Use of Zero- and One-Inflated Beta Regression to Model Availability of Loggerhead Turtles off the East Coast of the United States

Final Report

November 2014

Prepared by:

L.A.S. Scott-Hayward, D.L. Borchers and M.L. Burt
University of St. Andrews
S. Barco
Virginia Aquarium and Marine Science Center
H.L. Haas
Northeast Fisheries Science Center, NOAA's National Marine Fisheries Service
C.R. Sasso
Southeast Fisheries Science Center, NOAA's National Marine Fisheries Service
R.J. Smolowitz
Coonamessett Farm Foundation

Submitted to:

Naval Facilities Engineering Command, Atlantic
6506 Hampton Blvd.
Norfolk, VA 23508

Citation for this report is as follows:
Scott-Hayward, L.A.S., D.L. Borchers, M.L. Burt, S. Barco, H.L.Hass, C.R. Sasso and R.J. Smolowitz. 2014. *Use of Zero and One-Inflated Beta Regression to Model Availability of Loggerhead Turtles off the East Coast of the United States*. Final Report. Prepared for U.S. Fleet Forces Command. Submitted to Naval Facilities Engineering Command (NAVFAC) Atlantic, Norfolk, Virginia, under Contract No. N62470-10-D-3011, Task Order 40, issued to HDR Inc., Norfolk, Virginia. Prepared by CREEM, University of St. Andrews, St. Andrews, Scotland. July 2014.

The species experts (Barco, Haas, Sasso and Smolowitz) provided GPS and Argos data from satellite relay data loggers, filtered the satellite relayed data, assigned locations to behavior records, appended environmental predictor variables, provided the ecological context, reviewed the draft contract report, and made suggestions about how to refine this current product to make it applicable for direct use in turtle assessments. Scott-Hayward, Borchers and Burt selected the methods, formulated the models and fit the data.

Table of Contents

1. ABSTRACT **1**
2. INTRODUCTION **1**
3. METHODS **2**
3.1 BETA REGRESSION MODEL 3
3.2 ZERO AND ONE INFLATED BETA RANDOM EFFECTS MODEL 3
4. THE DATA **5**
4.1 MODEL FITTING AND SELECTION 5
4.2 MODEL PREDICTION 6
4.3 MODEL INFERENCE 8
5. RESULTS **8**
6. CONCLUSIONS **9**
7. POINTS FOR FURTHER CONSIDERATION **10**
8. ACKNOWLEDGEMENTS: **11**
9. REFERENCES **11**
10. TABLES **14**
TABLE 1: RESPONSE VARIABLES FOR LOGGERHEAD TURTLE TAG DATA. 14
TABLE 2: TABLE OF THE COVARIATES AVAILABLE FOR MODELING AND THEIR SOURCE. 14
TABLE 3: TABLE OF COVARIATES SELECTED FOR EACH MODEL AND THE ADJUSTED R^2. NOTE ALL COVARIATES ENTERED AS SMOOTH TERMS WITH ONE KNOT AT THE MEAN. 14
TABLE 4: TABLE OF RANDOM EFFECTS PARAMETERS, THE PRECISION PARAMETER FOR THE THREE MODELS AND THE CORRELATIONS BETWEEN THE RANDOM EFFECT PARAMETERS. STANDARD ERRORS ARE GIVEN IN PARENTHESES. 15
11. FIGURES **16**
FIGURE 1: MAPS OF THE STUDY REGION AND DATA POINTS FROM ALL OF THE TAGGED TURTLES (UPPER). 16
FIGURE 2: PLOTS OF THE RAW DATA FOR EACH OF THE THREE RESPONSE VARIABLES. 17
FIGURE 3: HISTOGRAMS OF EACH OF THE THREE RESPONSE VARIABLES; PROPORTION OF TIME SPENT AT THE SURFACE (TOP), IN THE TOP METER OF WATER (LT1; MIDDLE) AND TOP TWO METERS OF WATER (LT2; BOTTOM). 18
FIGURE 4: FIGURES SHOWING PREDICTIONS FOR A RANGE OF VALUES FOR EACH COVARIATE USING THE SURFACE MODEL. 18
FIGURE 5: FIGURES SHOWING PREDICTIONS FOR A RANGE OF VALUES FOR EACH COVARIATE USING THE LT1 MODEL. 19
FIGURE 6: FIGURES SHOWING PREDICTIONS FOR A RANGE OF VALUES FOR EACH COVARIATE USING THE LT2 MODEL. 20
FIGURE 7: SURFACE MODEL RAW AVAILABILITY DATA (UPPER) FOR SEASONS IN THE VIRGINIA AQUARIUM (VA) PREDICTION DATA ACROSS ALL YEARS, AND PREDICTED AVAILABILITY (LOWER) FOR THE VA LINE TRANSECT SURVEY DATA. 21

Figure 8: LT1 model raw availability data (upper) for seasons in the Virginia Aquarium (VA) prediction data across all years, and predicted availability (lower) for the VA line transect survey data. 22

Figure 9: LT2 model raw availability data (upper) for seasons in the Virginia Aquarium (VA) prediction data across all years, and predicted availability (lower) for the VA line transect survey data. 23

Figure 10: Figures showing CV scores for the Virginia Aquarium line transect surveys for each of the three model types; surface (a), LT1 (b), LT2 (c). 24

APPENDIX A 26

Table A1: Assessment of different Surface models and their AIC/BIC (Bayesian Information Criterion) scores where available. 26

Table A2: Assessment of different LT1 models and their AIC/BIC scores where available. 27

Table A3: Assessment of different LT2 models and their AIC/BIC scores where available. 28

Figure A1: Plots of the air temperature for each of the VA segments across surveys. 29

1. Abstract

The availability of an animal is defined as the animal being available at, or near, the surface so that it can be seen by an observer. Distance sampling methods (Buckland et al. 2001) adjust for the detectability of animals with respect to their distance from a point, or from a line travelled by an observer, but rarely adjust for their availability explicitly. To address this for loggerhead turtles (*Caretta caretta*), satellite tag data were collected from 156 turtles off the east coast of the United States. Location and depth were recorded in near continuous time, but returned as a single location and proportion of time spent in different depth bands for intervals between 4 to 6 hours. This non-binomial proportion data were modelled using a zero- and one-inflated beta regression, with a turtle random effect and smooth functions of covariates, to determine the availability of turtles at the surface, the top 1 m of water and top 2 m of water. Month, latitude and air temperature were chosen as covariates for the 1 and 2 m models and for the surface availability model, latitude and air temperature were chosen. In general, estimated availability was highest in the summer months, above Cape Hatteras (north of 38^{o}) and, when included in models, at air temperatures between 25 °C and 30 °C. Results from this initial study suggested that ignoring availability of turtles with respect to an observer (assuming it is 1) may substantially under estimate the population size. Development of the best models for predicting turtle availability is ongoing.

2. Introduction

As part of a coordinated effort to improve sea turtle density estimates off the East coast of the United States (US) several organizations assembled relevant data from satellite relayed data loggers deployed on loggerhead sea turtles (*Caretta caretta*). The intent of this collaboration was to provide the best available data to use to model the proportion of loggerheads near the ocean surface and within view of aerial observers.

Knowledge of the availability of animals to be detected by observers is important when determining abundance using distance sampling methods (Buckland et al. 2001). For example, by assuming that animals were always available to be detected we would underestimate the animal abundance twofold if animals were in fact only available for half the time. In aerial surveys, bias due to availability can be substantial because the plane is moving quickly so that animals do not have time to appear at the surface by the time the plane has passed. The motivation for this analysis arose from the need to estimate abundance of sea turtles in the coastal waters of Maryland and Virginia, including Chesapeake Bay, using aerial survey data.

Satellite-monitored radio transmitters ("satellite tags" for brevity) are attached to animals to collect data about their geographic locations in space and time, as well as their behavior. We are concerned here with modeling the availability of animals from satellite tag data that return proportions of time in different states (e.g., depth bands). Availability is defined as an animal on or near the surface such that it can be seen by an observer (aerial or boat). Although the data might be recorded continuously in time, when they are returned as proportions (as in the case of the data analyzed here), the sample units are the time intervals associated with each proportion and the responses are real numbers bounded on [0,1][1].

Although we are interested in modeling probabilities, because the responses are real numbers between 0 and 1, inclusive, they cannot be modeled as arising from a binomial distribution, which is only appropriate when responses are of the form of a count of the number of "successes" (*x*) out of *n* "trials". As a result, modeling the response using a binomial Generalized Linear Model (GLM; McCullough and Nelder 1989) is not appropriate. Rather we need a statistical distribution for responses that are proportions, not counts. The beta distribution is one possibility; it is extremely flexible and one can model its parameters as functions of explanatory variables in much the same way that one can model the probability parameter of a binomial distribution as a function of explanatory variables. This can be done in a GLM framework (Ferrari and Cribari-Neto 2004). However, because the beta distribution is only defined on (0, 1)[2] it alone is inadequate for modeling responses that are proportions that can be zeros or ones. One way of dealing with this is to transform the proportion data to lie on the (0, 1) range, but doing this introduces problems in interpreting results and/or introduces an ad-hoc aspect to the analysis (see Methods below). Instead, we combine a mixture of models to allow for the zero-one inflation (Ospina and Ferrari 2010, 2012), as detailed below.

The motivating data for this report come from tagged loggerhead turtles, where the aim was to determine the proportion of time turtles spend in various depth bands. The tag data enable us to determine the proportion of time turtles spend in different depth bands. We deal with three kinds of availability, according to the depths at which animals are believed to be visible to an aerial observer (which depends on prevailing environmental conditions). We consider situations in which animals are only available at the surface (S), animals are available when at depths shallower than 1m (LT1) and animals are available when at depths shallower than 2m (LT2). With a representative sample of turtles and good geographic coverage we can build an environmental model to predict the availability of turtles when aerial surveys were being conducted. In this report we develop a method and model the availability of animals conditional on a given sample. We do not address the issue of how representative the sample is, and indeed the data we analyze are likely not a random or even representative sample of animals that use the aerial survey region, because the selection of animals for tagging was done without consideration of their usage of this region.

We introduce below the beta regression model and then move onto zero- and one-inflated beta regression models with random effects (**Section 3**). In **Section 4** the data are described along with the specific fitting procedure to this application. **Section 5** summarizes the results, with some discussion in **Section 6** and points for future consideration in **Section 7**.

3. Methods

One way of dealing with proportion data that contains zeros and ones is to transform the data to restrict the range to exclude zero and one (Warton and Hui 2011; Smithson and Verkuilen 2006).

[1] Square brackets indicate the number given is included in the interval (here a zero and a one are possible).
[2] Round brackets indicate the number is excluded from the interval (here neither zero nor one are included).

In general it is better to use a mixture model to allow for the zero and one inflation (Ospina and Ferrari 2010, 2012).

Historically, the most frequently used method of analysis of percentage data was to utilize the arcsine square-root transform followed by linear modeling (Sokal and Rohlf 1995, Gotelli and Ellison 2004). This was largely surpassed by logistic regression (Zhao et al. 2001, Wilson & Hardy 2002). However, both methods were still used frequently as recently as 2008–2009 (Warton and Hui 2011). If data are binomial (of the form *x* out of *n*) then logistic regression is appropriate. In the case of non-binomial data, the data must be transformed to fulfill modeling assumptions, which in the case of a beta model involves transforming data to be between 0 and 1, but excluding the values 0 and 1. However, transforms are difficult to interpret as interpretation is only possible on the transformed scale. Paulino (2001) and Ferrari and Cribari-Neto (2004) proposed using beta regression to model rates and proportions that are continuous and bounded on (0,1). Using this kind of model means the regression parameters are easily interpretable in terms of the mean of the response. Smithson and Verkuilen (2006) further showed the benefits of beta regression in comparison to alternatives and suggested a transformation to compress data on a [0,1] scale to (0,1). They also allowed for variable dispersion in the beta regression model. The transformation of data is not ideal as it involves modeling something other than the data that were actually observed. Ospina and Ferrari (2010) showed that the beta distribution can be used to describe the continuous component and, mixed with a discrete distribution, can capture the probability mass at 0, 1, or both. This allows both zeros and ones to occur in the data and removes the need for the transformation described by Smithson and Verkuilen (2006).

3.1 Beta Regression Model

The beta distribution is a two-parameter function that describes the response and is bounded (0,1). It is therefore useful for modelling proportion data excluding the zeros and ones. The probability density function (pdf) of a beta-distributed random variable, y, is parameterized in terms of its mean, μ, and a parameter related to its variance, ϕ.

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, 0 < \mu < 1, \phi > 0 \qquad (1)$$

where $\Gamma(.)$ is the gamma function, $E(y) = \mu$ and $Var(y) = \frac{\mu(1-\mu)}{\phi+1} \equiv \sigma^2$. Larger values of ϕ correspond to less heterogeneity in the data (i.e., a decrease in $Var(y)$).

To map the covariate vector to the real line the mean is modeled using a suitable link function (e.g., logit, probit, or complementary log-log). The precision parameter may be assumed constant (constant variance assumption) or regressed onto the covariates by another link function. The link function for the precision parameter must result in a positive estimate as variance cannot be negative (e.g., log, square root).

3.2 Zero and One Inflated Beta Random Effects Model

Random effects have been added to the beta regression using a likelihood-inference model (Bonat et al. 2014; Verkuilen and Smithson 2012) and to the beta regression component in a zero-one-inflated beta model in a Bayesian framework (Galvis et al. 2013). To the best of our knowledge, no one has implemented random effects in both the continuous and discrete elements of the inflated beta model and in a classical likelihood framework.

Here we describe the zero and one-inflated beta regression model with a random effect in all three components. The pdf (*beinf*(.)) is a mixture of Bernoulli and Beta distributions:

$$beinf(y_i|\pi_{0i},\pi_{1i},\mu_i,\phi)=\begin{cases}\pi_{0i} & y_i=0\\ \pi_{1i} & y_i=1\\ (1-\pi_{0i}-\pi_{1i})f(y_i|\mu_i,\phi) & 0<y_i<1\end{cases}$$

where π_0 accounts for the probability of observations at zero and π_1 accounts for the probability of observations at 1. The function $f(y_i|\mu_i,\phi)$ is the probability density function for the beta distribution shown in $f(y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\quad 0<y<1, 0<\mu<1, \phi>0$ (1), parameterized in terms of its mean, μ, and precision, ϕ. The mean and the variance of y_i is given by:

$$E(y_i)=(1-\pi_{0i}-\pi_{1i})\mu_i+\pi_{1i} \qquad (2)$$

$$Var(y_i)=\pi_{1i}(1-\pi_{1i})+(1-\pi_{0i}-\pi_{1i})\left[\frac{\mu_i(1-\mu_i)}{1+\phi}+(\pi_{0i}+\pi_{1i})\mu_i^2-2\mu_i\pi_{1i}\right]$$

The log-likelihood for the i^{th} observation from the beta-inflated distribution is:

$$\log\mathcal{L}(\omega,\tau,\pi_0,\pi_1,y_i)=\log\Gamma(\omega+\tau)-\log\Gamma(\omega)-\log(\tau)+(\omega-1)\log(y_i)\ +$$
$$(\tau-1)\log(1-y_i)+\log(1-\pi_0)+\log(1-\pi_1)$$

where $\omega=\phi\mu$ and $\tau=\phi(1-\mu)$. Both ω and τ are shape parameters, with τ pulling the density towards one and ω pulling density toward zero.

Parameters μ,π_0 and π_1 for data point *i* and turtle *j* are found as follows:

$$\begin{aligned} g_1(\mu_{ij}) &= \boldsymbol{X}_{1ij}^T\boldsymbol{\beta_u}+\boldsymbol{Z}_{ij}\boldsymbol{b}_{1j}\\ g_2(\pi_{0ij}) &= \boldsymbol{X}_{2ij}^T\boldsymbol{\beta_o}+\boldsymbol{Z_{ij}}\boldsymbol{b}_{2j}\\ g_3(\pi_{1ij}) &= \boldsymbol{X}_{3ij}^T\boldsymbol{\beta_1}+\boldsymbol{Z}_{ij}\boldsymbol{b}_{3j}\end{aligned}$$

where $g(a)=\log\left(\frac{a}{1-a}\right)$, $\boldsymbol{\beta_u}$ is the vector of the fixed-effects regression coefficients of the beta distribution mean, μ_i. Similarly, $\boldsymbol{\beta_o}$ and $\boldsymbol{\beta_1}$ are the regression coefficients for π_{0ij} and π_{1ij}. $\boldsymbol{X}_{1ij}^T$, $\boldsymbol{X}_{2ij}^T$ and $\boldsymbol{X}_{3ij}^T$ are design matrices corresponding to the vectors of fixed effects. $\boldsymbol{Z_i}$ is the design matrix for the random effects with corresponding coefficient vectors, $\boldsymbol{b_{1j}}$, $\boldsymbol{b_{2j}}$ and $\boldsymbol{b_{3j}}$.

$$b_1,b_2,b_3\sim N(\boldsymbol{0},\boldsymbol{\Sigma}) \quad \text{, where } \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{b_1}^2 & \sigma_{b_1}\sigma_{b_2}\rho_1 & \sigma_{b_1}\sigma_{b_3}\rho_2\\ \sigma_{b_1}\sigma_{b_2}\rho_1 & \sigma_{b_2}^2 & \sigma_{b_2}\sigma_{b_3}\rho_3\\ \sigma_{b_1}\sigma_{b_3}\rho_2 & \sigma_{b_2}\sigma_{b_3}\rho_3 & \sigma_{b_3}^2\end{pmatrix} \quad (3)$$

The variance of the random effect is denoted σ^2 and ρ is the correlation coefficient.

4. The Data

The data are a summary of surfacing behavior from satellite-relay data loggers that were attached to 156 loggerhead sea turtles off the East Coast of the United States (**Figure 1**) and monitored from June 2010 to January 2014. Data cleaning and filtering, and collation of environmental covariates was completed by NOAA (National Oceanic and Atmospheric Administration) prior to analysis by CREEM. Specifically, data from the first 24 hours of deployment were deleted to exclude possible erratic behavior associated with the tagging process. Furthermore, due to the limited data offshore and the differing environmental conditions in the Gulf of Mexico, data from locations deeper than 200 meters and from the Gulf of Mexico and Bahamas were removed (**Figure 1**). The remainder contained information on the percentage of time turtles spent at, or near, the surface during a summary period (4–6 hours) and consisted of 32,792 data records from 152 turtles. Three response variables were provided by the tag data: proportion of time spent at the surface (S), proportion of time spent in the top 1 m of the water column (LT1) and proportion of time spent in the top 2 m (LT2) (**Table 1**). During the analyses of these responses it was assumed that the tag recordings were accurate. Whilst this may not necessarily be the case, there was no quantitative information available to provide a correction or inclusion of bias or additional uncertainty in recording.

Figure 2 shows the spatial distribution of each response. There were more data to the north of the study region and an indication that animals were spending more time in the depth bands in the north compared to the south. As ectothermic reptiles, the distribution, biology and behavior of sea turtles are strongly linked to the thermal regimes of their environment (Bell and Richardson 1978, Spotila et al. 1997). For this reason many of the environmental covariates relate to temperature. The covariates available for modeling are shown in **Table 2**. For the purposes of developing the methods, only those covariates that did not have missing values were evaluated (e.g. curved carapace length included missing values). Cloud cover was not evaluated as NOAA considered it to be a substantial contributer to estimation of the downward solar radiation covariate, which was preferred.

4.1 Model Fitting and Selection

The distributions of the LT1 and LT2 responses show data at both zero and one and so models were fitted using a zero and one-inflated beta regression (**Figure 3**). The surface model is very right skewed with few data at one and so a zero-inflated beta regression was fitted (**Figure 3**). In fact the 9 data points (out of 32,792) that had S=1 were removed to allow the zero-only inflated model to be fitted to the surface response (9 data points were too few to fit a one-model reliably).

To summarize, the response models are a mixture of three (zero- and one-inflated beta) or two (zero-inflated beta) sub-models:

1. A "beta" model that models the expected proportion of time an animal is available when this is neither 0 nor 1, as a function of some suitable set of covariates (fixed effects) and a turtle random effect. (It is referred to as the "beta" model because the observed proportion is assumed to have a beta distribution.)

2. A "zero" model (zero-inflated) that models the expected proportion of responses that have zero availability, as a function of some suitable set of covariates (fixed effects) and a turtle random effect,
3. A "one" model (one-inflated) that models the expected proportion of responses that have 100% availability, as a function of some suitable set of covariates (fixed effects) and a turtle random effect,

From here we describe the fitting of the zero-one-inflated beta model. The zero-only beta model (surface availability) is a simplification of this model (sub-model 3 not included for zero-inflated model).

The effects of the covariates in each of the above were modeled using nonparametric smooth functions of the covariates, and this was implemented by means of regression splines, using spline basis functions. This approach allowed the models to be formulated as generalized linear mixed models (GLMMs) (McCulloch & Neuhaus 2001), and hence allowed GLMM software to be used for model fitting. For this implementation, each GLMM requires a design matrix with columns containing the spline basis functions evaluated at relevant covariate values; these were constructed before calling the GLMM fitting function. Each column of each design matrix has a regression coefficient parameter associated with it. In addition there are random effect parameters for each model.

The covariates and spline basis functions might be different for each of the sub-models, or shared across sub-models. All smooth terms were specified as B-splines with one knot at the mean, except for Month which was specified as a cyclic cubic spline with boundary knots at 0 and 12. The software program R, version 3.1.0 (R Core Team 2014) was used to construct the design matrix for each model (and sub-model component).

The following is an example of the three sub-models with two smooth covariates (month and latitude) and a random effect for the intercept in each model:

$$\begin{array}{ll}
\mu \text{ model:} & \beta_o + b_1 + \beta_1\text{month1} + \beta_2\text{month2} + \beta_3\text{month3} + \beta_4\text{lat1} + \beta_5\text{lat2} + \beta_6\text{lat3} \\
\pi_0 \text{ model:} & zero_o + b_2 + zero_1\text{month1} + zero_2\text{month2} + zero_3\text{month3} + zero_4\text{lat1} + zero_5\text{lat2} + zero_6\text{lat3} \\
\pi_1 \text{ model:} & one_o + b_3 + one_1\text{month1} + one_2\text{month2} + one_3\text{month3} + one_4\text{lat1} + one_5\text{lat2} + one_6\text{lat3}
\end{array} \quad (26)$$

The random effect for the data in this example was tag number, which identified individual turtles.

Initially, models were fitted for each individual covariate (smooth or linear) and the covariate was used for all parts of the likelihood. The order of best predictors was determined from the AIC scores and covariates were added to the model in this order until there was no improvement in AIC or parameterization issues occurred.

The models were fitted using SAS 9.4 and macros from Swearingen et al. (2011 and 2012) adapted for the inclusion of the random effect (SAS® `PROC NLMIXED`).

4.2 Model Prediction

Predictions of availability were required to estimate loggerhead abundance within the coastal waters of Maryland and Virginia, including Chesapeake Bay. Six aerial surveys had been conducted; in 2011 (in spring, summer and fall), in 2012 (in spring and summer) and in 2013 (summer), referred to hereafter as 'VA aerial surveys'. Availability was required for each segment of survey effort.

Three model objects are required for prediction, given the fitted models described above:

1. Model specifications for each sub-model (explanatory variables, degrees of freedom of smooths, link functions, and error models)
2. Vector of coefficients for beta, zero-inflated, and one-inflated model components (including random effect parameters)
3. Covariance matrix for these coefficients.

Before prediction is possible, the design matrices with columns corresponding to the basis functions evaluated at the relevant covariate values at every prediction grid point must be constructed. If any of the covariates in the model change with time, the design matrix will need to be calculated for every time point of interest.

Due to the presence of a (turtle) random effect in the GLMM, predictions were averaged over the random effects distribution, i.e., we required population average predictions. We calculated a population average estimated availability for each prediction location for each model as follows:

$$\hat{\mu}_{gj} = g_{\mu}^{-1}\left(\sum_{g} \boldsymbol{x}_{\boldsymbol{\mu},gj}\widehat{\boldsymbol{\beta}_{\boldsymbol{\mu}}} + b_{1,j}\right)$$

$$\hat{\pi}_{0,hj} = g_{0}^{-1}\left(\sum_{h} \boldsymbol{x}_{\boldsymbol{0},hj}\widehat{\boldsymbol{\beta}_{\boldsymbol{0}}} + b_{2,j}\right)$$

$$\hat{\pi}_{1,ij} = g_{1}^{-1}\left(\sum_{i} \boldsymbol{x}_{\boldsymbol{1},ij}\widehat{\boldsymbol{\beta}_{\boldsymbol{1}}} + b_{3,j}\right)$$

where $g_{\mu}^{-1}()$, $g_{0}^{-1}()$, and $g_{1}^{-1}()$ are the inverse logit link functions for the three sub-models, $\boldsymbol{x}_{ij}$ is the i^{th} row of the relevant design matrix, $\boldsymbol{\beta}$'s are the estimated coefficient vectors. These lengths of these vectors will be equal if all three submodels contain the same covariates specified in the same way (e.g. same degrees of freedom per smooth). The *b* parameters are the random effects for turtle, *j*, sampled from a multivariate normal using the estimated covariance matrix. These three sub-models can be combined using Equation $\mathrm{E}(\mathrm{y_i}) = (1 - \pi_{0i} - \pi_{1i})\mu_i + \pi_{1i}$ (2) to give a single population-averaged prediction of availability probability in each cell:

$$\widehat{Y}_i = E(y_i|\mu_i, \pi_{0i}, \pi_{1i}) = (1 - \hat{\pi}_{0i} - \hat{\pi}_{1i})\hat{\mu}_i + \hat{\pi}_{1i}$$

This process was repeated 1,000 times and averaged to obtain the average availability at each location. Predicting for the surface model is slightly different because it is only zero-inflated (and not zero- and one-inflated). Thus, there is no calculation for π_1 and the random effects covariance matrix is (2 x 2) rather than (3 x 3).

4.3 Model Inference

To make confidence intervals for the predictions we calculated a prediction interval. Predictions were made for a random sample of 500 individuals (random effect sampled from the multivariate normal) and for each individual, 1000 sets of regression coefficients (i.e. β's) sampled from a multivariate normal using the estimated coefficients and their covariance matrix (to include parameter uncertainty). Prediction intervals were calculated by taking 2.5 and 97.5 percentiles for each prediction cell (of 500,000 sets of predictions) to give 95% prediction intervals for the predicted availability surface. A Coefficient of Variation (CV) was also calculated for each cell using the mean and variance across bootstraps (CV = standard deviation/mean).

5. Results

The final model covariates chosen using AIC for each response are shown in **Table 3** along with adjusted R^2 values. Month, Latitude and air temperature were chosen for LT1 and LT2 responses. The surface model was nested within the other two as air temperature was not selected. Tables of all the models trialled and their AIC scores can be found in **Appendix A, Tables A1-A3**. The LT2 model had the highest adjusted R^2 (0.38) and was therefore the best fitting model of the three. The surface model was the poorest (adj. R^2 = 0.10).

Figures 4, 5 and 6 show the relationship of the selected covariates to each of the responses. These show that the tagged loggerhead turtles spend more time in the surface waters in the summer months (May to July) than at other times of year. There also seems to be a preference for the north of Cape Hatteras or south of the tip of Florida (above 38° and below 26° latitude) although uncertainty increases at either end of the latitude range where, particularly in the south, there are few observations. Air temperature is not in the surface model but turtles show a preference for the top 1 and 2 meters when the temperature is between 25 °C and 30 °C.

The random effects parameters are presented in **Table 3**: Table of covariates selected for each model and the adjusted R2. Note all covariates entered as smooth terms with one knot at the mean.

Model	Covariates	Adjusted R2
Surface	Month and Latitude	0.10
LT1	Month, Latitude, and Air Temperature	0.34
LT2	Month, Latitude, and Air Temperature	0.38

Table 4 provides the estimated correlations between the three random effects. For the surface model, there is a high, negative correlation between the beta and zero components (-0.689). This indicates that if a turtle has a small random effect coefficient for the beta component (below average), then the same turtle will have a large random effect coefficient for the zero component (above average). In the LT1 model, the beta/0 and beta/1 correlations are weakly positive (0.128 and 0.139; if above average, then above in both), while the 0/1 correlation is negative (-0.558); a turtle that is above the average in the 1 component (i.e. always available in time period) will be below the average in the zero component (i.e. never available). The beta/0 and 0/1 correlations are very weakly positive (0.046) and very weakly negative (-0.009) for the LT2 model. There is a weak positive beta/1 correlation (0.218) indicating if a turtle is above average in the beta component it may also be above average in the 1 component.

Visual inspection suggests that the predictions for the VA aerial surveys match the corresponding season's raw availability data (**Figure 7, 8 and 9**) reasonably well. For the surface model results, while the lowest predicted and observed availability is in Fall 2011, the predictions are generally higher than the observations. The surface model contains only two terms, for month and latitude, compared to three terms for the LT1 and LT2 models. This means that, for a given month, predictions made using the surface model can only change due to latitude. With the relatively small latitudinal range of the VA survey region (36.5^{o}N – 38.5^{o}N) compared to the large latitudinal range of the data used to fit the model (see Figure 2a), the corresponding range in predicted availability is also relatively small, approximately 0.125 to 0.175 (e.g. see **Figure 4**). For LT1 and LT2, the data suggest a difference in availability between offshore areas and within Chesapeake Bay, however, there are few data within the bay compared with offshore. The dynamic variable, air temperature, is not able to pick up this change due, in part, to the lack of spatial variability in air temperatures within surveys (**Figure A1**). Consequently, the predictions of availability for the Chesapeake Bay region seem to be too high on average.

Spatially explicit CV scores for each model are on average highest for the surface model (30-50%), however, there is one survey in Fall 2011 that has very high CV scores in the LT1 model (**Figure 10)**.

6. Conclusions

The modeling process has suggested that time of the year (month), latitude, and, to a lesser extent, air temperature were the most important explanatory variables for the availability of loggerhead turtles. The maps produced do not indicate presence of turtles, but rather the availability of turtles in differing parts of the water column, should turtles be found there. They show that availability of turtles varies both spatially and temporally, which makes it very important to know where and when surveys took place when estimating animal abundance so that the appropriate availability can be taken into account.

One outcome of this modeling process was to incorporate the availability of turtles into an analysis of data collected during aerial line transect surveys of loggerhead turtles (conducted by Virginia Aquarium Foundation). The area of study was the Atlantic coasts of Virginia and Maryland and in Chesapeake Bay. Predicted availability was taken from the surface model for the Chesapeake Bay region, due to the high turbidity of the water, and the LT2 model for the rest of the survey region. With these adjustments the probability of turtles being available was lower on average in Chesapeake Bay compared with the Atlantic Ocean. This led to a substantial adjustment in the abundance of turtles and highlighted the importance of including information about availability. See the report by Burt et al. (2014) for more details.

This report outlines an appropriate statistical method for analyzing proportional data that include zeros and ones. These methods provide a basis for developing further models, for comparison with other methods and for determining appropriate methods to estimate availability for turtle stock assessment. With this in mind, the following section lists some points for further consideration.

7. Points for further consideration

Possible future research includes: evaluating possible effects of biased sampling of turtles to attach tags to; errors in the tag data; the spatial scope of data and model predictions (particularly the north-eastern and south-western extremes); patterns of outliers (particularly spatial patterns including a comparison of coastal and offshore strata); whether combinations of variables that are expected a-priori to drive turtle availability (such as surface and bottom temperature) could replace proxy variables (such as latitude); the possible relationship between turtle size and availability (which requires dealing with the issue of missing size observations); the practicality of using derived environmental predictor variables (such as an index of thermocline strength) rather than interaction terms and the comparison of the utility of the models developed here with simpler models for purposes such as turtle stock assessment.

The analyses performed here assumed no errors in tag depth recording. There may be recording errors in the tag data and this may result in, for example, false zeros (i.e., an animal may be incorrectly recorded as never spending time within surface waters during the time period). Further work could include investigation of tag recording error, for example, are there are more likely to be errors with zero values or ones, and then in light of this, revising the modeling of the tag data, if necessary.

The tag data has been assumed to come from a representative sample of turtles. There has been no attempt to address how representative the sample is or what the possible effects of biased sampling of turtles would have on this analysis.

The definition of the proportion of time at the surface (S) and how it applies to seeing turtles from survey planes should be considered. An S reading only happens when the salt water switch is dry and water splashing onto the switch (especially with bio fouling) could lead to non-S readings even if the turtle is at the surface. Rather than using S, LT1 (within 1m of the surface) could be used instead. In addition, wind direction may have a substantial effect on whether a turtle can be seen at, or under, the surface.

Due to limitations associated with optimization, the methods described here failed to converge for some mechanistic environmental variables deemed important for turtle availability (for example, solar radiation, bottom temperature, and surface temperature). Other methods for dealing with proportion data, that contain zeros and ones, transform the data so that it lies between zero and one, and these methods may not suffer from such convergence problems. There are drawbacks to transforming variables but the complexity of the modelling approach implemented here may outweigh the drawbacks of other approaches. Potentially, simpler methods may allow these mechanistic variables and interactions to be investigated. Transformation of the data may also be appropriate for dealing with possible errors (false zeros and false ones) in the tag data.

The covariates included in the model were selected on the basis of the AIC scores and included covariates such as latitude and month which may be thought of as proxy variables for some unmeasured variable. In order to explain turtle behavior (rather than purely describe which was the aim here), a biologically-driven procedure for covariate selection may be more appropriate than the objective approach used.

Curved carapace length (CCL) could be considered as an explanatory variable (possibly as a random, rather than a fixed, effect) to establish if it is biologically important. It was not included here because some values were missing. Consideration would need to be given to values assigned to CCL for prediction (both for missing values and for application of predicted availability to surveys).

The methods described here were implemented using SAS and R. The R package zoib allows a zero- and one-inflated beta model in a Bayesian framework and this may be worth investigating as an alternative approach and allow a more streamlined implementation which would be useful for future updates as more tag data becomes available.

8. Acknowledgements:

This study was funded in part by the U.S. Department of the Interior, Bureau of Ocean Energy Management, Environmental Studies Program, Washington DC, through Inter-Agency Agreement; the Atlantic Sea Scallop Research Set Aside Program, Virginia Maryland Section 6 Program, and funds from Coonamessett Farm Foundation, National Marine Fisheries Service, and the Virginia Aquarium and Marine Science Center. We are also grateful to the vessel crew and captains who made tag deployments possible and to Beth Josephson and Arliss Winship for the projected grid. Also, thanks to M Mackenzie (CREEM, University of St Andrews) for providing technical expertise and advice.

9. References

Bell, R., & Richardson, J. I. (1978). An analysis of tag recoveries from Loggerhead Sea turtles (*Caretta caretta*) nesting on Little Cumberland Island, Georgia. Florida Department of Natural Resources, 33: 19.

Bonat, W. H., & Ribeiro Jr, P. J. (2014). Bayesian analysis for a class of beta mixed models. arXiv preprint arXiv:1401.2957.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2001) Introduction to Distance Sampling. Oxford University Press, Oxford.

Burt, M.L., Scott-Hayward, L.A.S. and Borchers, D.L. (2014). Analysis of aerial surveys conducted in coastal waters of Maryland and Virginia, including Chesapeake Bay, 2011-2013; Loggerhead turtles. Unpublished technical report, CREEM, University of St Andrews, UK.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31(7), 799-815.

Figueroa-Zuniga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. (2013). Mixed beta regression: A Bayesian perspective. Computational Statistics & Data Analysis 61, 137-147.

Galvis, D. M., Bandyopadhyay, D., & Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data. Statistics in Medicine. DOI: 10.1002/sim.6179

Gotelli, N. J. & Ellison, A. M. (2004). A primer of ecological statistics. Sinauer Associates, Sunderland, Massachusetts, USA.

McCullagh, P. and J.A. Nelder (1989). Generalized Linear Models (2nd ed.). Chapman & Hall, London

McCulloch, C. E., & Neuhaus, J. M. (2001). Generalized Linear Mixed Models. John Wiley & Sons, Ltd.

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. Statistical Papers, 51(1), 111-126.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis, 56(6), 1609-1623.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. Political Analysis, 9(4), 325-346.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods, 11(1), 54.

Sokal, R. R., & Rohlf, F.J. (1995). The principles and practice of statistics in biological research. Third Edition. W.H. Freeman, New York, New York, USA.

Spotila, J.R., O'Connor, M.P. & Paladino, F.V. (1997) Thermal biology. Biology of Sea Turtles (ed. by P.L. Lutz and J.A. Musick), pp. 297–314. CRC Press, Boca Raton, Florida.

Swearingen, C.J., Melguizo Castro, M.S. and Bursac, Z. (2012) Inflated beta regression: Zero, one, and everything in between. SAS® Global Forum Proceedings 2012; Paper 325: 1-11

Swearingen CJ, Melguizo Castro, M.S., and Bursac, Z. (2011). Modeling percentage outcomes: The %Beta_Regression macro. SAS® Global Forum Proceedings 2011; Paper 335:1–12

Verkuilen, J., & Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. Journal of Educational and Behavioral Statistics, 37(1), 82-113.

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology, 92(1), 3-10.

Wilson, K., & Hardy, I. C. (2002). Statistical analysis of sex ratios: an introduction. Sex ratios: concepts and research methods, 48-92.

Zhao, L., Chen, Y., & Schaffner, D. W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. Applied and environmental microbiology 67.5; 2129-2135.

THIS PAGE INTENTIONALLY LEFT BLANK

10. Tables

Table 1: Response variables for loggerhead turtle tag data.

Response	
S	Proportion of time at the surface (0 meter)
LT1	Proportion of time in the top 1 meter of water column
LT2	Proportion of time in the top 2 meters of water column

Table 2: Table of the covariates available for modeling and their source.

Covariate	Unit	Source
Surface Temperature	°C	MGET (Marine Geospatial Ecology Tools) , HyCom (HYbrid Coordinate Ocean Model)
Bottom Temperature	°C	MGET (HyCom)
Surface solar radiation downwards	J/m^2	Movebank (European Centre for Medium-Range Weather Forecasts [ECMWF])
Air temperature	Kelvin	Movebank (ECMWF)
Cloud cover	(0-1)	Movebank (ECMWF)
Curved carapace length	cm	n/a
Water Depth	m	Movebank (National Oceanic and Atmospheric Administration)
Distance to Coast	Km	Movebank (National Aeronautics and Space Administration Ocean Biology Processing Group)
Month		Tag
Latitude	Degrees	Tag
Longitude	Degrees	Tag
PIT tag number		Tag

Table 3: Table of covariates selected for each model and the adjusted R^2. Note all covariates entered as smooth terms with one knot at the mean.

Model	Covariates	Adjusted R^2
Surface	Month and Latitude	0.10
LT1	Month, Latitude, and Air Temperature	0.34
LT2	Month, Latitude, and Air Temperature	0.38

Table 4: Table of random effects parameters, the precision parameter for the three models and the correlations between the random effect parameters. Standard errors are given in parentheses. b_1 is the random effect for the beta component, b_2 for the zero component and b_3 for the one component (see Equation (3) for details). The bottom three rows show the correlation coefficients; between the beta and 0 components (ρ_1), the beta and 1 components (ρ_2) and the 0 and 1 components (ρ_3). There are no standard errors given for the correlation coefficients as the correlation coefficients cannot be isolated, for example, ρ_1 cannot be isolated from $\sigma_{b1}\ \sigma_{b2}\rho_1$.

Parameter	Surface	LT1	LT2
σ^2_{b1}	0.180 (0.024)	0.236 (0.030)	0.200 (0.026)
σ^2_{b2}	4.782 (0.875)	6.341 (1.85)	6.056 (1.78)
σ^2_{b3}	-	28.37 (21.84)	34.10 (22.37)
$\sigma_{b1}\ \sigma_{b2}\rho_1$	-0.639 (0.115)	0.157 (0.170)	0.051 (0.156)
$\sigma_{b1}\ \sigma_{b3}\rho_2$	-	0.360 (0.440)	0.574 (0.433)
$\sigma_{b2}\ \sigma_{b3}\rho_3$	-	-7.489 (6.75)	-0.128 (2.74)
ρ_1	-0.689	0.128	0.046
ρ_2	-	0.139	0.218
ρ_3	-	-0.558	-0.009

11. Figures

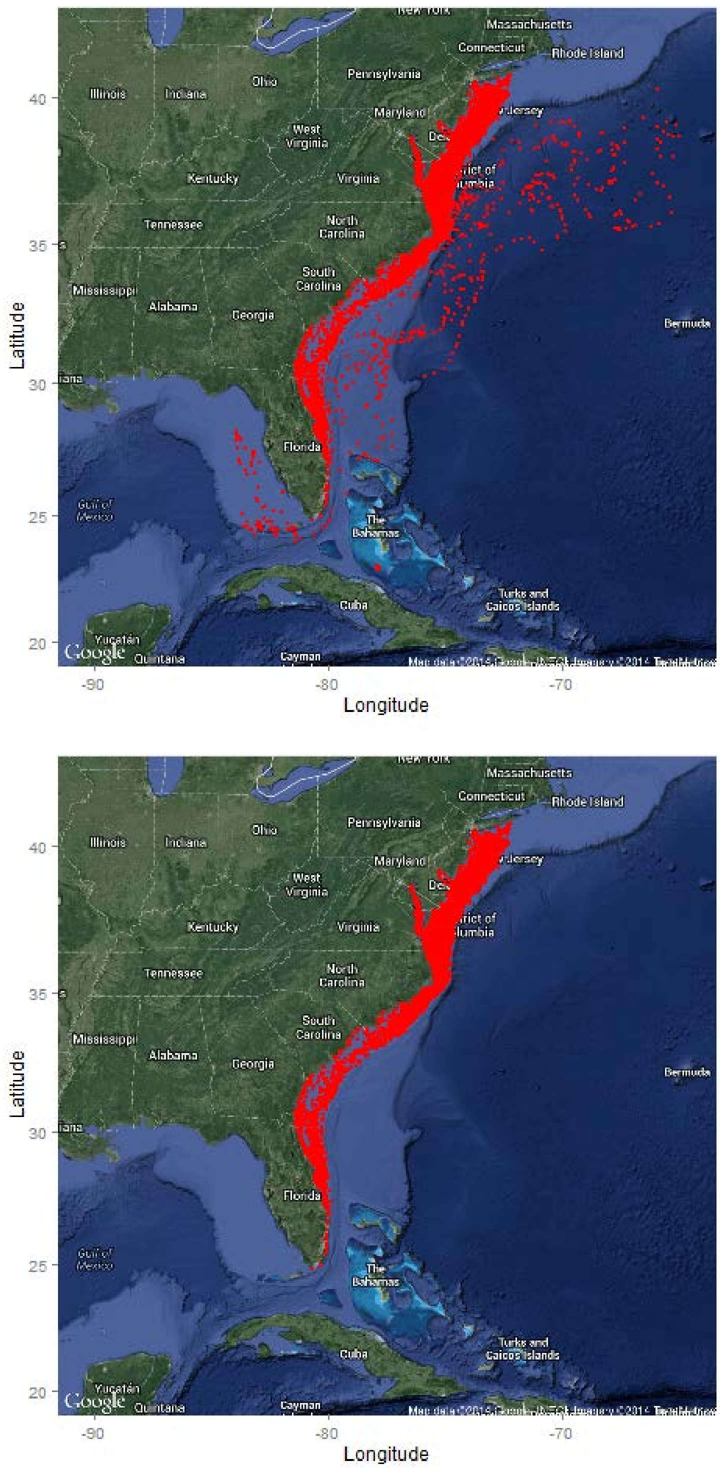

Figure 1: Maps of the study region and data points from all of the tagged turtles (upper). The lower map is the reduced data set where data beyond the 200-meter contour and from the Gulf of Mexico and Bahamas regions have been excluded.

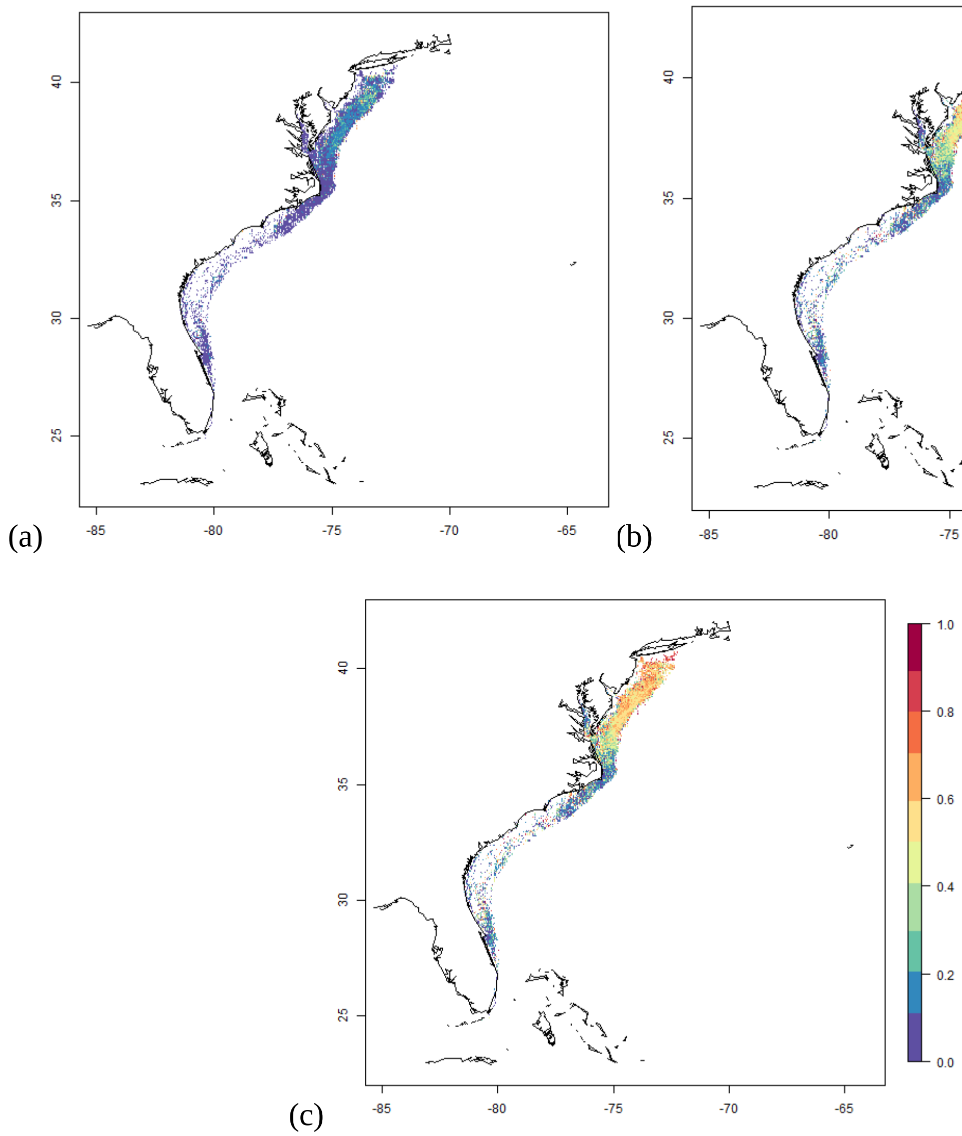

Figure 2: Plots of the raw data for each of the three response variables. The data represent (a) the proportion of time spent at the surface, (b) within 1 meter of the surface and (c) within two meters. In heavily sampled areas, the colors represent a mean of a number of records, because plotting individual records in these areas results in many records being obscured.

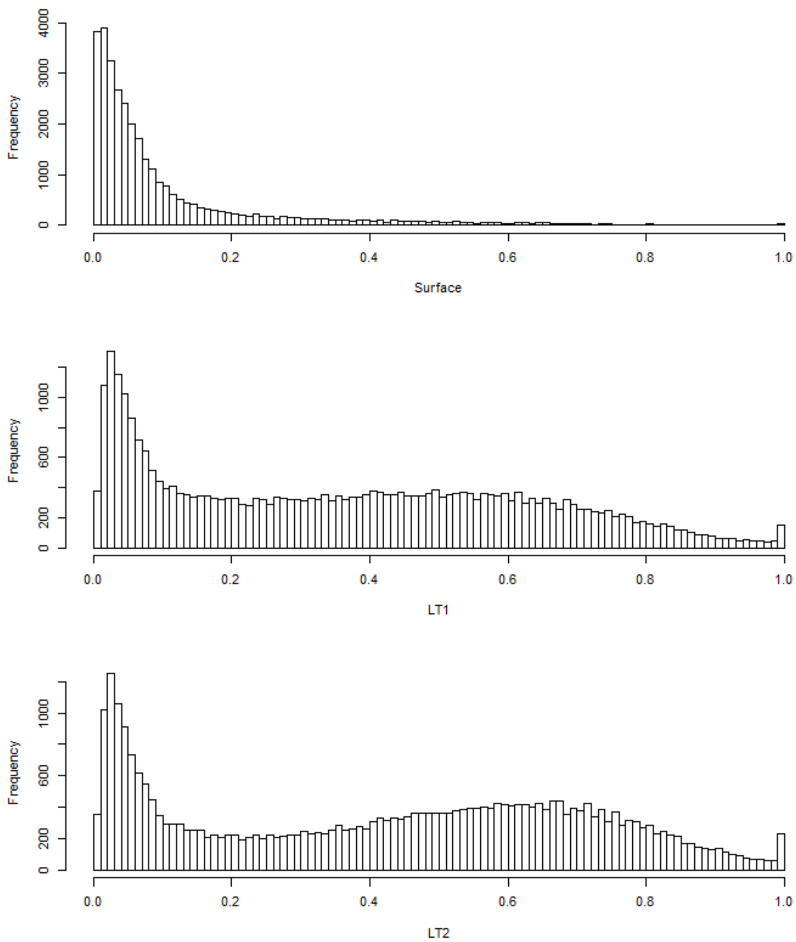

Figure 3: Histograms of each of the three response variables; proportion of time spent at the surface (top), in the top meter of water (lt1; middle) and top two meters of water (lt2; bottom).

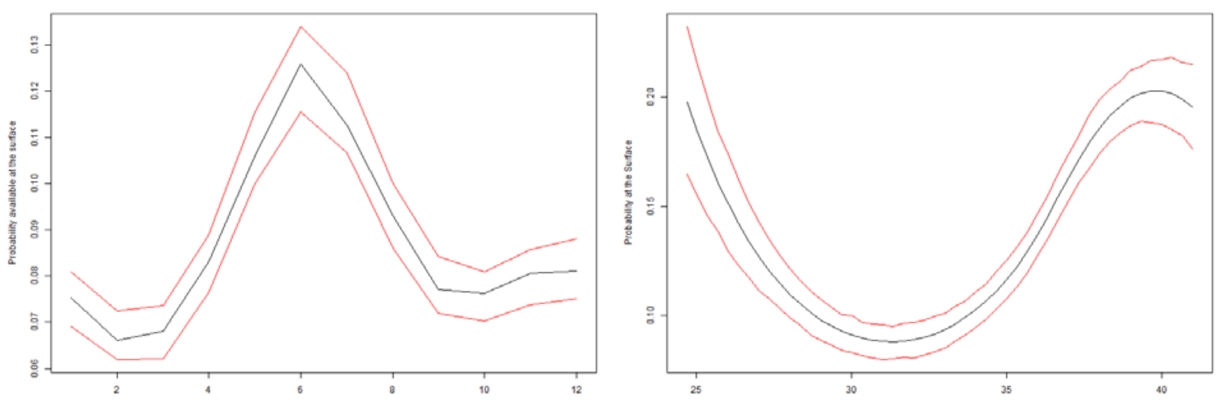

Figure 4: Figures showing predictions for a range of values for each covariate using the surface model. The black line is the mean of 500 bootstraps and the red lines are upper and lower 95 percentile confidence intervals. For month, latitude is fixed at the mean of the prediction region (35.43°N) and for latitude, month is fixed at 6.

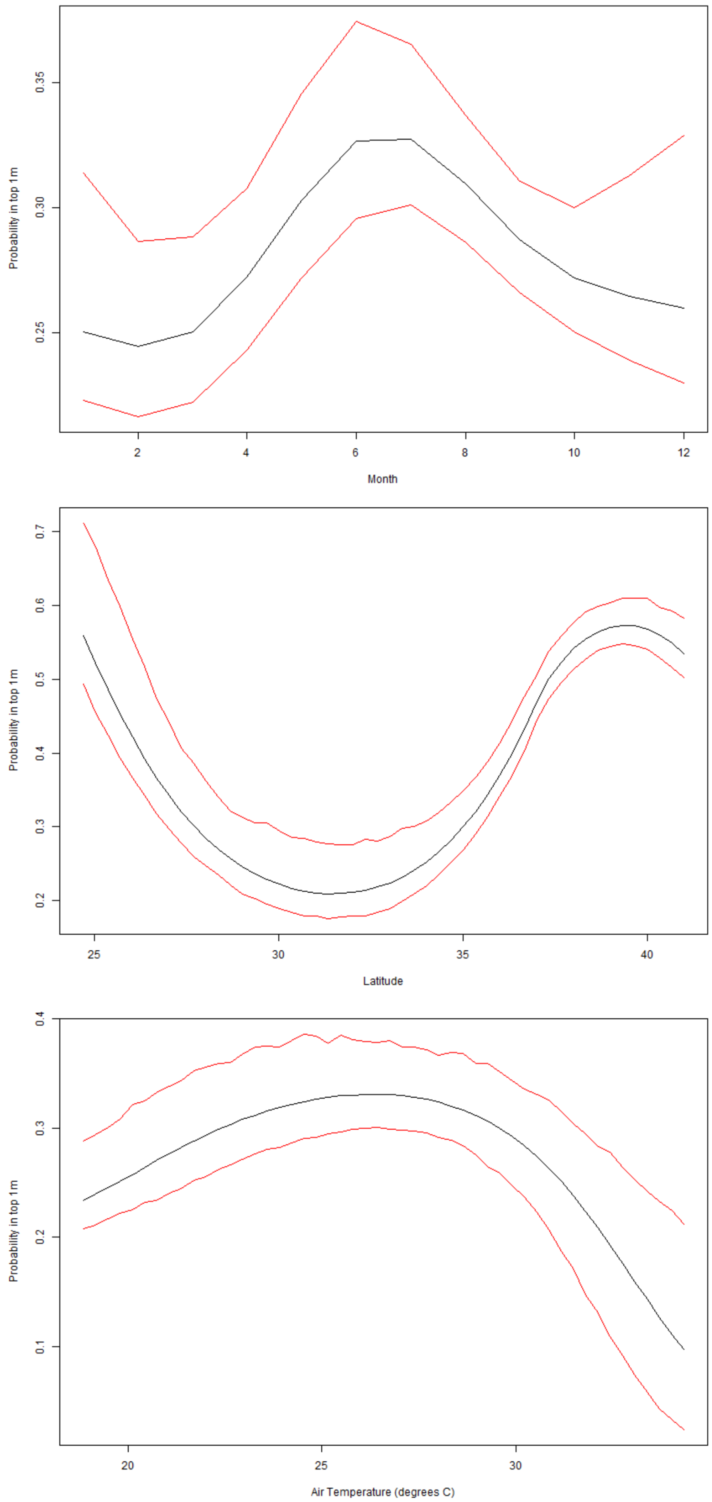

Figure 5: Figures showing predictions for a range of values for each covariate using the LT1 model. The black line is the mean of 500 bootstraps and the red lines are 95 percentile confidence intervals. When non-varying, month was fixed at 6, latitude at the mean of the prediction region (35.43°N) and air temperature the mean of the prediction set (27.7°C).

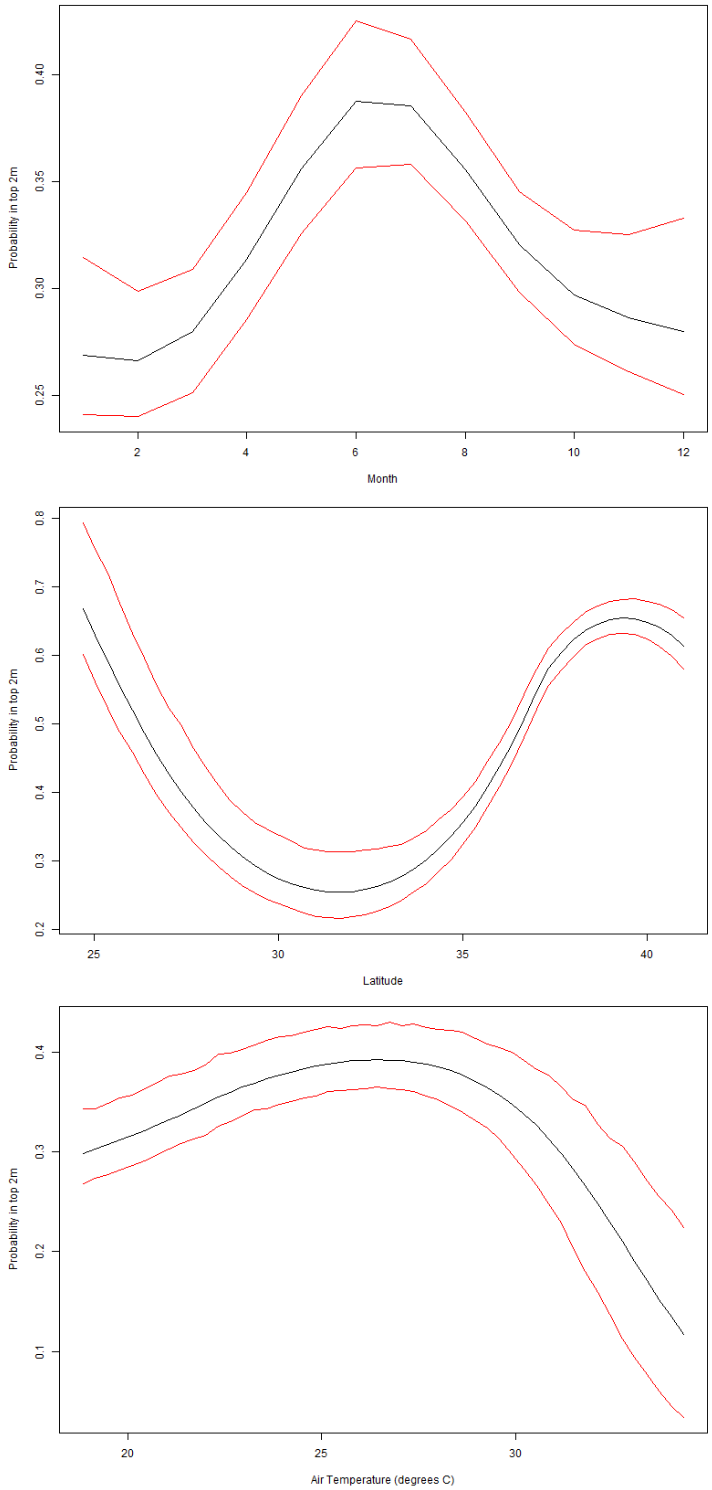

Figure 6: Figures showing predictions for a range of values for each covariate using the LT2 model. The black line is the mean of 500 bootstraps and the red lines are 95 percentile confidence intervals. When non-varying, month was fixed at 6, latitude at the mean of the prediction region (35.43) and air temperature the mean of the prediction set (27.7^{o}C).

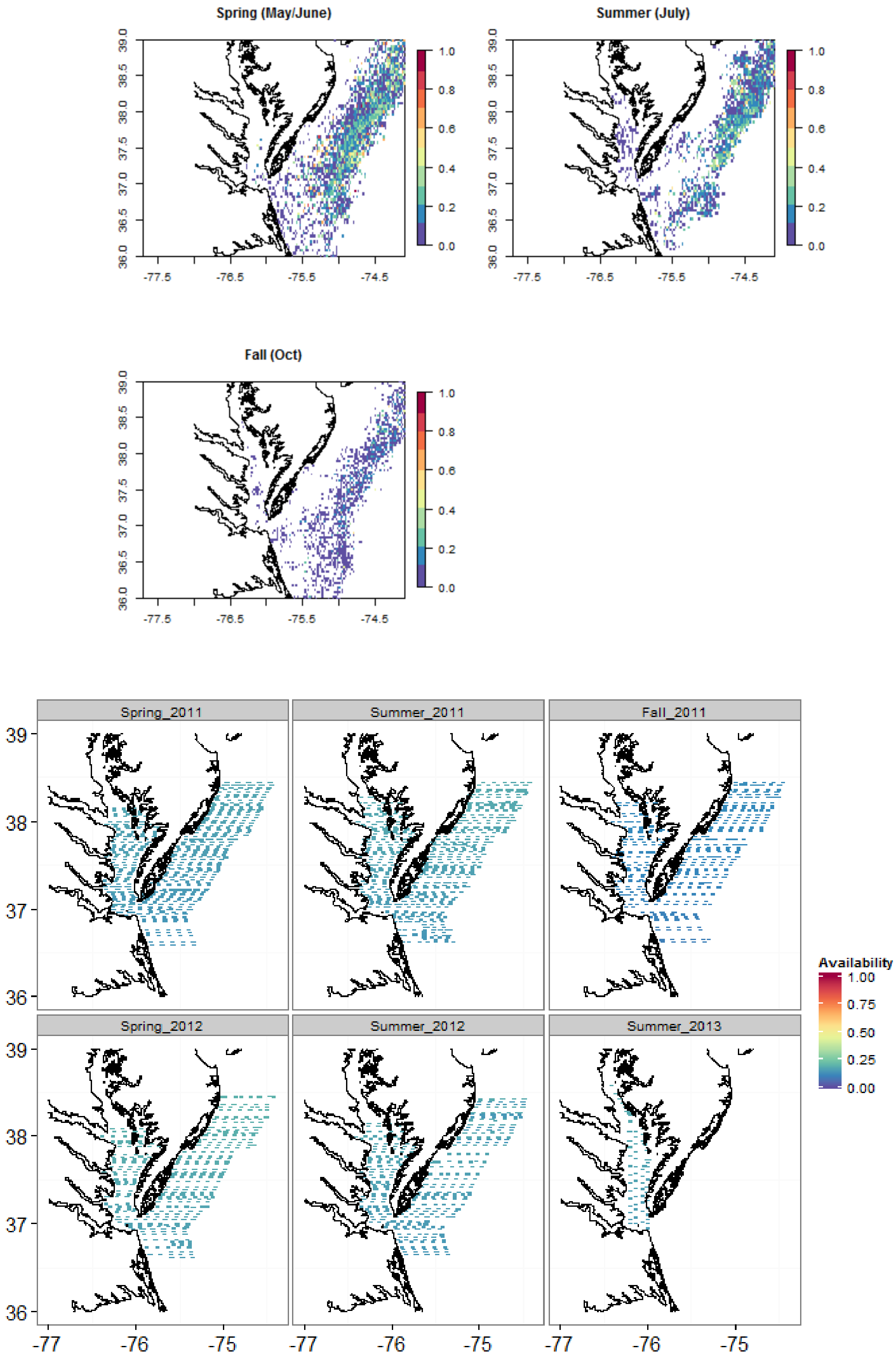

Figure 7: Surface model raw availability data (upper) for seasons in the Virginia Aquarium (VA) prediction data across all years, and predicted availability (lower) for the VA line transect survey data.

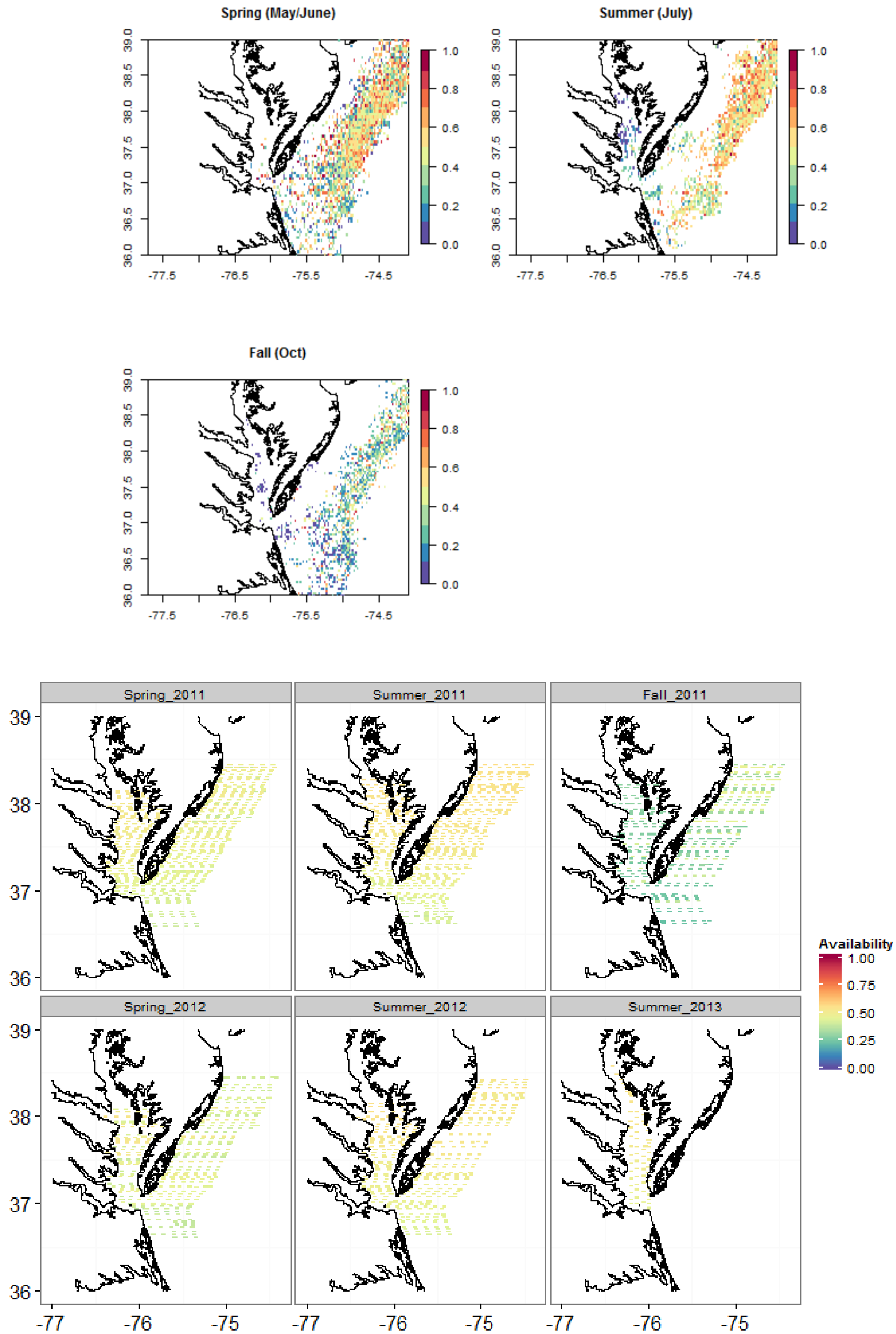

Figure 8: LT1 model raw availability data (upper) for seasons in the Virginia Aquarium (VA) prediction data across all years, and predicted availability (lower) for the VA line transect survey data.

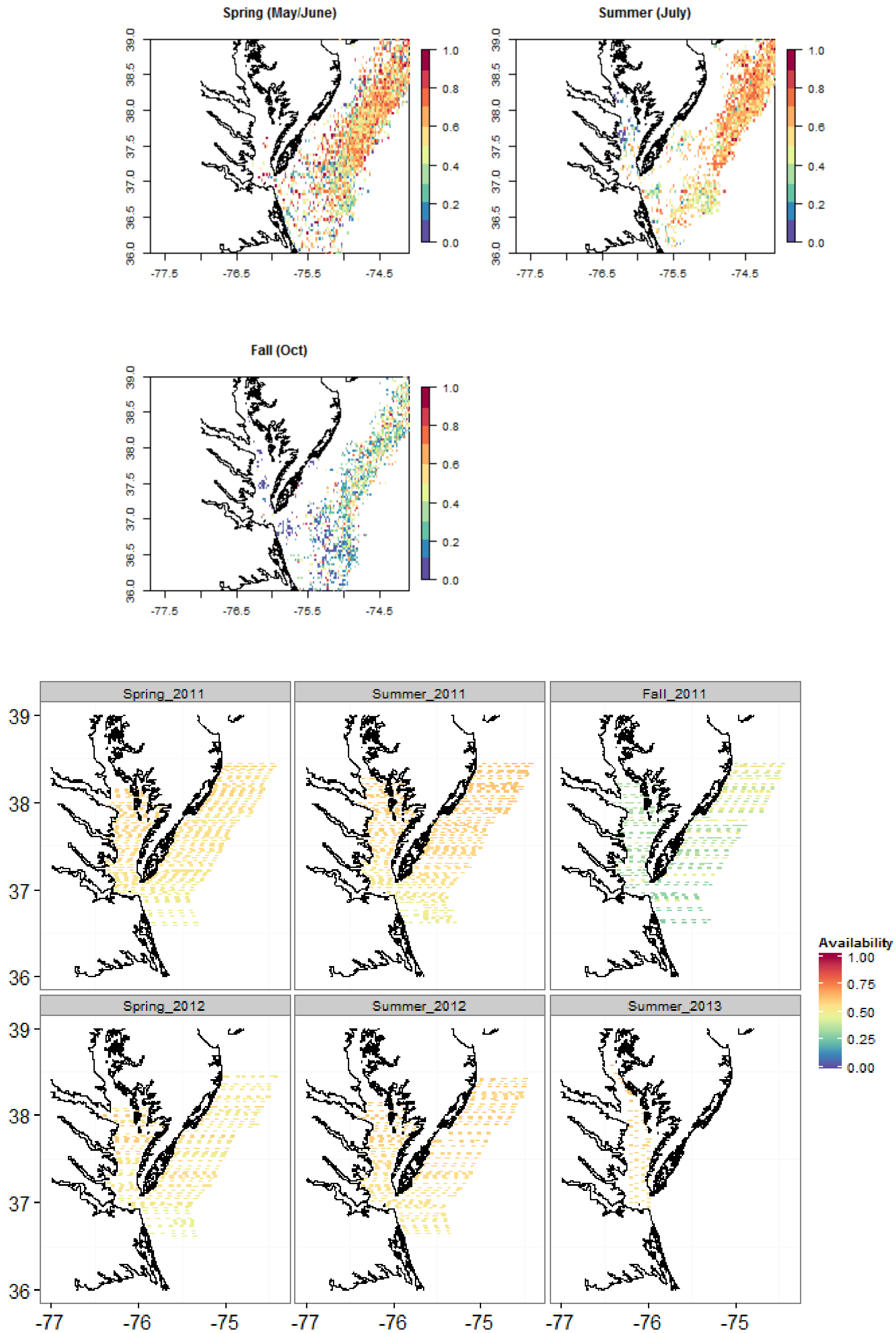

Figure 9: LT2 model raw availability data (upper) for seasons in the Virginia Aquarium (VA) prediction data across all years, and predicted availability (lower) for the VA line transect survey data.

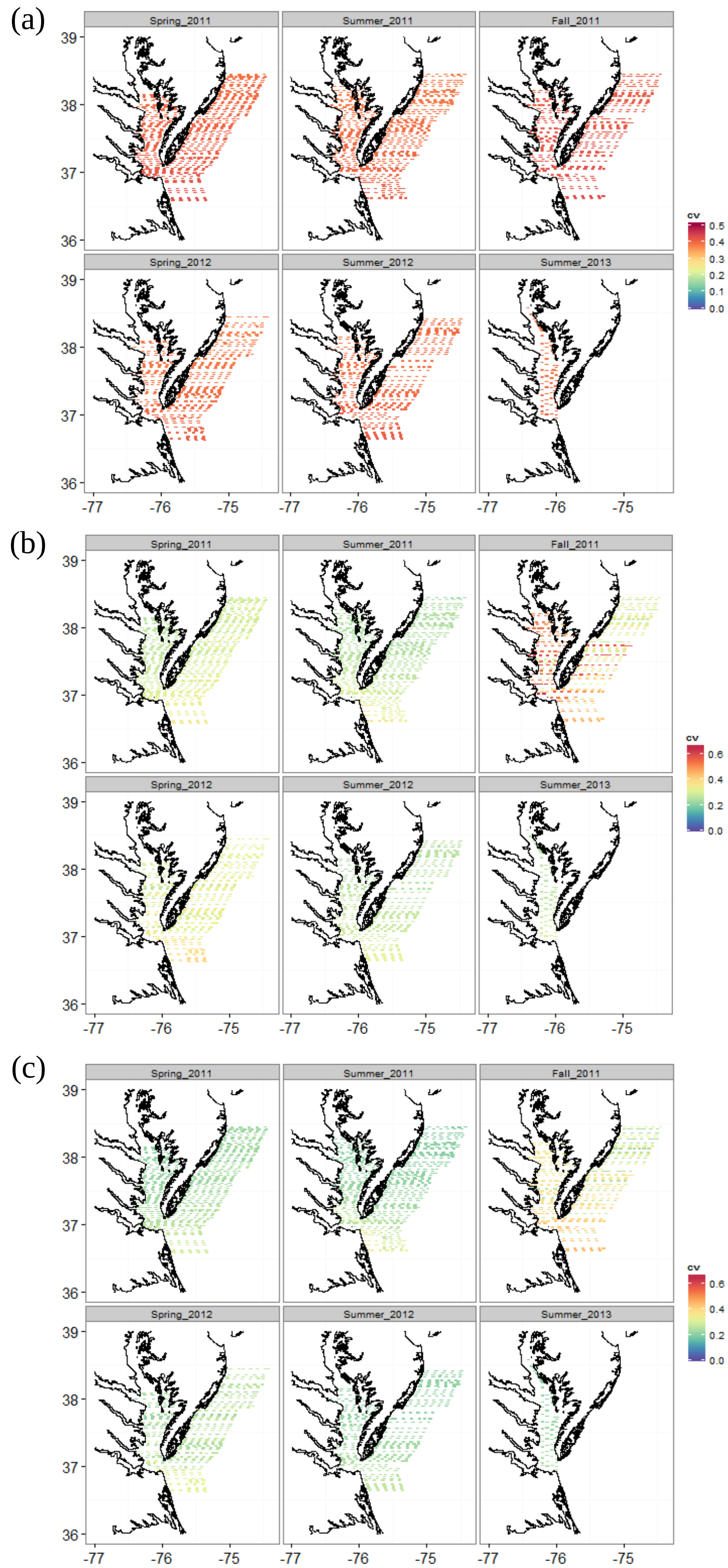

Figure 10: Figures showing CV scores for the Virginia Aquarium line transect surveys for each of the three model types; surface (a), LT1 (b), LT2 (c).

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Table A1: Assessment of different surface models and their AIC/BIC (Bayesian Information Criterion) scores where available. The best (lowest) AIC score model is highlighted in green and was used for the results in this paper. BIC scores are shown but were not used in the selection process. The 'df' column contains the degrees of freedom for each model.

Response	Covariate	df	AIC	BIC	Notes
Surface (zero Inf only)	s(Month)	12	-78508	-78439	
	s(LAT)	12	-77601	-77565	
	s(HySur,HyBot); s(HySur) for 0	16	-77236	-77187	
	s(HySur,HyBot); s(airTC) for 0	16	-77164	-77115	
	LAT	8	-77054	-77030	
	s(airTC)	12	-76612	-76575	
	airTC	8	-76238	-76214	
	s(Raddown)	12	-75932	-75895	
	s(DistC)	12	-75533	-75497	
	Raddown	8	-74822	-74798	
	s(wdepth)	12	-74735	-74699	
	s(HyBot)	12	-74729	-74629	
	s(HySur)	12	-74498	-74461	
	HyBot	8	-74471	-74447	
	HySur	8	-74354	-74330	
	wdepth	8	-74119	-74095	
	DistC	8	NA	NA	
	s(HySur,HyBot)	20	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(Hysur, HyBot, s(HySur))	28	-79653	-79565	hessian full rank but negative eigen value
	s(Month)+s(LAT) + s(airTC)	24	-79502	-79429	won't converge
	s(Month)+s(LAT)	18	-79261	-79193	
	s(Month)+s(LAT) + s(Hysur, HyBot, beta only)	25	NA	NA	optimization not completed (no maxima)
	s(Month)+s(LAT) + s(airTC) + s(Raddown)	30	NA	NA	optimization not completed
	s(Month) + s(LAT) + s(airtc, beta only)	21	NA	NA	
	s(Month) + s(air)	18	-78767	-78713	

Table A2: Assessment of different LT1 models and their AIC/BIC scores where available. The best (lowest) AIC score model is highlighted in green and was used for the results in this paper. The 'df' column contains the degrees of freedom for each model.

Response	Covariate	df	AIC	BIC	Notes
LT1	s(LAT)	19	-25907	-25850	
	s(Month)	19	-25831	-25774	
	s(airTC)	19	-22405	-22347	
	s(HySur,HyBot)	31	-21632	-21539	
	s(HySur,HyBot); s(HySur) for 0/1	23	-21597	-21528	
	s(HySur,HyBot); s(airTC) for 0/1	23	-21482	-21412	
	s(HySur)	19	-16267	-16210	
	s(Raddown)	19	-15697	-15639	
	s(DistC)	19	-15409	-15352	
	s(wdepth)	19	-13424	-13366	
	s(Hybot)	19	-13264	-13207	
	LAT	13	NA	NA	optimization not completed
	Raddown	13	NA	NA	optimization not completed
	airTC	13	NA	NA	optimization not completed
	wdepth	13	NA	NA	optimization not completed
	DistC	13	NA	NA	optimization not completed
	HySur	13	NA	NA	optimization not completed
	Hybot	13	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC)	37	-28960	-28854	
	s(Month)+s(LAT) + s(Hysur, Hybot, s(HySur))	41	-28101	-27974	hessian full rank but negative eigen value even with 1000 iter
	s(Month)+s(LAT) + s(Hysur, Hybot, s(air))	41	-28015	-27888	hessian full rank but negative eigen value
	s(Month)+s(LAT)	28	-27996	-27911	
	s(Month)+s(LAT) + s(Hysur, Hybot, beta only)	35	-27844	-27775	hessian full rank but negative eigen value
	s(Month)+s(LAT) + s(airTC) + s(HySur, beta only)	44	NA	NA	optimization not completed

Table A3: Assessment of different LT2 models and their AIC/BIC scores where available. The best (lowest) AIC score model is highlighted in green and was used for the results in this paper. The 'df' column contains the degrees of freedom for each model.

Response	Covariate	df	AIC	BIC	Notes
LT2	s(Month)	19	-22167	-22110	
	s(LAT)	19	-21954	-21896	
	s(airTC)	19	-17603	-17545	
	s(HySur,HyBot); s(HySur) for 0/1	23	-16459	-16390	
	s(HySur,HyBot); s(airTC) for 0/1	23	-16334	-16265	
	s(Raddown)	19	-11132	-11074	
	s(HySur)	19	-10737	-10680	
	s(DistC)	19	-9979	-9922	
	s(wdepth)	19	-7842	-7784	
	LAT	13	NA	NA	optimization not completed
	Raddown	13	NA	NA	optimization not completed
	airTC	13	NA	NA	optimization not completed
	wdepth	13	NA	NA	optimization not completed
	DistC	13	NA	NA	optimization not completed
	s(HySur,HyBot)	31	NA	NA	optimization not completed
	s(Hybot)	19	NA	NA	optimization not completed
	Hybot	13	NA	NA	optimization not completed
	Hysur	13	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC)	37	-25433	-25331	
	s(Month)+s(LAT)	28	-24624	-24539	
	s(Month)+s(LAT) + s(Hysur, Hybot, s(Hysur))	41	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(Hysur, Hybot, s(air))	41	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(Hysur, Hybot, beta only)	35	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC) + s(Raddown)	46	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC) + s(Raddown) + s(Hysur)	66	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC) + s(Raddown, betaonly)	40	NA	NA	optimization not completed
	s(Month)+s(LAT) + s(airTC) + s(Raddown, betaonly) + s(Hysur, betaonly)	47	NA	NA	optimization not completed

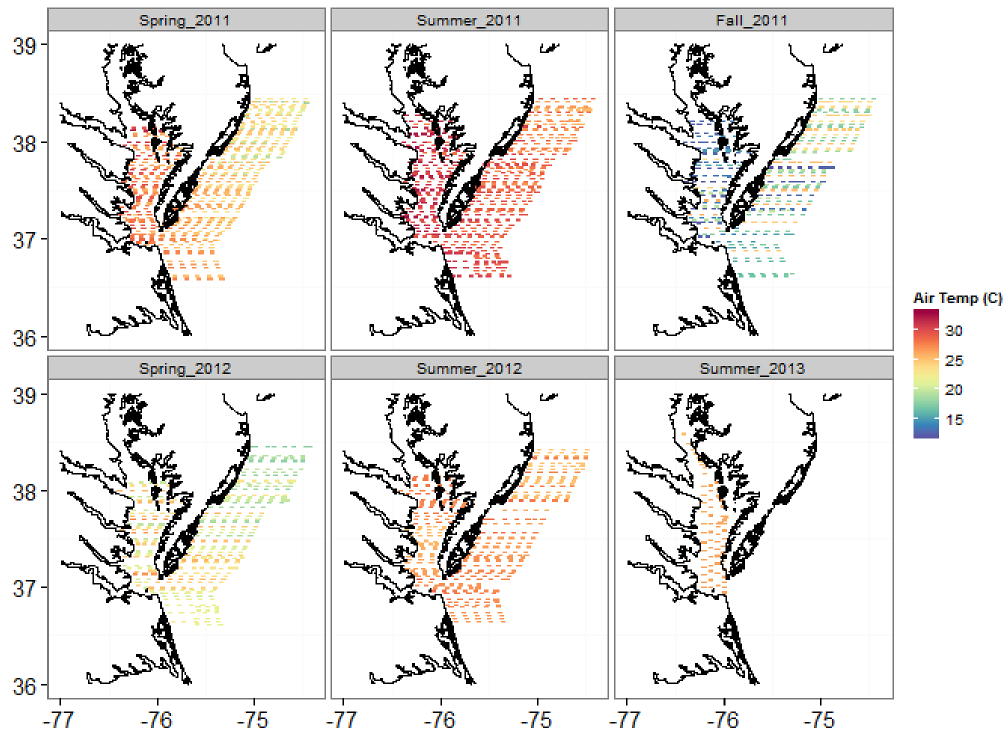

Figure A1: Plots of the air temperature for each of the VA segments across surveys.