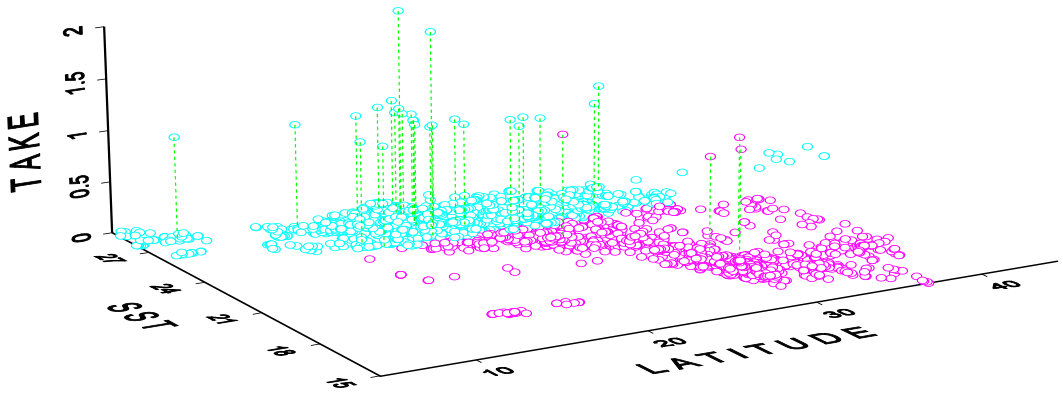

Figure 8. Takes of olive ridley versus latitude North and sea surface temperature degrees Celsius (SST). The different colors represent the two categories of sea surface temperature split at $24.22°C$ used in the prediction models (See Table 3).

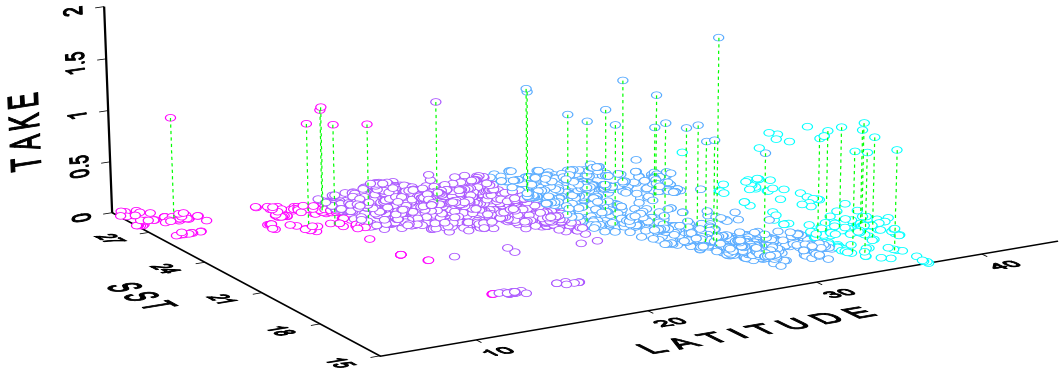

Figure 9. Takes of leatherback versus latitude North and sea surface temperature degrees Celsius (SST). The different colors represent the four categories of latitude split at $14.95°N$, $24.84°N$, and $33.82°N$ used in the prediction models (See Table 3).

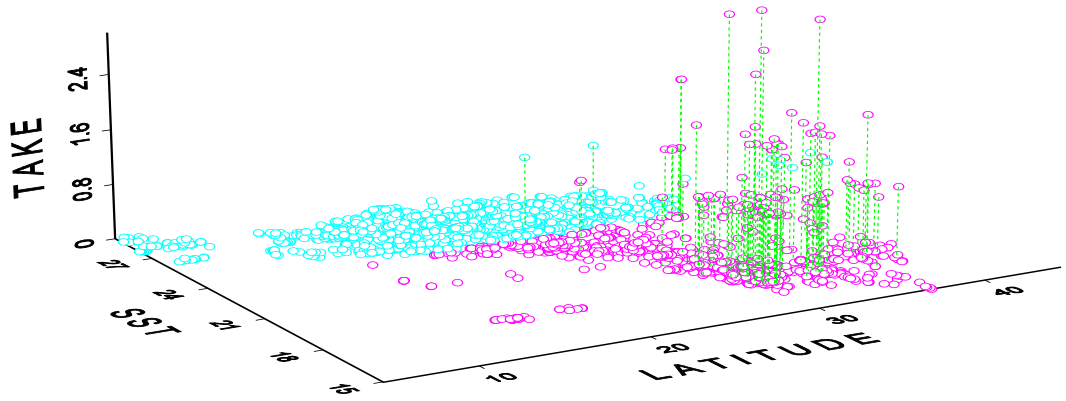

Figure 10. Takes of loggerhead versus latitude North and sea surface temperature degrees Celsius (SST). The different colors represent the two categories of sea surface temperature split at $23.77°C$ used in the prediction models (See Table 3).

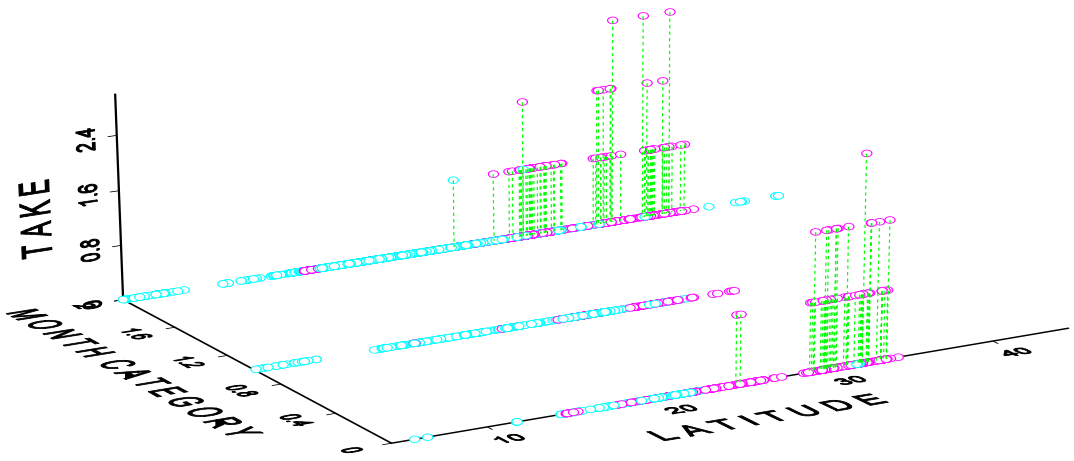

Figure 11. Takes of loggerhead versus latitude North and month classification ($0 = \{1, 2\}, 1 = \{5, 6\}, 2 = \{3, 4, 7 - 12\}$). The different colors represent the two categories of sea surface temperature used in the prediction models (red=$SST \leq 23.77°C$, blue=$SST > 23.77°C$).

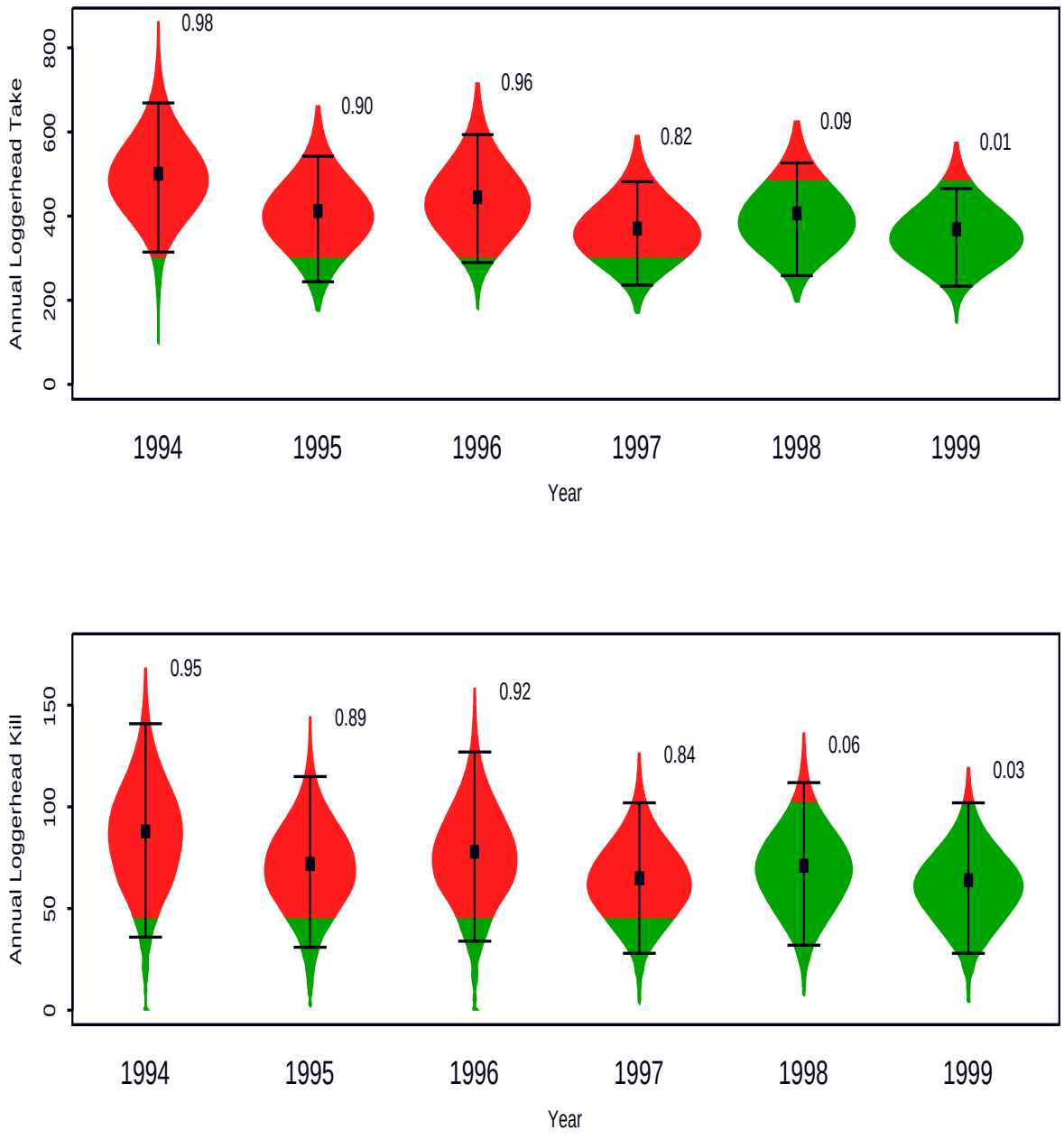

Figure 12. Take and kill estimates for loggerhead turtles. The approximate distributions of estimated take and kill are mirrored around the prediction intervals (dark line with bars at each end). The dark squares in the plots represent the point estimates. The proportion of the distribution above the trigger level is shaded in red and the estimated probability that the trigger level is exceeded is given above each year's plot.

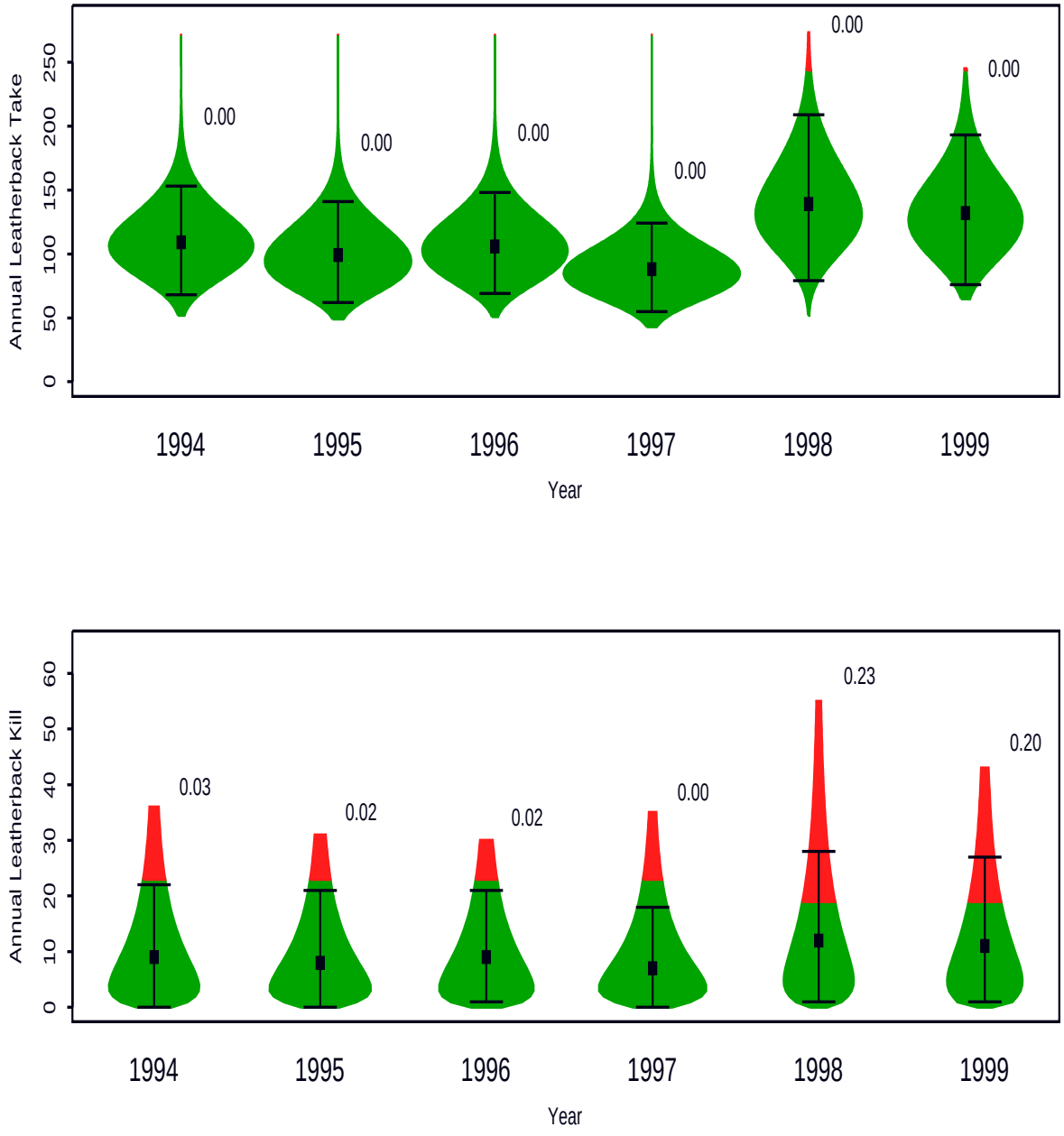

Figure 13. Take and kill estimates for leatherback turtles. The approximate distributions of estimated take and kill are mirrored around the prediction intervals (dark line with bars at each end). The dark squares in the plots represent the point estimates. The proportion of the distribution above the trigger level is shaded in red and the estimated probability that the trigger level is exceeded is given above each year's plot.

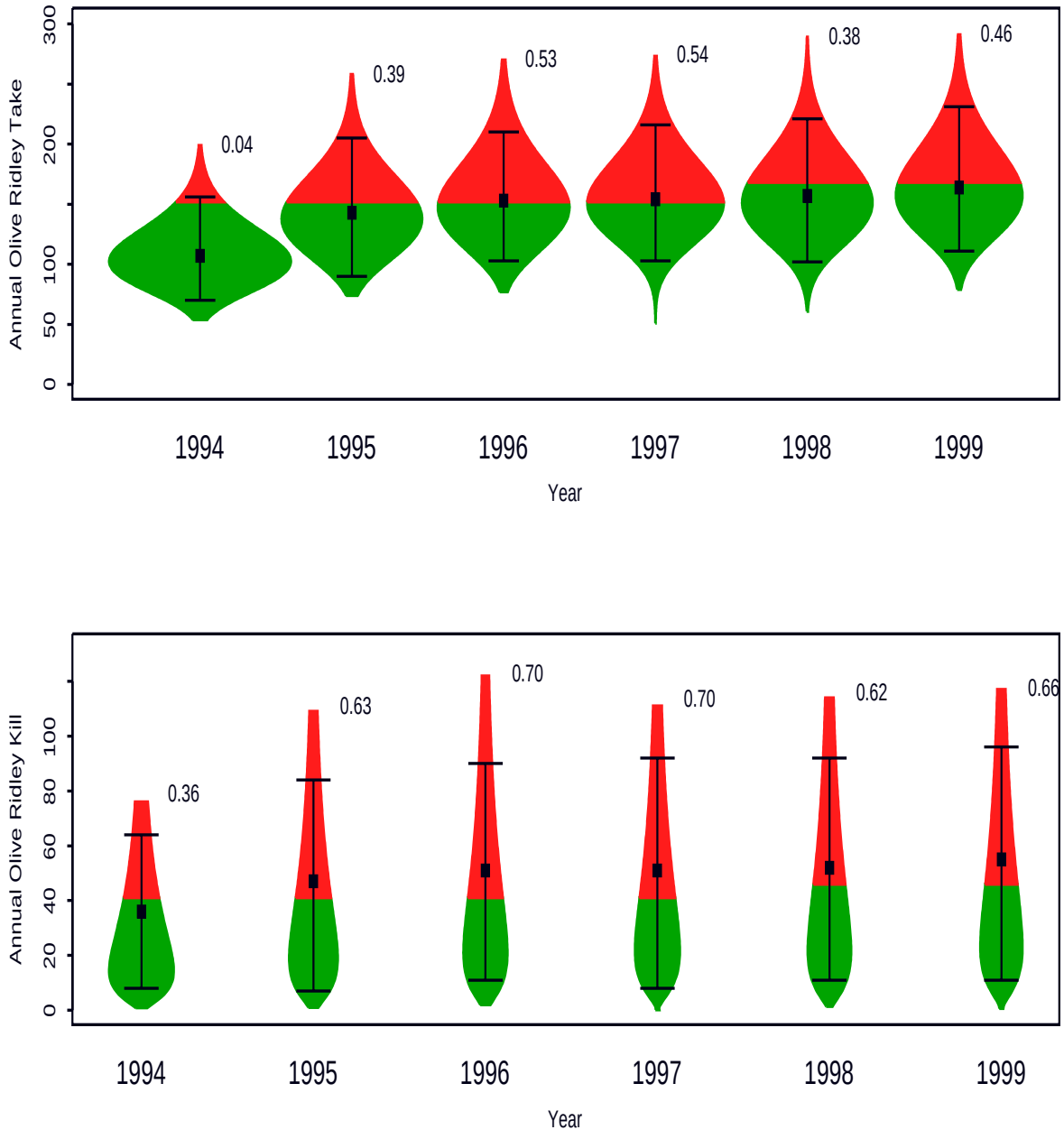

Figure 14. Take and kill estimates for olive ridley turtles. The approximate distributions of estimated take and kill are mirrored around the prediction intervals (dark line with bars at each end). The dark squares in the plots represent the point estimates. The proportion of the distribution above the trigger level is shaded in red and the estimated probability that the trigger level is exceeded is given above each year's plot.

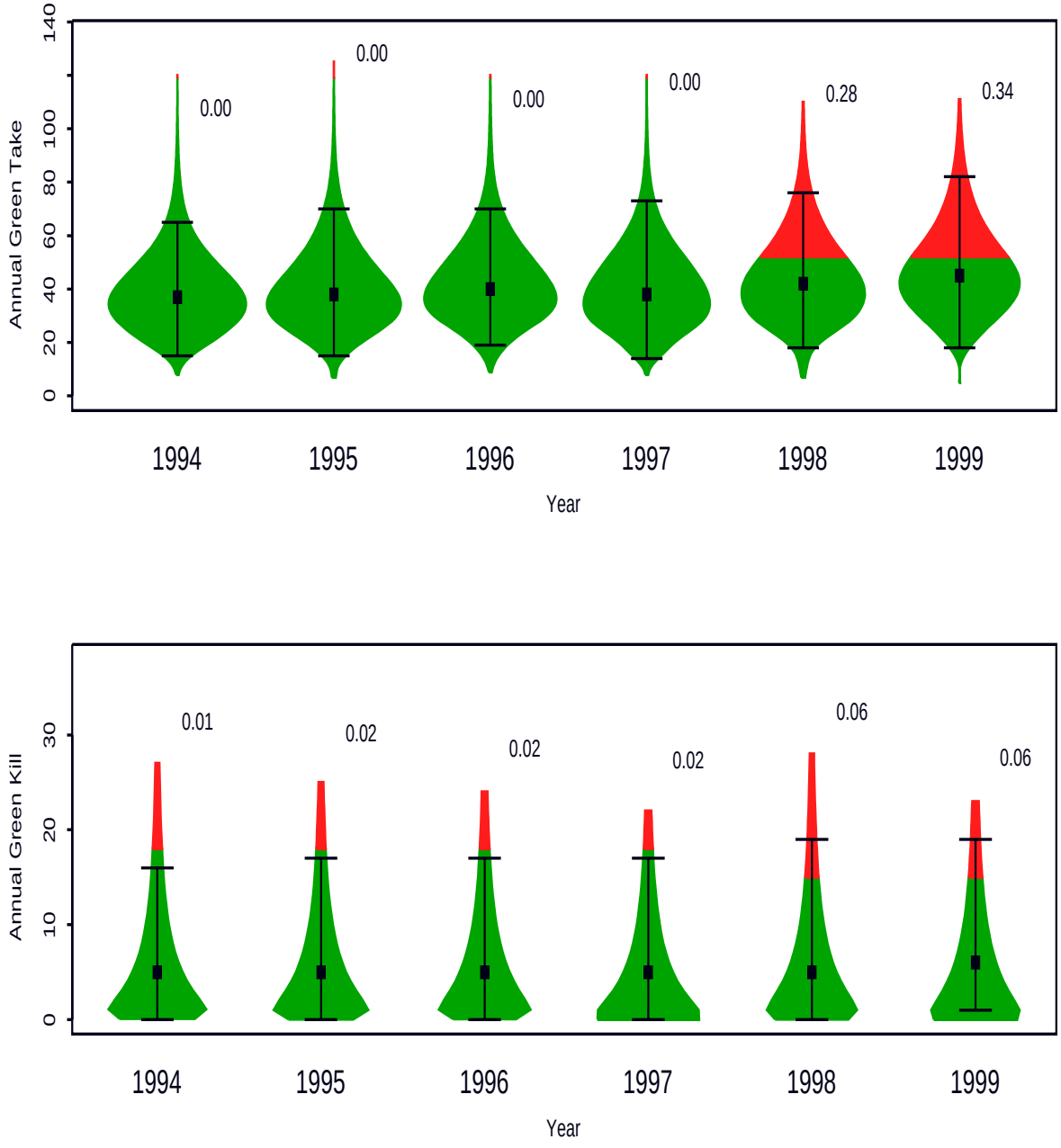

Figure 15. Take and kill estimates for green turtles. The approximate distributions of estimated take and kill are mirrored around the prediction intervals (dark line with bars at each end). The dark squares in the plots represent the point estimates. The proportion of the distribution above the trigger level is shaded in red and the estimated probability that the trigger level is exceeded is given above each year's plot.