

# Nantucket Sound Sea Ducks

*Adam D. Smith, Benjamin Hofner, Jason E. Osenkowski, Taber Allison,  
Scott R. McWilliams, and Peter W. C. Paton*

*15 Juni, 2015*

## **Didn't get around to modifying this introduction framework much based on our previous discussions**

A basic tenet of ecology is that biophysical features at multiple spatial and temporal scales affect the distribution and abundance of organisms and their prey (MacArthur 1972, Andrewartha and Birch 1986, Maurer 1999). Human activities can likewise influence the distribution and abundance of organisms (general citations; we mentioned oil, onshore wind, shale, others?), and quantifying the effect of these activities on animal distribution and abundance requires a flexible approach that accommodates potentially complex associations between animal distributions and environmental covariates. Model-based approaches (cf. design-based approaches) provide this flexibility and, given an appropriate model, facilitate the spatially-explicit estimation of animal abundance and changes in animal distribution independent of abundance (Borchers et al. 2002, see also, e.g., Certain and Bretagnolle 2008, Nur et al. 2011, Kinlan et al. 2012, Winiarski et al. 2014).

Despite increasing anthropogenic impacts on marine ecosystems (**citations; any suggestions**), the environmental factors associated with the distribution and abundance of sea ducks (Tribe Mergini), top consumers in these ecosystems, are notably understudied. Moreover, current evidence suggests that populations of most species of North American sea ducks are below historic levels with negative or indeterminate (given current data) population trends (Merkel 2004, Bowman et al. 2015). Thus there exists considerable interest in identifying the biophysical characteristics of the marine environment that influence the distribution and abundance of sea ducks.

## **Some of this still seems relevant, but maybe we don't need this much ecology?**

At macro-scales (e.g., wintering range), large-scale ocean characteristics (e.g., Flint 2013), climatic conditions (e.g., Zipkin et al. 2010), and static or persistent habitat features (e.g., bathymetry, substrate, current and frontal systems; Hyrenbach et al. 2000) typically associate with large-scale patterns in distribution of marine birds (e.g., Nur et al. 2011). At the micro- and meso-scales of acute anthropogenic disturbances (e.g., energy development, oil spills, fisheries), however, the abundance and distribution of prey, and thus predators, can be extremely ephemeral and dynamic (Hyrenbach et al. 2000, Ortega Cisneros et al. 2011), even relating to prevailing meteorologic conditions (e.g., Garthe et al. 2009), potentially hindering the utility of static or persistent biophysical covariates (or those measured too coarsely in time or space; **e.g. human activity such as ferry/fisheries/etc.?**). Marine birds (sea ducks) distribution is patchy (**citations**); abundance can be extremely variable; often interested in evaluating the relevance of many potential covariates with possibly complex (i.e., nonlinear) associations with animal presence and abundance (**citations**).

Seems a little disingenuous to decry previous methods given the rapid evolution of distribution and abundance modelling, so how about we describe challenges and then suggest our approach as a means of handling these things? Some of these are mentioned above, but the most relevant things are:

- Lots of zeros (Martin et al. 2005), overdispersion (Richards 2008).
- Complex/non-linear associations between species distribution/abundance and the environment
- Other components of the assumed response distribution (e.g., overdispersion), and not just the conditional mean, can vary with covariates response
- The typical necessity of identifying a few important covariates (variable selection) out of many potentially correlated covariates (high dimensionality; collinearity) while avoiding overfitting
- Evaluating competing expectations of associations between covariates and the response (model selection; e.g., linear vs. non-linear effects)
- Spatial and temporal autocorrelation (Hoeting 2009).

We demonstrate a model-based approach to predict animal distribution and abundance that incorporates these considerations using aerial survey data of sea ducks in Nantucket Sound, Massachusetts, USA. We implemented generalized additive models (GAMs; Hastie and Tibshirani 1990, Wood 2006) and their recent extension (GAM for location, scale, and shape; GAMLSS; Rigby and Stasinopoulos 2005) in a gradient descent boosting framework (Hothorn et al. 2010, Mayr et al. 2012, Hofner et al. 2015b). GAMLSS methods extend GAMs to allow all parameters of the conditional response distribution to be modeled as a function of relevant covariates (Rigby and Stasinopoulos 2005); this approach thus grants insight not only into the covariates associated with sea duck distribution and abundance but also their variability. The boosting algorithm accommodates the inclusion of various effect types (e.g., linear, smooth, spatial, spatiotemporal, and random effects; Hofner et al. 2014c) and correlated covariates (i.e., parameter estimates are regularized), identifies the most relevant subset among potentially many covariates (i.e., variable selection), and evaluates competing representations of continuous covariates (e.g., linear vs. nonlinear effects of covariates; Kneib et al. 2009, Maloney et al. 2012).

## Methods

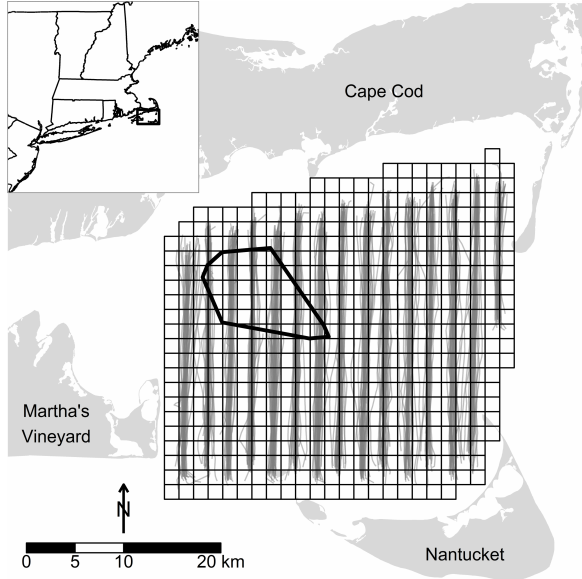
***Study area and species.***—Maybe a short paragraph regarding Nantucket Sound’s importance in particular, the Atlantic Coast in general, and the plight of sea ducks fits here rather than in the introduction given the current focus?

*Aerial strip transect surveys.*—During the winters (late October to mid-April) of 2003-2005, we conducted 30 (2003/2004: 13, 2004/2005: 10, 2005/2006: 7) standardized aerial strip-transect surveys (e.g., Pollock et al. 2006, Certain and Bretagnolle 2008) throughout an 1,100 km<sup>2</sup> study area in Nantucket Sound, Massachusetts, USA (Figure 1). Surveys occurred

primarily from November - March ( $n = 27$ ), with an occasional October ( $n = 1$ ) or April ( $n = 2$ ) survey. During each survey, we flew along 15 parallel (ca. 2.3 km apart), roughly north-south transects (Figure 1) using a high-wing, twin-engine aircraft (Cessna Skymaster 337) at an average altitude of 152 m and speed of 167 kmh<sup>-1</sup> (90 kts). This altitude allowed us to identify most birds at the sea surface and reduced sea disturbance (i.e., flushing birds to another part of the study area and, thus, potential double counting). The airspeed was the slowest at which the aircraft could safely fly. Surveys occurred only on days with wind speeds  $\leq 15$  kts and good visibility ( $>15$  km). Surveys occurred between 0900 to 1600 h (average duration of  $\sim 2.5$  h) to ensure that birds had completed their post-dawn movements (e.g., Davis 1997) but had yet to initiate pre-sunset movements from feeding to roosting areas; this time window also reduced glares due to low sun angles.

On each survey flight, two observers used their unaided eyes to continuously detect individuals or flocks, identifying sea ducks to species with the aid of binoculars as needed. Observers monitored the sea surface on their side of the plane in a  $\sim 91$  m wide transect between  $\sim 56$  - 147 m from the plane; observers could not see beneath the plane out to 56 m on each side. The narrow strip width ensured birds were detectable and identifiable with the naked eye and allowed us to avoid situations in which ducks were too abundant or spread over too wide an area to count accurately. This last point was especially relevant because sea ducks occasionally congregated in extremely large aggregations. We thus assumed inconsequential perception bias (i.e., few missed individuals present to be counted; Marsh and Sinclair 1989) given that our flight altitude and narrow transect width (e.g., Certain and Bretagnolle 2008) within the strip transect (Buckland et al. 2012).

Observers verbally communicated duck sightings to a recorder, who entered the number of ducks, species, behavior (i.e., on the water or flying), and geographic location into a laptop computer linked to the plane's onboard GPS; we report results only from ducks observed sitting on the water. Sea duck species of interest included Common Eider (*Somateria mollissima*; hereafter eider), White-winged Scoter (*Melanitta deglandi*), Surf Scoter (*M. perspicillata*), Black Scoter (*M. americana*), and Long-tailed Duck (*Clangula hyemalis*). We considered scoter species collectively (hereafter, scoter), as it was regularly difficult to identify scoter to species. We subsequently consolidated counts for each species (eider and Long-tailed Duck) or species group (scoter) into 2.25km<sup>2</sup> segments (Figure 1); this resolution (1.5 km x 1.5 km) corresponded approximately to the coarsest level of resolution of biophysical covariates (see below).



**Figure 1.** Actual aerial strip transect tracks (gray lines) during winter (October - April, 2003 - 2005) sea duck surveys ( $n = 30$ ) in Nantucket Sound, Massachusetts, US. The grid indicates the extent of the 1100 km<sup>2</sup> study area and its division into 504 2.25km<sup>2</sup> segments. The polygon in northwest Nantucket Sound indicates the 62 km<sup>2</sup> area of permitted wind energy development on Horseshoe Shoal.

*Modeling approach.*—We related spatiotemporal variation in sea duck occupancy (i.e., probability of presence) and abundance to potentially relevant biophysical and spatiotemporal covariates. We applied a negative binomial hurdle model that separately modeled (1) the probability of occurrence of at least one individual (hereafter, the occupancy model) in a given segment using a logistic regression model and (2) the abundance of sea ducks in that segment conditional on their presence (hereafter, the count model) using a truncated negative binomial model. The prevalence of zero counts (e.g., 75% of eider segment observations) prompted the use of a hurdle model. We generated separate hurdle models for each sea duck group (i.e., eider, scoter, and Long-tailed Duck; Zipkin et al. 2010).

To model sea duck occupancy, we used a generalized additive model (GAM; Hastie and Tibshirani 1990, Wood 2006) that flexibly accommodated potential nonlinearities between covariates and sea duck occupancy. We used generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos 2005) to fit the count models; these models accommodated potential nonlinearities between covariates and the conditional mean and overdispersion of sea duck abundance.

We implemented GAMs and GAMLSS in a gradient descent boosting framework (Mayr et al. 2012, Hofner et al. 2015b). More specifically, we used component-wise functional gradient descent boosting to fit the models. Briefly, one begins by computing the negative gradient of a user-specified loss function (typically the negative log-likelihood of the underlying model); the negative gradient can be viewed as a proxy for residuals. We used the negative binomial log-likelihood as the loss function for occupancy models and the (negative) truncated negative binomial log-likelihood as the loss function for count models. In the next step, all modeling alternatives (i.e., functional forms of the covariates relative to the response, called base-

learners; Hofner et al. 2014c) are fitted separately to the negative gradient and a fraction of the best fitting base-learner is added to the current model fit. The negative gradient is then reevaluated at the current model fit and the procedure iterated until for a fixed number of iterations,  $m_{stop}$ , specified by the user (see Bühlmann and Hothorn 2007 for additional details). Boosted GAMLSS models are fitted similarly, although in each iteration the negative gradient is computed separately with respect to each GAMLSS parameter (e.g., mean and overdispersion) while holding other parameters as fixed constants. Base-learners are then fitted to the resulting parameter-specific negative gradient vector and model updates are computed separately for each GAMLSS parameter as for boosted GAMs (see Mayr et al. 2012 for additional details).

As only the single best-fitting base-learner is selected in each iteration, the algorithm thus integrates intrinsic selection of the most relevant covariates and their functional form (i.e., some base-learners may never be selected). Variable selection is further fostered by stopping the algorithm prior to convergence to maximum likelihood estimates (early stopping; Mayr et al. 2012, Maloney et al. 2012). Furthermore, early stopping aims at maximizing predictive accuracy while avoiding model overfitting. We used 25-fold subsampling to determine the optimal stopping iteration for each model. Specifically, we randomly drew (without replacement) 25 samples of size  $n/2$  from the original data set. We used the selected sample to estimate the model and the balance of the data in each sample to determine the out-of-bag prediction accuracy (empirical risk) measured by the negative log-likelihood of each model; the optimal stopping iteration ( $\hat{m}_{stop}$ ) is the iteration with the lowest average empirical risk. In boosted GAMLSS models we used multi-dimensional subsampling to determine the stopping iteration for each of the GAMLSS parameters while allowing for potentially different model complexities in the parameters; a detailed explanation of this cross-validation (subsampling) scheme is given in Hofner et al. (2014b).

Despite these agreeable features, boosting methods typically produce “rich” models relying to some extent on many base-learners (Hofner et al. 2015a). Thus, we additionally applied stability selection (Meinshausen and Bühlmann 2010, Shah and Samworth 2013) to identify those base-learners, and thus covariates, most commonly selected in each model (i.e., selected on the majority of 100 random subsamples of the data); see Appendix 1 and Hofner et al. (2015b) for more details.

*Covariates.*—We evaluated biophysical covariates expected to influence the distribution, abundance, and movements of sea ducks or, more likely, the distribution and availability of their benthic prey (e.g., mollusks and crustaceans; see Appendix 2); we did not have information related directly to the distribution of preferred prey. Biophysical covariates could be characterized as strictly spatial or temporal effects (i.e., varying only among segments or within or among winters, respectively) or as spatiotemporal effects (i.e., varying among segments and within or among winters; see Appendix 2). Additionally, we included interactions that allowed the effects of two covariates to vary over time within a given winter (see Appendix 2). We standardized (i.e., mean centered and scaled) all continuous covariates. Appendix 3 describes a function to visualize the spatial and temporal distribution of covariates in Nantucket Sound.

For each continuous covariate, we specified two base-learners: a linear base-learner and a

penalized nonlinear base-learner (i.e., P-splines; Eilers and Marx 1996, Schmid and Hothorn 2008). We dummy-coded categorical covariates, resulting in  $K - 1$  linear base-learners for a  $K$ -level covariate. To address potential spatial autocorrelation, we included a smooth surface function of the spatial coordinates of segment centers (Kneib et al. 2008); this surface comprised four base-learners – linear base-learners for the easting and northing, their linear interaction, and a penalized nonlinear tensor product P-spline (Kneib et al. 2008, 2009, Maloney et al. 2012). We also allowed this surface to vary over (interact with) time within a winter. Hofner et al. (2014c) describe base-learners in detail.

The decomposition of continuous covariates into centered linear and penalized orthogonal nonlinear base-learners allowed us to weight base-learners equally and thus allowed unbiased model choice (i.e., prevented the preferential selection of smooth base-learners; Kneib et al. 2009, Hofner et al. 2011). Specifically, we restricted each base-learner to a single degree of freedom and omitted the intercept term from each base-learner (Kneib et al. 2009, Hofner et al. 2011). Consequently, we added an explicit linear base-learner for the model intercept. The occupancy and count models had the following structure (see Appendix 2 for descriptions of abbreviated covariates):

$$\begin{aligned}
g(\cdot) = & \text{int} + \text{time} + f(\text{time}) + \text{SSTw} + f(\text{SSTw}) + \text{SSTm} + f(\text{SSTm}) + \text{SSTrel} + \\
& f(\text{SSTrel}) + \text{SSTrel} \cdot \text{time} + f(\text{SSTrel}, \text{time}) + \text{SBT} + f(\text{SBT}) + \text{NAOw} + \\
& \text{depth} + f(\text{depth}) + \text{depth} \cdot \text{time} + f(\text{depth}, \text{time}) + \text{d2land} + f(\text{d2land}) + \\
& \text{chla} + f(\text{chla}) + \text{cdom} + f(\text{cdom}) + f(\text{cdom}, \text{chla}) + \text{meanphi} + f(\text{meanphi}) + \\
& \text{SAR} + f(\text{SAR}) + \text{tidebmean} + f(\text{tidebmean}) + \text{tidesd} + f(\text{tidesd}) + \text{strat} + \\
& f(\text{strat}) + \text{ferry} + \text{y2004} + \text{y2005} + \text{xkm} + \text{ykm} + \text{xkm} \cdot \text{ykm} + \\
& f(\text{xkm}, \text{ykm}) + \text{xkm} \cdot \text{time} + \text{ykm} \cdot \text{time} + \text{xkm} \cdot \text{ykm} \cdot \text{time} + \\
& f(\text{xkm}, \text{ykm}) \cdot \text{time} + \text{obs\_window} + f(\text{obs\_window}).
\end{aligned} \tag{1}$$

In our occupancy models,  $g(\cdot)$  is  $g(\pi_{\text{sea ducks}})$ , the occupancy probability of a given duck species in a segment and  $g$  is the logit link. In our count models,  $g(\cdot)$  took two forms within the GAMLSS framework – the (conditional) mean count of sea ducks,  $g(\mu_{\text{sea ducks}})$ , and the (conditional) overdispersion in sea duck counts,  $g(\sigma_{\text{sea ducks}})$ ;  $g$  is the log link in both cases. Base-learners denoted as  $f(\cdot)$  indicate the penalized nonlinear deviations from the corresponding linear base-learner (e.g.,  $f(\text{time})$ ). The explicit intercept ( $\text{int}$ ) was a necessary byproduct of our decomposition of base-learners (see above; Kneib et al. 2009, Hofner et al. 2011). Small values of  $\text{obs\_window}$ , our measure of survey effort (see Appendix 2), in some segments impaired the estimability of models including  $\text{obs\_window}$  as an offset; we thus included it as a covariate.

To evaluate the approximate explanatory power of our final models, we calculated a pseudo  $R^2$  measure of the explained variation (Nagelkerke 1991, Maloney et al. 2012). **Benjamin: I think I've correctly adapted Maloney et al.'s (2012) code; see pseudoR2.R. I calculate the pseudo  $R^2$  only from final model rather than multiple times in bootstrap samples. Can you possibly check my calculations of the log-likelihoods?**

All analyses were conducted in R (Version 3.1.3; R Core Team 2014) with the add-on packages *gamboostLSS* (Hofner et al. 2015b), *mboost* (Hothorn et al. 2015), and *stabs* (Hofner and Hothorn 2015). The data and code for reproducing this manuscript and analyses are given as an online electronic supplement at *TBD*.

## Results

As the occupancy and the hurdle model are independent, we fitted two models per sea duck species. Bootstrapped empirical risk suggested that occupancy models for all species converged on the maximum likelihood estimates (i.e., occupancy models failed to stop early; see Appendix 4). Conversely, bootstrapped empirical risks prescribed early stopping for both parameters (i.e., conditional mean and overdispersion) in all count models (see Appendix 4). Consequently, final occupancy models generally contained more covariates than did count models, particularly among stably selected covariates and their interactions. This suggests that the effects of the environmental variables on sea duck occupancy were more complex than their effect on sea duck abundance. Nonetheless, final occupancy and count (conditional mean) models included only a subset (12% to 38%) of the 48 base-learners initially specified for selection (see equation 1). This translated into the selection of only 8-10 (of 23) covariates and their interaction in occupancy models and 3-6 covariates and their interactions in conditional count (mean) models (Figure 2).

*Sea duck occupancy.*—The suite of stably selected covariates associated with occupancy was relatively consistent among sea duck species (Figure 2). When interpreting Figure 2, the influence of univariate effects (i.e., lines) on the response is reflected in the range of the effect over the Y-axis and, due to standardization, can be compared among species and covariates within a model. For example, monthly sea surface temperature ( $SST_m$ ) associated more strongly with eider occupancy than did distance to land ( $d2land$ ) because it spanned a larger range of the Y-axis (Figure 2). However, monthly sea surface temperature ( $SST_m$ ) associated much more strongly with occupancy of Long-tailed Duck than eider for the same reason (Figure 2). Covariate interactions, illustrated with bivariate plots, are similarly comparable within a model. Only the general association (i.e., positive or negative) with the additive predictor is given for factor variables. Comparing univariate, bivariate, and categorical effects is accomplished using the detailed covariate plots for eider, scoter, and Long-tailed Duck (Appendices 5 - 7, respectively); these plots are ordered in descending order of effect size within a model, and provide values for the partial contribution to the additive predictor for covariates and their interactions.

Spatiotemporal effects (i.e., occupancy associated with the  $xkm$ - $ykm$  location of segments and the change over time within winter) were the dominant explanatory feature in sea duck occupancy models, although the spatiotemporal patterns of occupancy varied considerably among species (Figure 2). Sea duck occupancy increased, but at a decreasing rate, with survey effort ( $obs\_window$ ) in a given segment (Figure 2). Among biophysical covariates, sea duck occupancy increased at intermediate monthly sea surface temperature ( $SST_m$ ), particularly (and at somewhat warmer temperatures) in Long-tailed Ducks. Sea duck occupancy also increased modestly at greater distances from land ( $d2land$ ), although eider occupancy was

lowest at intermediate distances. Scoter and especially eider occupancy was greater in areas with coarser sediments (i.e., smaller *meanphi*; Figure 2). Furthermore, eider occupancy associated negatively with chromomorphoc dissolved organic material (*cdom*) and positively with sea floor surface:planimetric area (*SAR*), our measure of the topographic variability of the sea floor (Figure 2); scoter occupancy likewise related to *SAR* and *cdom*, but in the opposite direction in both cases (Figure 2). Scoter occupancy was modestly greater in deeper waters (*depth*); Long-tailed Duck occupancy was greatest in shallow waters early in the winter but in deeper waters later in the winter (Figure 2). Other effects were relatively minor and inconsistent among species.

Spatially-explicit maps of median occupancy over the course of the study (Figure 3, top row) convey the dominant influence of strictly spatial (i.e., segment *xkm-ykm* location) effects. The result is quite distinct patterns of occupancy among species despite the relative similarity of occupancy associations with biophysical covariates (Figure 2). Occupancy was typically highest for eider in northwest and southwest Nantucket Sound, in interior Nantucket Sound for scoter, and in northeast and south Nantucket Sound for Long-tailed Duck (Figure 3, top row). All species relatively avoided the western edge of the Sound northeast of Martha's Vineyard (see Figure 1). Generally, the areas of highest occupancy exhibited the lowest relative variability (Figure 3, bottom row), defined as the median absolute deviation (MAD) of occupancy relative to median occupancy within a segment (a measure analogous to the coefficient of variation).

*Sea duck conditional abundance and overdispersion.*—In contrast to occupancy, fewer covariates associated consistently with the abundance of sea ducks, conditional on their presence, and the suite of stably selected effects varied more markedly among sea duck species (Figure 2; see also Appendices 5 - 7). However, most selected variables fell within the suite of covariates selected in occupancy models. The primary exception to this pattern was the increase in eider and scoter abundance, conditional on presence, associated with relatively warm or cool sea surface temperatures (*SST<sub>rel</sub>*; Figure 2). Spatial effects (*xkm-ykm*) were again the dominant explanatory feature in scoter and Long-tailed Duck, but they were not selected in the eider model (Figure 2). For a given species, biophysical covariates selected in occupancy and count models exhibited broad agreement with the exception of sediment grain size (*meanphi*) in scoter and more complex relationships between eider conditional abundance and dissolved organic material (*cdom*) and sea floor topography (*SAR*; Figure 2).

Spatially-explicit patterns of median conditional abundance (Figure 4, top row) did not necessarily reflect patterns of occupancy (Figure 3, top row). Some areas of Nantucket Sound exhibited mutually high conditional abundance and occupancy for a given species (e.g., eider in the southwest, scoter in the interior, and Long-tailed Ducks in parts of the northeast). In some cases, however, conditional abundance was low despite relatively high occupancy (e.g., eider in the northeast and Horseshoe Shoal, scoter in the northeast and southeast, and Long-tailed Duck along the northern margin). Conversely, other areas of Nantucket Sound exhibited lower occupancy but sea ducks, when present, were more abundant (e.g., eider along the eastern margin, and scoter and Long-tailed Duck in the southwest). As in occupancy models, sea ducks were relatively absent from the middle western margin of Nantucket Sound (i.e., northeast of Martha's Vineyard; see Figure 1). In contrast to sea duck



occupancy, however, areas of highest conditional sea duck abundance typically exhibited the highest relative variability over time (Figure 4, bottom row).

Overdispersion in conditional sea duck abundance also varied with biophysical covariates, although there was less consistency in the associated covariates among species (Figure 2; see also Appendices 5 - 7). Variability (i.e., overdispersion) in sea duck counts was heterogeneous in space (Figure 5, top row) and time (Figure 5, bottom row) in Nantucket Sound, particularly for eider and scoter (as indicated by the magnitude of the overdispersion parameter values). **I still prefer this figure in an Appendix, but if you feel strongly that it should also be in the main text I'll drop Appendix 8.** *BH: If you ask me I can live with both. But for a concise paper I think it would be beneficial to move this to an appendix*

**What else re: overdispersion? This heterogeneity contrasts sharply with the typical assumption in overdispersed (negative binomial) count models – that overdispersion is fixed – and suggests that in practice it is useful to consider that other parameters of the distribution (e.g., the overdispersion parameter) and not simply the conditional mean, depend on covariates. In other words, and in cases similar to the current study, certain biophysical/ecological features might influence the abundance of animals/organisms and the variability in those counts relative to the expectations of the underlying model (negative binomial in this case).** [Maybe this is better discussion?] *BH: Same here. I think this is a very valuable discussion. I also like your exposition very much. Put it where it fits well.*

*Expected sea duck abundance.*—We consolidated occupancy and conditional count models (see, e.g., Equation 6 in Zeileis et al. 2008) to estimate expected (average) sea duck abundance in the study area over the survey period (Figure 6, top row). Final models of expected sea duck abundance explained moderate amounts of variation in observed counts of eider, scoter, and Long-tailed Duck (pseudo  $R^2 = 0.31, 0.48, \text{ and } 0.32$ , respectively). Conditional abundance (Figure 4) strongly influenced the spatially-explicit patterns of expected abundance. Sea duck species exhibited relatively distinct patterns of abundance in Nantucket Sound. Eider were most abundant, and relatively consistently so over time, in southwestern Nantucket Sound (Figure 6); eider were also relatively abundant in northeastern part of the sound but less consistently so as evidenced by the relatively high MAD/median abundance over time (Figure 6, bottom row). Scoter were also most abundant, occasionally extremely so, in southwestern Nantucket Sound, although this was also the area of highest relative variation in scoter abundance (Figure 6); relatively high abundances of scoter also occurred in interior Nantucket Sound (Figure 6). Long-tailed Ducks were consistently most abundant in northeastern Nantucket Sound, as well as along its southern margin (Figure 6). No species' highest abundances occurred in the permitted Nantucket Shoal area, although expected eider and scoter abundances were consistently elevated in some parts of the Shoal (west and southeast, respectively; Figure 6).

**Here are some thoughts about the kind of information we could present relevant to estimation of total abundance. Pardon the poor prose; I simply wrote my train of thought since we haven't finalized our focus in this section.**

Summing the spatially explicit estimates of expected sea duck abundance (Figure 6) provides

an estimate of total animal abundance in the study area. However, we compare the total number of each sea duck species observed in aerial strip transects with the corresponding total (summed) expected abundance for each of the 30 aerial surveys (Figure 7).

Maybe this goes in discussion?: We elected not to estimate of total expected abundance in Nantucket Sound for several reasons. *BH: Didn't you do this just above when you derived the R2? Do I understand you correctly here that the difference is that you depend on the transect for figure 7 and wouldn't depend on the transect here?* First, unknown and unquantified availability bias (e.g., ducks diving during survey; Marsh and Sinclair 1989) possibly resulted in underestimated sea duck occupancy (aggregating observations into 2.25 km<sup>2</sup> segments likely reduced the impact on occupancy) and almost certainly underestimated abundance. Second, expected abundance changed dynamically over time (see next section or not, if in discussion). Finally, the moderate amounts of variation explained by the fitted models suggest the estimates would be imprecise had we incorporated the code to estimate them. *BH: I do not understand the last sentence. Sorry.*

Some reasoning behind the approach:

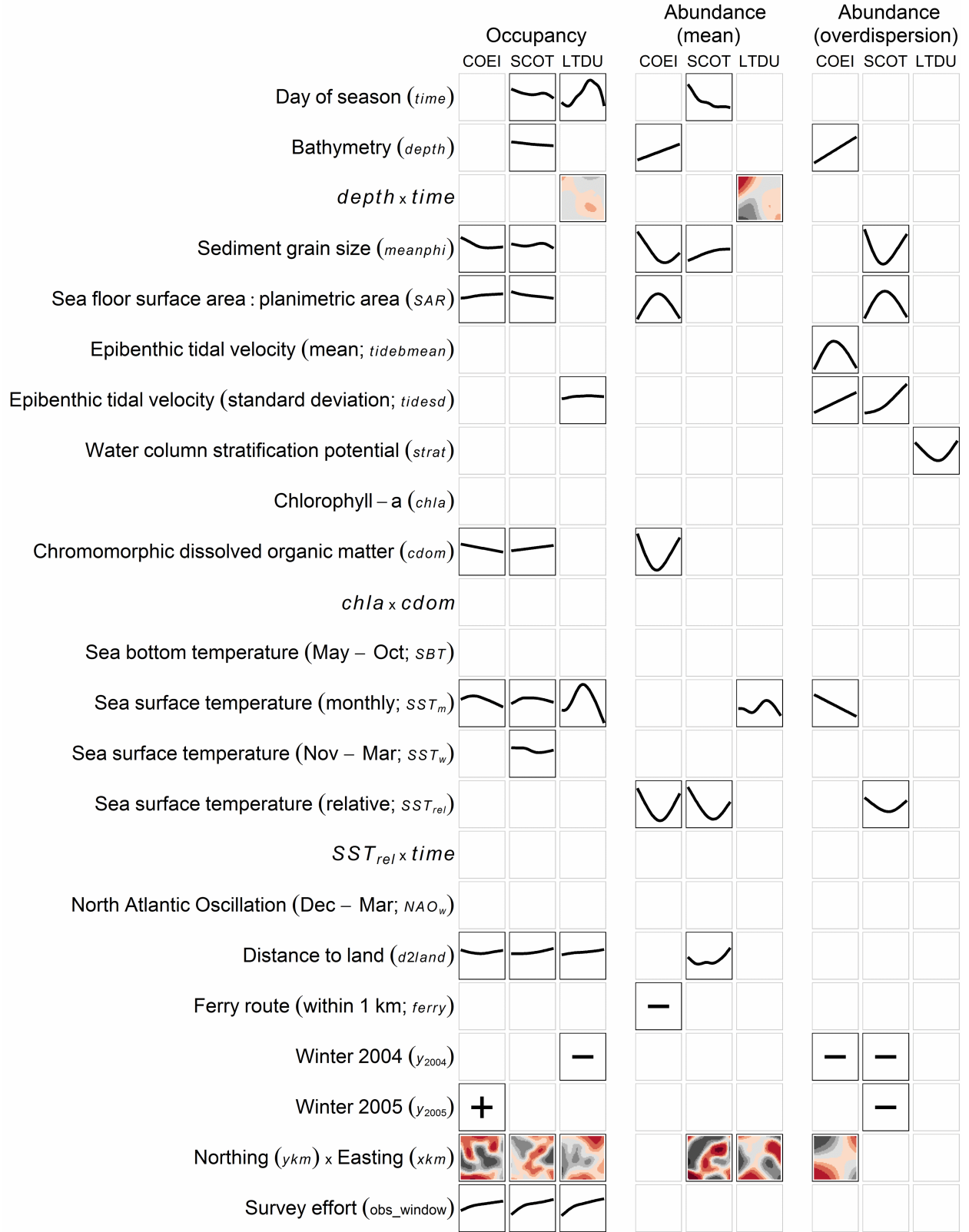
- I restricted the comparison to surveyed segments because (1) there was more variation in the completeness of surveys than I remembered and doing so allowed me to keep all 30 surveys, and (2) the comparison seems more concrete/relevant than extrapolating from the aerial survey to the whole study area and comparing *that* assumption-ridden estimate with another estimate.

Some notes on/observations from the figure:

- The increasing variation with abundance is expected. Counts and the distributions used to model them expect this feature. We could show on a log scale and this would largely vanish but the interpretation is more abstract.
- Scoter abundance is occasionally extreme relative to typical counts (but we knew this).
- The model tends to overestimate sea duck abundance when actual counts are relatively low.
- Eider abundance seems a bit prone to overestimation; scoter abundance seems a bit prone to underestimation.
- We should be careful in making comparisons of (relative) abundance among species because we haven't quantified potential differences in availability bias (e.g., diving rates, commuting of LTDU, visibility differences, etc.).
- We could possibly partition these predictions by winter, but I'm afraid that will take more space to explain (without good justification) and make the figure more complicated

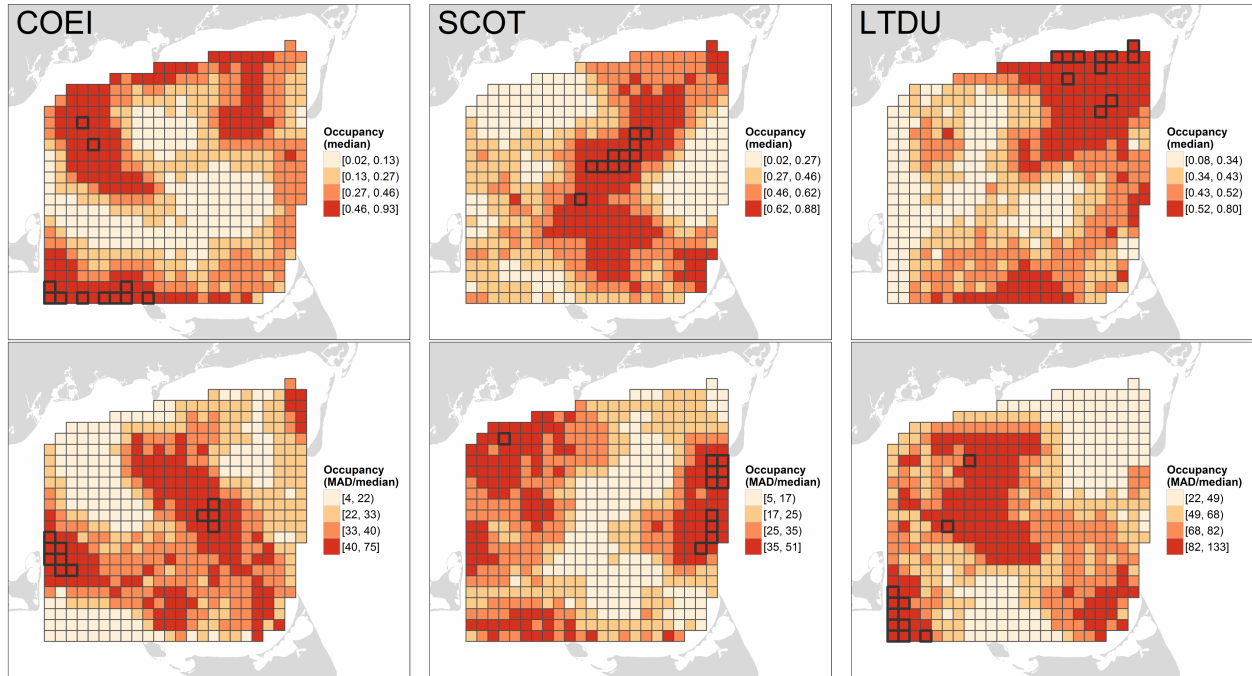
**BH: I think it is a good display as is: The only thing one could consider to use a log scale for plotting as you suggested. But this is not a must in my eyes.**

*Temporal dynamics in wintering sea ducks.*—The MAD/median maps (bottom rows of Figures 3-5) emphasize something important but perhaps not obvious – that our spatially-explicit estimates of occupancy, abundance, and overdispersion invariably change over time, either explicitly via the selection of a within- or among-winter temporal effect (*time* and *y2004/y2005*, respectively) or implicitly via the selection of biophysical covariates that change within or among winters. The temporal dynamic of the wintering sea duck system in Nantucket Sound was one of its most striking attributes. Appendix 9 illustrates the dynamic nature of occupancy and average abundance in Nantucket Sound using scoter data during the 2005 winter.

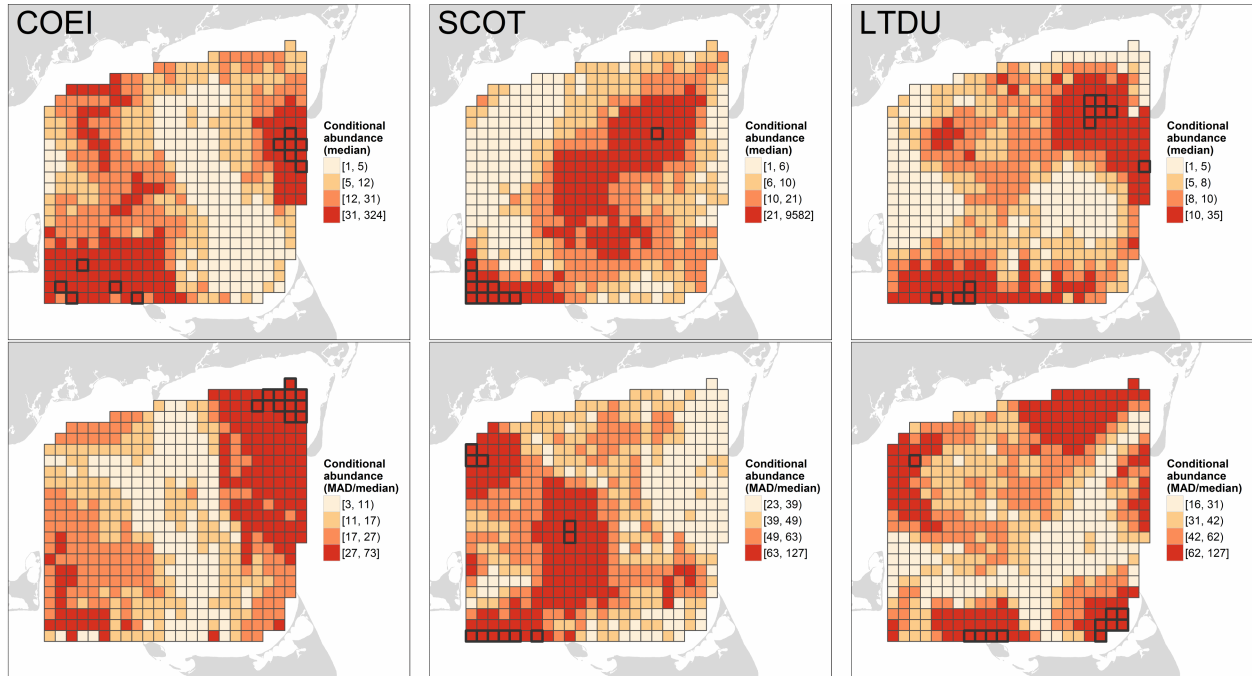


**Figure 2.** Marginal functional plots for stably selected covariates in the occupancy (probability of presence) and conditional abundance (mean and overdispersion of count model) of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) in Nantucket Sound

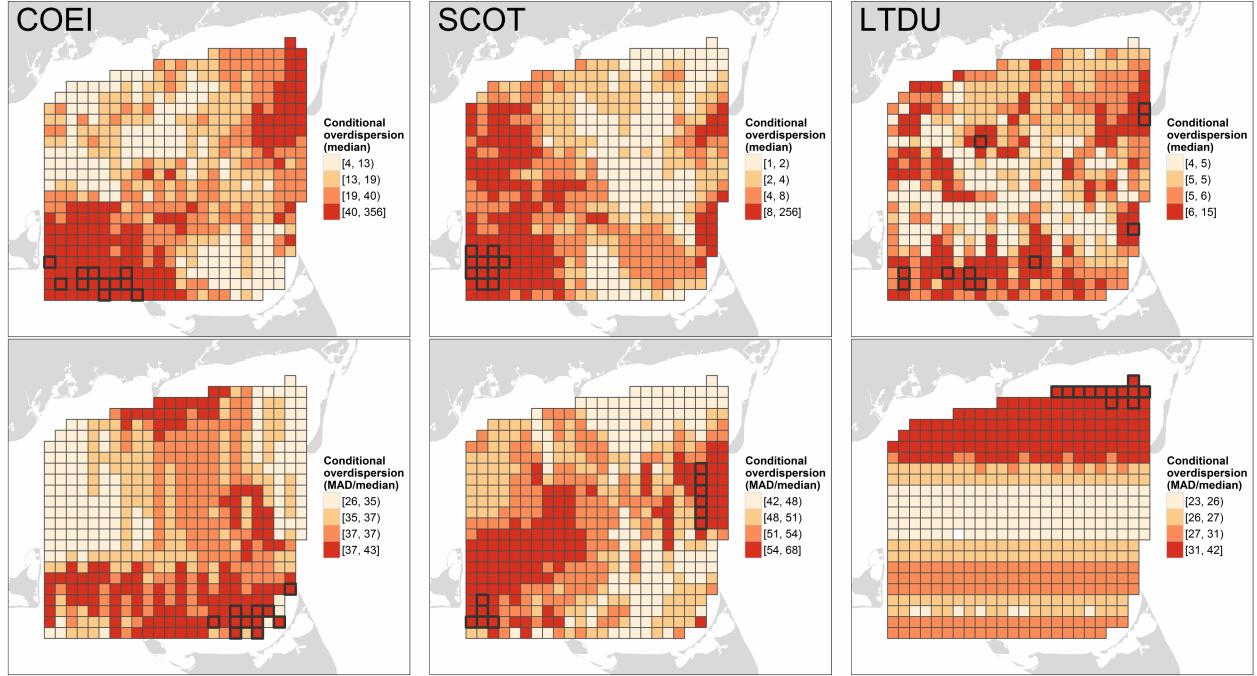
during winters 2003 - 2005. Plots illustrate the partial contribution to the additive predictor of a covariate holding all other covariates at their mean. Within a model, univariate plots (i.e., lines) share a Y-axis scale, enabling direct comparisons of effect sizes among covariates and species. For bivariate plots, the Y-axis and X-axis reflect the first and second variables listed in the interaction, respectively; colors indicate the direction and magnitude of the partial contribution (blacks = negative, reds = positive; darker colors = larger effect) and are likewise comparable within a model. Northing by easting effects are given only at 31 December. For factor variables, only the general association (i.e., positive or negative) with the additive predictor is given. Covariate abbreviations correspond to Equation 1.



**Figure 3.** Median (first row) and median absolute deviation (MAD) relative to median (%; second row) of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) occupancy in Nantucket Sound based on predictions at 10 dates evenly-spaced from 15 November through 1 April in three winters (2003 - 2005). Predicted values are categorized based on their quartiles; segments with predictions  $\geq 98$ th percentile are outlined in black. Occupancy values represent the predicted probability of sea duck presence in a 1.5 km x ca. 180 m transect through the segment.

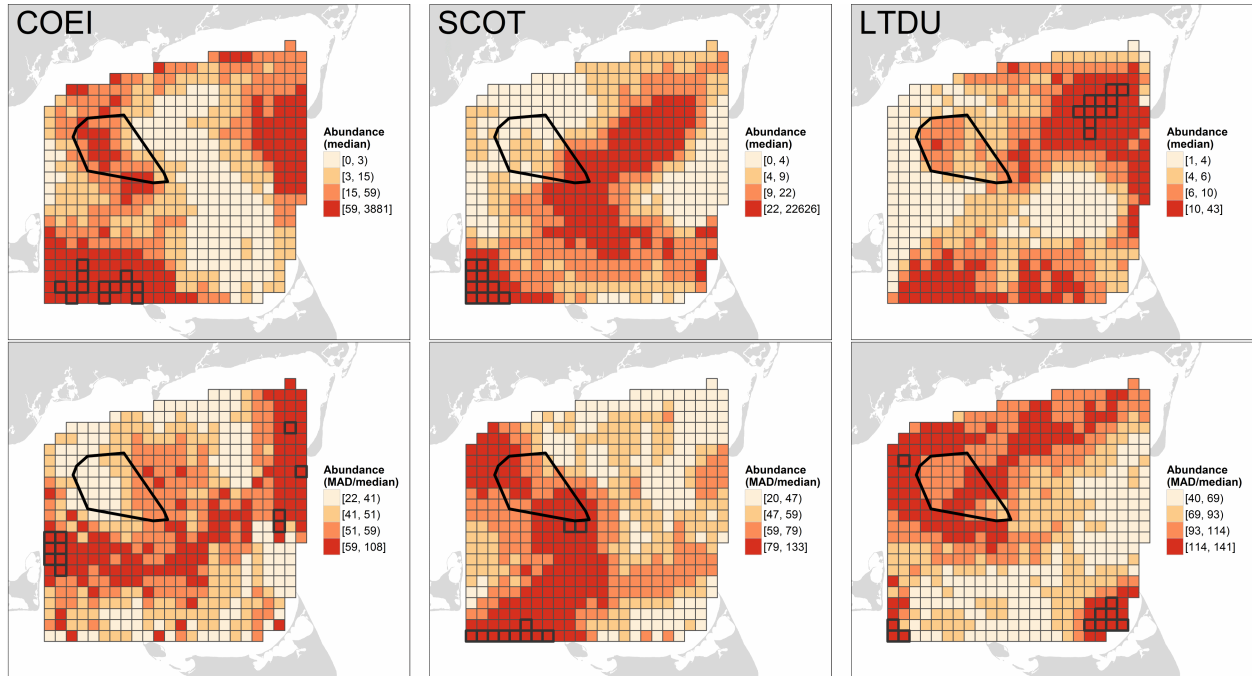


**Figure 4.** Median (first row) and median absolute deviation (MAD) relative to median (%) (second row) of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) conditional abundance in Nantucket Sound based on predictions at 10 dates evenly-spaced from 15 November through 1 April in three winters (2003 - 2005). Predicted values are categorized based on their quartiles; segments with predictions  $\geq 98$ th percentile are outlined in black. Conditional abundance values represent the expected number of sea ducks, conditional on their presence, in a 1.5 km x ca. 180 m transect through the segment.

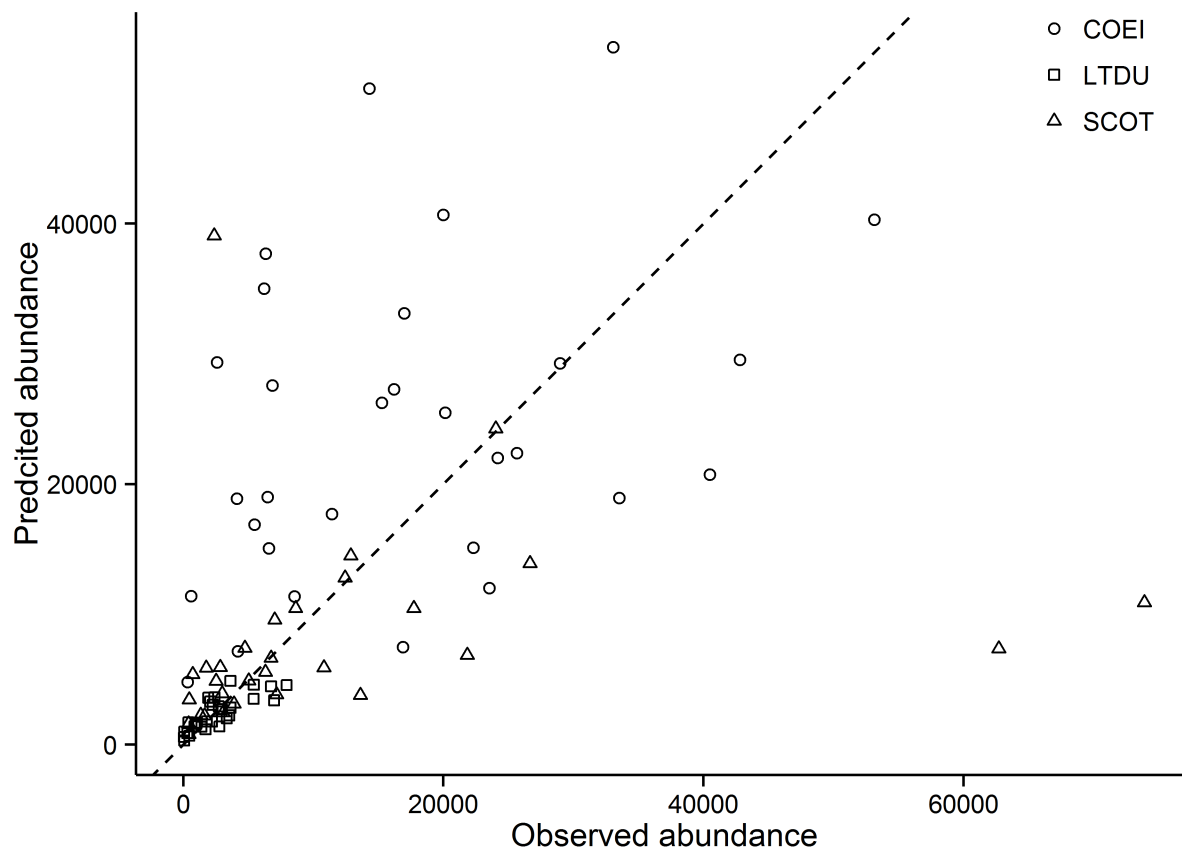


**Figure 5.** Median (first row) and median absolute deviation (MAD) relative to median (%) (second row) of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) conditional overdispersion in Nantucket Sound based on predictions at 10 dates evenly-spaced from 15 November through 1 April in three winters (2003 - 2005). Predicted values are categorized based on their quartiles; segments with predictions  $\geq 98$ th percentile are outlined in black. Higher conditional overdispersion values indicate increased overdispersion (variance in excess of the mean in the negative binomial model) in counts of sea ducks, conditional on their presence, in a 1.5 km x ca. 180 m transect through the segment.





**Figure 6.** Median (first row) and median absolute deviation (MAD) relative to median (%; second row) of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) average abundance in Nantucket Sound based on predictions at 10 dates evenly-spaced from 15 November through 1 April in three winters (2003 - 2005). Predicted values are categorized based on their quartiles; segments with predictions  $\geq 98$ th percentile are outlined. Conditional abundance values represent the expected number of sea ducks in a 1.5 km x ca. 180 m transect through the segment. The polygon in the northwest corner of Nantucket Sound indicates the location of permitted wind energy development on Horseshoe Shoal.



**Figure 7.** Comparison of predicted with observed abundance of Common Eider (COEI), scoter (SCOT), and Long-tailed Duck (LTDU) summed over all surveyed segments during 30 aerial surveys of Nantucket Sound. The dashed line indicates a 1:1 relationship between predicted and observed abundances; points below and above this line indicate predicted abundances below and above the observed totals, respectively.

## Discussion notes

Points of emphasis:

- open with paragraph recapping our methodology and its current application, and major conclusions, which possibly include:
  - Something regarding the apparent benefit of modelling approach? variable reduction? selection among competing functional forms (linear vs. non-linear)?
  - how much variation was explained?
  - disparate spatiotemporal patterns of sea duck species abundance in Nantucket Sound?
  - others?
- while occupancy is part of the picture in terms of describing species distribution, it potentially misses important abundance information (if it's available). In the current case, it was this conditional abundance information that largely controlled the spatiotemporal abundance of sea ducks in Nantucket Sound. In other words, occupancy does not necessarily equal abundance so occupancy models alone (e.g., Winiarski et al. 2014, ???) may be inadequate for assessing risk from anthropogenic disturbances or describing species distribution.
- the extensive temporal variability in sea duck abundance and distribution is often ignored as a potential source of variation (e.g., before/after studies; long-term studies of distribution/abundance); this considerable temporal variability may complicate the assessment of anthropogenic effects - for example, potentially something as trivial as inconsistency in the timing of surveys could influence pre/post comparisons
- Still much unexplained variation, especially considering that our measure of predictive ability/variation explained (pseudo- $R^2$ ) (1) was based on the data used to fit the model (in-sample) and (2) may be misleading because much of the explanatory power attributable to marginal spatial effect (i.e.,  $xkm-ykm$  position), which suggests we're most likely missing some important variable(s) relevant to the distribution of sea ducks or their prey ~~may be misleading because much of the explanatory power attributable to marginal spatial effect (i.e.,  $xkm-ykm$  position)~~. Corrolary: models are only as good as the input spatial data (e.g., crude sediment grain size). We need better proxies (or higher resolution of proxies?) for dynamic prey distribution or, better yet, concurrent prey distribution information (e.g., sea BC sea duck work, Kaiser et al. 2006, Ainley et al. 2009). I think there's some good discussion in CREEM guidance documents. Moreover, importance of spatiotemporal (unexplained heterogeneity) covariates complicates extrapolation/generalization to all but the immediate study area.
- Briefly mention the spatiotemporal heterogeneity in the overdispersion parameter in all species and how this contrasts sharply with the typical assumption in overdispersed (negative binomial) count models – that overdispersion is fixed – and suggests that in practice it is useful to consider that other parameters of the distribution (as GAMLSS

models allow) and not simply the conditional mean, depend on covariates. In other words, and in cases similar to the current study, certain biophysical/ecological features might influence the abundance of animals/organisms *and* the variability in those counts relative to the expectations of the underlying model (negative binomial in this case).

- Method extensibility; could be modified to generate precision estimates of abundance estimates, incorporate observation models (i.e., detection functions; distance sampling), better incorporate temporal or spatial autocorrelation via random effects ***BH: splines are similar though not equal to random effects models; in the spline case we have a smoothness penalty which takes the spatial pattern into account and can be related to a certain Gaussian distribution, i.e., random effect; a random effect usually imposes a Gaussian distribution but neglects the spatial structure***, use more current smoothing approaches (CREEM), accommodate double hurdle models (e.g., Earvin Balderama work) for extreme counts with zero inflation. Should I mention here that I checked for spatial and temporal autocorrelation in the residuals of the model? I found modest residual spatial (but not temporal) autocorrelation. ***BH: You can but usually that is not necessary as we explicitly allow for spatio-temporal variation in the model via smooth effects. The remaining auto-correlation would most likely vanish if we wouldn't use cross-validation approaches for early stopping.*** While the spatial smooth is typically *expected* to capture this “spatially correlated stochastic process” (Maloney et al. 2012), I *think* another approach could be incorporating appropriate random effects (**Benjamin?**). In any case, I could produce another brief Appendix illustrating the autocorrelation checks if you think it necessary. It certainly could come up in review. ***BH: I just had a short email discussion with a colleague regarding my thoughts that I expressed inline above and he totally agrees with me. From my point of view everything is fine as is and is solely a result from using optimal stopping to prevent overfitting.***
- Benjamin, we have to raise and assuage this issue if we expect people to use this for assessing impacts: How can we get precision estimates of predictions and/or derived estimates (e.g., change in abundance)? For example, how can we derive the precision of the spatially-explicit (i.e., segment-level) difference between a model constructed from data collected post-impact with the corresponding pre-impact model? Alternatively, if we included a factor variable in a comprehensive model (e.g., dichotomous pre/post indicator), stability selection would give us some indication of its general importance, but how would we translate that to a spatially-explicit effect? Would an appropriate interaction do it (e.g.,  $x_{km} \cdot y_{km} \cdot pre/post$ )? ***BH: I am sorry, I currently do not fully understand this item. Do you want to obtain prediction intervals (i.e., variability of the prediction) or confidence intervals (i.e., variability of the estimate)? The former could be obtained from the model as we modeled both, variability and mean, the latter could be obtained via bootstrap confidence intervals as given in (Hofner et al. 2014a).***
- Benjamin, your feedback here is also appreciated: drawbacks of the approach:

requisite computing power; difficulty of deriving precision estimates (if this is difficult to incorporate); difficulty to specify a good candidate model, i.e., which effects should be included and how?; missing significance tests, and confidence intervals that are relatively difficult and time-consuming to derive (but this is true for all penalization approaches); others? **BH: I've added some which you or me can now work with**

## Literature cited

- Ainley, D. G., K. D. Dugger, R. G. Ford, S. D. Pierce, D. C. Reese, R. D. Brodeur, C. T. Tynan, and J. A. Barth. 2009. Association of predators and prey at frontal features in the California Current: competition, facilitation, and co-occurrence. *Marine Ecology Progress Series* 389:271–294.
- Andrewartha, H. G., and L. C. Birch. 1986. *The Ecological Web: More on the Distribution and Abundance of Animals*. University of Chicago Press, Chicago, Illinois, USA.
- Borchers, D. L., S. T. Buckland, and W. Zucchini. 2002. *Estimating animal abundance: closed populations*. Springer, London.
- Bowman, T. D., E. D. Silverman, S. G. Gilliland, and J. B. Leirness. 2015. Status and trends of North American sea ducks: reinforcing the need for better monitoring. Pages 1–27 in J.-P. L. Savard, D. V. Derksen, D. Esler, and J. M. Eadie, editors. *Ecology and Conservation of North American Sea Ducks*. Studies in Avian Biology. CRC Press, New York, NY, USA.
- Buckland, S. T., M. L. Burt, E. A. Rexstad, M. Mellor, A. E. Williams, and R. Woodward. 2012. Aerial surveys of seabirds: the advent of digital methods. *Journal of Applied Ecology* 49:960–967.
- Bühlmann, P., and T. Hothorn. 2007. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22:477–505.
- Certain, G., and V. Bretagnolle. 2008. Monitoring seabirds population in marine ecosystem: the use of strip-transect aerial surveys. *Remote Sensing of Environment* 112:3314–3322.
- Davis, J., W.E. 1997. The Nantucket oldsquaw flight: New England's greatest bird show? *Bird Observer* 25:16–22.
- Eilers, P. H. C., and B. D. Marx. 1996. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* 11:89–121.
- Flint, P. L. 2013. Changes in size and trends of North American sea duck populations associated with North Pacific oceanic regime shifts. *Marine Biology* 160:59–65.
- Garthe, S., N. Markones, O. Hppop, and S. Adler. 2009. Effects of hydrographic and meteorological factors on seasonal seabird abundance in the southern North Sea. *Marine Ecology Progress Series* 391:243–255.

- Hastie, T. J., and R. J. Tibshirani. 1990. Generalized Additive Models. CRC Press, Boca Raton, FL, USA.
- Hoeting, J. A. 2009. The importance of accounting for spatial and temporal correlation in analyses of ecological data. *Ecological Applications* 19:574–577.
- Hofner, B., and T. Hothorn. 2015. stabs: stability selection with error control, R package version 0.5-1.
- Hofner, B., L. Boccuto, and M. Göker. 2015a. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics* in press.
- Hofner, B., T. Hothorn, T. Kneib, and M. Schmid. 2011. A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics* 20:956–971.
- Hofner, B., T. Kneib, and T. Hothorn. 2014a. A Unified Framework of Constrained Regression. *Statistics and Computing*:1–14.
- Hofner, B., A. Mayr, and M. Schmid. 2014b. gamboostLSS: An R Package for Model Building and Variable Selection in the GAMLSS Framework.
- Hofner, B., A. Mayr, N. Fenske, and M. Schmid. 2015b. gamboostLSS: boosting methods for GAMLSS models, R package version 1.1-3.
- Hofner, B., A. Mayr, N. Robinsonov, and M. Schmid. 2014c. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics* 29:3–35.
- Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. 2015. mboost: model-based boosting, R package version 2.4-2.
- Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner. 2010. Model-based boosting 2.0. *Journal of Machine Learning Research* 11:2109–2113.
- Hyrenbach, K. D., K. A. Forney, and P. K. Dayton. 2000. Marine protected areas and ocean basin management. *Aquatic Conservation: Marine and Freshwater Ecosystems* 10:437–458.
- Kaiser, M. J., M. Galanidi, D. A. Showler, A. J. Elliott, R. W. G. Caldow, E. I. S. Rees, R. A. Stillman, and W. J. Sutherland. 2006. Distribution and behaviour of Common Scoter *Melanitta nigra* relative to prey resources and environmental parameters. *Ibis* 148:110–128.
- Kinlan, B., C. Menza, and F. Huettmann. 2012. Chapter 6: predictive modeling of seabird distribution patterns in the New York Bight. Pages 87–224 *in* C. Menza, B. Kinlan, D. Dorfman, M. Poti, and C. Caldow, editors. A biogeographic assessment of seabirds, deep sea corals and ocean habitats of the New York Bight: science to support offshore spatial planning. NOAA Technical Memorandum NOS NCCOS 141. NOAA, Silver Spring, MD.
- Kneib, T., T. Hothorn, and G. Tutz. 2009. Variable selection and model choice in geoaddditive regression models. *Biometrics* 65:626–634.

- Kneib, T., J. Müller, and T. Hothorn. 2008. Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics* 15:343–364.
- MacArthur, R. H. 1972. *Geographical Ecology: Patterns in the Distribution of Species*. Joanna Cotler Books, New York, NY, USA.
- Maloney, K. O., M. Schmid, and D. E. Weller. 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. *Methods in Ecology and Evolution* 3:116–128.
- Marsh, H., and D. F. Sinclair. 1989. Correcting for visibility bias in strip transect aerial surveys of aquatic fauna. *The Journal of Wildlife Management* 53:1017–1024.
- Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8:1235–1246.
- Maurer, B. A. 1999. *Untangling Ecological Complexity: The Macroscopic Perspective*. University Of Chicago Press, Chicago, Illinois, USA.
- Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid. 2012. Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61:403–427.
- Meinshausen, N., and P. Bühlmann. 2010. Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:417–473.
- Merkel, F. R. 2004. Evidence of population decline in Common Eiders breeding in western Greenland. *Arctic* 57:27–36.
- Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78:691–692.
- Nur, N., J. Jahncke, M. P. Herzog, J. Howar, K. D. Hyrenbach, J. E. Zamon, D. G. Ainley, J. A. Wiens, K. Morgan, L. T. Ballance, and D. Stralberg. 2011. Where the wild things are: predicting hotspots of seabird aggregations in the California Current System. *Ecological Applications* 21:2241–2257.
- Ortega Cisneros, K., A. J. Smit, J. Laudien, and D. S. Schoeman. 2011. Complex, dynamic combination of physical, chemical and nutritional variables controls spatio-temporal variation of sandy beach community structure. *PLoS ONE* 6:e23724.
- Pollock, K. H., H. D. Marsh, I. R. Lawler, and M. W. Alldredge. 2006. Estimating animal abundance in heterogeneous environments: an application to aerial surveys for dugongs. *Journal of Wildlife Management* 70:255–262.
- R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology* 45:218–227.

- Rigby, R. A., and D. M. Stasinopoulos. 2005. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54:507–554.
- Schmid, M., and T. Hothorn. 2008. Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis* 53:298–311.
- Shah, R. D., and R. J. Samworth. 2013. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75:55–80.
- Winiarski, K. J., D. L. Miller, P. W. Paton, and S. R. McWilliams. 2014. A spatial conservation prioritization approach for protecting marine birds given proposed offshore wind energy development. *Biological Conservation* 169:79–88.
- Wood, S. N. 2006. *Generalized additive models: an introduction with R*. Chapman; Hall/CRC, Boca Raton, FL, USA.
- Zeileis, A., C. Kleiber, and S. Jackman. 2008. Regression models for count data in R. *Journal of Statistical Software* 27:1–25.
- Zipkin, E. F., B. Gardner, A. T. Gilbert, A. F. O. Jr, J. A. Royle, and E. D. Silverman. 2010. Distribution patterns of wintering sea ducks in relation to the North Atlantic Oscillation and local environmental characteristics. *Oecologia* 163:893–902.