Spatiotemporal Models of an Estuarine Fish Species to Identify Patterns and Factors Impacting Their Distribution and Abundance

Citation for published version:
Polansky, L, Newman, K, Nobriga, M & Mitchell, L 2018, 'Spatiotemporal Models of an Estuarine Fish Species to Identify Patterns and Factors Impacting Their Distribution and Abundance' Estuaries and Coasts, vol 41, no. 2, pp. 572-581. DOI: 10.1007/s12237-017-0277-3

Digital Object Identifier (DOI):
10.1007/s12237-017-0277-3

Published In:
Estuaries and Coasts

Estuaries and Coasts

Spatiotemporal models of an estuarine fish species to identify patterns and factors impacting their distribution and abundance

--Manuscript Draft--

Manuscript Number:	ESCO-D-16-00288R2
Full Title:	Spatiotemporal models of an estuarine fish species to identify patterns and factors impacting their distribution and abundance
Article Type:	Original Article
Keywords:	delta smelt; geostatistical models; population ecology; soap film smoothers; San Francisco Estuary
Corresponding Author:	Leo Polansky US Fish and Wildlife Service UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	US Fish and Wildlife Service
Corresponding Author's Secondary Institution:	
First Author:	Leo Polansky
First Author Secondary Information:	
Order of Authors:	Leo Polansky
	Ken B Newman
	Matthew L Nobriga
	Lara Mitchell
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Understanding the distribution and abundance of organisms can be exceedingly difficult for pelagic fish species that live in estuarine environments. This is particularly so for fish that cannot be readily marked and released or otherwise tracked, such as the diminutive delta smelt, Hypomesus transpacificus, endemic to the San Francisco Estuary. The environmental factors that influence distribution operate at multiple scales, from daily tidal cycles and local perceptual fields to seasonal and annual changes in dominant environmental gradients spanning the entire San Francisco Estuary. To quantify scale specific patterns and factors shaping the spatiotemporal abundance dynamics of adult delta smelt, we fit a suite of models to an extensive, spatially resolved, catch survey time series from 13 annual cohorts. The best model included cohort-specific abundance indicators and daily mortality rates, a regional spatial adjustment, and haul-specific environmental conditions. The regional adjustment identified several density hotspots that were persistent across cohorts. While this model did include local environmental conditions, the gain in explained variation was relatively slight compared to that explained by the regional adjustment. Total abundance estimates were derived by multiplying habitat volume by catch density (design-based) and modeled density (model-based), with both showing severe declines in the population over the time period studied. The design-based approaches had lower uncertainty but potentially higher bias. We discuss the implications of our results for advancing the science and improving management of delta smelt, and future data collection needs.

Spatiotemporal models of an estuarine fish species to identify patterns and factors impacting their distribution and abundance

Leo Polansky[1,*], Ken B. Newman[2], Matthew L. Nobriga[1], Lara Mitchell[2]

[1]U.S. Fish and Wildlife Service, Bay Delta Fish and Wildlife Office, Sacramento, CA 95814, USA

[2]U.S. Fish and Wildlife Service, Lodi Fish and Wildlife Office, Lodi CA 95240, USA

[*]Corresponding author email: leo_polansky@fws.gov

Running page head: Distribution and abundance of delta smelt

ABSTRACT: Understanding the distribution and abundance of organisms can be exceedingly difficult for pelagic fish species that live in estuarine environments. This is particularly so for fish that cannot be readily marked and released or otherwise tracked, such as the diminutive delta smelt, *Hypomesus transpacificus*, endemic to the San Francisco Estuary. The environmental factors that influence distribution operate at multiple scales, from daily tidal cycles and local perceptual fields to seasonal and annual changes in dominant environmental gradients spanning the entire San Francisco Estuary. To quantify scale specific patterns and factors shaping the spatiotemporal abundance dynamics of adult delta smelt, we fit a suite of models to an extensive, spatially resolved, catch survey time series from 13 annual cohorts. The best model included cohort-specific abundance indicators and daily mortality rates, a regional spatial adjustment, and haul-specific environmental conditions. The regional adjustment identified several density hotspots that were persistent across cohorts. While this model did include local environmental conditions, the gain in explained variation was relatively slight compared to that explained by the regional adjustment. Total abundance estimates were derived by multiplying habitat volume by catch density (design-based) and modeled density (model-based), with both showing severe declines in the population over the time period studied. The design-based approaches had lower uncertainty but potentially higher bias. We discuss the implications of our results for advancing the science and improving management of delta smelt, and future data collection needs.

KEY WORDS: delta smelt; geostatistical models; population ecology; soap film smoothers; San Francisco Estuary

INTRODUCTION

Determining how and why an organism's population is distributed in space and time is a fundamental organizing problem in population ecology (Krebs 1994). For small pelagic species in tidal river estuaries, drawing inference about their distribution and abundance is especially challenging because they cannot be tagged and tend to aggregate in schools that are small relative to the size of their open-water habitats. Tidal river estuaries are ecotones characterized by almost continuous multi-scale changes in environmental factors, from tidal to annual time scales and with spatial scale changes ranging from the perceptual field of the organism up to the entire span of the estuary (Odum 1988). In general, we can expect to need to apply models that can disentangle the relative effects of processes acting at different spatiotemporal scales.

Multi-scale environmental variability can be especially important for small resident pelagic species (Peebles et al. 2007; Reum et al. 2011). For example, tidal currents can influence vertical and horizontal distributions so that organisms can maintain or change geographic position within the estuary (Kimmerer and McKinnon 1987; Bennett and Burau 2015). Also, pelagic species will move in response to temperature, turbidity, salinity and prey density gradients, all of which can directly influence vital rates (Peebles et al. 2007; Reum et al. 2011; Rose et al. 2013) and shape estuary wide and regional distributions. A practical consequence for model-based analysis of distribution and abundance is that care must be taken to appropriately match the spatiotemporal resolution of the data and the model. Models must include factors, and allow for predictions, across multiple spatial and temporal scales simultaneously in order to provide useful insight into spatiotemporal variability in abundance.

The San Francisco Estuary (SFE, Fig. 1) is a tidal river estuary ecotone with habitat composition and structure that changes at multiple temporal and spatial scales (Cloern and Jassby 2012). One of the largest tidal river estuaries on the west coast of the Americas, the SFE provides habitat for delta smelt (*Hypomesus transpacificus*), an endemic annual pelagic fish species that inhabits the low salinity and freshwater portions of the estuary upstream of San Pablo Bay (Fig. 1). Substantial declines in the cohort abundance of delta smelt during the 1980s and early 1990s led to protection under both the U.S. and California Endangered Species Acts in 1993, and new fish monitoring programs, including one for the adult life stage. In addition to these spatially and temporally extensive fish surveys, measurements of several salient environmental metrics have also been collected in the SFE.

Delta smelt habitat preferences are relatively well understood (Moyle et al. 2016). The species distribution is constrained down-estuary by salinity while up-estuary a variety of life stage specific factors operate, including landward extent of tides, water clarity, salinity, temperature, and risk of entrainment into water diversions (Sweetnam 1999; Bennett 2005; Kimmerer 2008; Kimmerer 2008; Nobriga et al. 2008; Feyrer et al. 2011). Nevertheless, more precisely understanding the spatiotemporal changes (or lack of changes) in abundance within the broader range limits has been a focal point of conservation discussions (Brown et al. 2009; Feyrer et al. 2011), highlighting the need for statistical analysis at finer spatial and temporal scales than has been typically carried out.

Our primary motivation was to gain insight into patterns of the distribution and abundance of adult delta smelt. Specifically, we addressed three questions: Where do adult delta smelt distribute themselves during their spawning season, and how variable is this distribution across time (both within and between cohorts)? What factors operating at what scales most strongly influence the spatial distributions? What are the year-over-year population growth rates?

To answer these questions, we constructed spatiotemporal models of catch density with three different levels of spatiotemporal scale that we label global, regional, and local. Year of the survey and cohort-specific mortality rates were global-level (i.e. population wide) components to the model that described overall cohort specific trends in time. Regional (approximately 5 km and larger) spatial variation is apparent from exploratory data analyses (Fig. 2) of catch per unit volume (CPUV), the sum of all fish caught at given survey location divided by the sum of all water sampled at that location, and this spatial variation was modeled using nonparametric techniques. The importance of both within- and across-cohort changes in the regional spatial distribution patterns were tested. At the local (individual sample) level, we estimated how much of the variability in catch density was explained by three environmental covariates: water clarity, salinity, and tide. Increased turbidity and decreased salinity are expected to have positive effects on catch density based on *in situ* studies of earlier life stages (Nobriga et al. 2008; Feyrer et al. 2011). Flood relative to ebb tide was expected to increase catch densities as fish position themselves within the water column and channel to either move upstream or to otherwise maintain position (Feyrer et al. 2013; Bennett and Burau 2015). We also compared design-based and model-based estimates of abundance for February of each year (definitions of design-based and model-based inference are given in the supplementary material [SM] Section 4). Here the aim was to quantify inter-annual changes and long-term trends, to assess how different abundance estimates would be when standardizing effort for tide and to evaluate whether the two approaches have any qualitative differences.

METHODS

Data

The California Department of Fish and Wildlife established the Spring Kodiak Trawl (SKT) in 2002 to collect data on the distribution and reproductive stage of spawning delta smelt. The SKT survey usually visits 40 locations monthly from January through May (Fig. 1) over a several day period. During each location visit a 10 minute surface trawl of the approximately top 2m of water is taken. Three quarters of all trawls were made before noon. All delta smelt retained by the gear are counted and measured, and the volume of water sampled (*Vol*, m^3) is estimated. We used data from 2002-2014. Of the 2396 records used here, 1706 (71%) had zero catch. Of the 690 samples with positive catch, 227 recorded a single adult delta smelt, with a maximum catch of 375.

The local tow-specific environmental covariate data included Secchi disk depth (*Sec*, cm), a proxy for water clarity; specific conductance (*Cond*, microSiemens per centimeter, μS cm-1), a proxy for salinity; and tide stage (*Tide*) which is categorically recorded as ebb (1500 observations), low slack (68 observations), high slack (97 observations), flood (731 observations). Although water temperature is also recorded, for this analysis we did not include it in the models because the range of observed temperatures (min=6.6°C, max=23.6°C, mean=12.9°C) were well within the tolerance of spawning and post-spawn adult delta smelt (Swanson et al. 2000; Komoroske et al. 2014). Earlier versions of the model that did include temperature never identified it as statistically significant. In contrast, measures of salinity up to 21ppt, high enough to constrain distribution and affect survival (Komoroske et al. 2014; Lisa M. Komoroske et al. 2016), have been recorded in the SKT survey.

Spatiotemporal catch density models

The catch $y_{t,c,l}$ on Julian day t of cohort c at location l was modeled using a negative binomial distribution $y_{t,c,l} \sim \text{NegBin}(\mu_{t,c,l},\theta)$ parameterized to have expected value μ and variance $\mu+\mu^2/\theta$ (Venables and Ripley 2002). The negative binomial was selected given evidence for overdispersion relative to a Poisson distribution and from model residual diagnostics. The different models for $\mu_{t,c,l}$ are described next and summarized in Table 1.

Most generally, the expected catches $\mu_{t,c,l}$ were modeled using a semi-parametric, spatiotemporally explicit model within a generalized additive model (GAM) framework (Hastie and Tibshirani 1986; Wood 2006; Augustin et al. 2013). The expected catch is the product of the volume of water sampled, $Vol_{t,c,l}$, the true density $\delta_{t,c,l}$ in a spatially local region around l, and the catchability $q_{t,c,l}$,

$$\mu_{t,c,l} = q_{t,c,l}\delta_{t,c,l}Vol_{t,c,l}. \quad (1)$$

Catchability $q_{t,c,l}$ has recently (Maunder et al. 2014) been conceptualized as a function of availability (i.e. whether fish are in the tow path in the first place) and contact selectivity (the probability that the net will catch and retain the fish given availability) (see Arreguín-Sánchez 1996 for other classic definitions). The catchability parameter $q_{t,c,l}$ is confounded with the overall density, so it is assumed equal to1 for all the models. Further discussion of $q_{t,c,l}$ in the context of adult delta smelt surveying is provided in the Discussion.

Modifications to Eqn. (1) were made to study different sources of variability in $\delta_{t,c,l}$. The first, which is labeled global scale effects, and was included in all models and intended to capture temporal trends in the overall density (total fish over total water volume), was to rewrite Eqn. (1) as

$$\mu_{t,c,l} = \delta_{0,c}\exp(\beta_c t)Vol_{t,c,l} \qquad (2)$$

Eqn. (2) describes an exponential decline (assuming $\beta_c<0$) in density from an overall initial density $\delta_{0,c}$, and the expected catch is simply this density times the volume sampled on a given tow.

An extension of the global density model of Eqn. (2) was to add a regional scale factor, namely a dependency on space to the predictions,

$$\mu_{t,c,l} = \delta_{0,c}\exp(\beta_c t + s_{m,c})Vol_{t,c,l} \quad (3)$$

where $s_{m,c} = s_{m,c}(\text{UTMX}_l, \text{UTMY}_l)$ is a nonparametric spatial smooth. A total of four different hypotheses about how $s_{m,c}$ changed through time were considered: (1) it did not change in time, $s_{m,c} = s$; (2) it depended only on the month of the survey, $s_{m,c} = s_m$; (3) it depended only on the year of the survey, $s_{m,c} = s_c$; and (4) it depended on both the month and the year of the survey. Because the spatial adjustments to the density vary at scales larger than the water surveyed in a single trawl, these adjustments can be thought of as capturing spatially regional changes in density.

The global and regional effects model given by Eqn. (3) was further extended to include local scale effects. For each assumption about $s_{m,c}$, the effects of local environmental conditions on $\delta_{t,c,l}$ were estimated with the model

$$\mu_{t,c,l}=\delta_{0,c}\exp(\beta_c t + s_{m,c} + \beta_{Sec}Sec_{t,c,l} + \beta_{Cond}Cond_{t,c,l} + \beta_{Tide(t,c),t,c,l}Tide_{t,c,l})Vol_{t,c,l}. \qquad (4)$$

The importance of Secchi and conductivity was also considered in the absence of a regional spatial adjustment component, i.e. extending Eqn. (2) with these covariates.

In total fifteen different models were fit and evaluated (Table 1). Model fitting was done in the R environment (R Core Team 2016) primarily using the glm.nb (Venables and Ripley 2002) and gam (Wood 2004; Wood 2011) functions. Other functions and packages used are documented in the model code provided in the SM. Soap film smoothers (Wood 2008) were used to make spatial smooths $s_{m,c}$ follow large-scale habitat boundary features (SM Fig. S1). The boundaries were set up in particular to avoid an influence of catch between Montezuma Slough and either Cache Slough or Suisun Bay. Smoothing parameter estimation was done using maximum likelihood (Wood 2011), but other criteria used for estimating the smooth parameter such as generalized cross-validation did not qualitatively change the results. Secchi and conductivity measurements were standardized to their z-scores prior to model fitting. A wide range of smooth basis dimensions were considered to ensure results were not predicated on this choice, and standard model residual diagnostics were investigated, including semivariograms (Clark 2007) of residuals by month and year. Model comparison was done by assessing residual diagnostics, Akaike's information criterion AIC (Burnham and Anderson 2002), fitted negative log-marginal-likelihoods (NLML, see Eqn. 5 in Wood 2011).

Model evaluation of the effects of the locally measured covariates Secchi and conductivity was partly complicated because of their global spatial structure. On average, more easterly (upstream) regions of the delta smelt habitat are clearer and less saline (SM Fig. S2), leading to the possibility that local environmental covariates will be confounded with the spatial terms in the model. To approximate an upper bound on the most variability that local environmental conditions might explain in the absence of spatial terms in the model, we computed the proportion of null deviance explained by models of the form of Eqn. (2) but including each of these covariates one at a time (Table 1 models 2-4). The proportion of the deviance explained by each locally measured covariate when fitting the full model in Eqn. (4) (Table 1 models 13-15) was also calculated by dropping each term individually and refitting the model while fixing the smoothing parameters at the values estimated in model 9. This helped ensure that no changes in the smoothing penalty upon refitting resulted in a "mopping up" of variation previously accounted for by the removed covariate, thereby diminishing the estimated proportion of deviance explained by the dropped covariate under consideration.

Abundance estimation

Total abundance estimates for the month of February for each year were made using both design-based and model-based approaches (SM Sec. 4). Both approaches rely on volumetric expansions of density estimates. The volumes were calculated by multiplying the area of water with at least 2 meters depth (provided by the United States Geological Survey) by 2 to compute the volume of habitat V_{tot} over which the density estimates might reasonably be extrapolated. This volume excludes water deeper than 2 meters as well as shoal habitats. Thus our estimates are likely underestimating the total population size depending on unknown densities in these unsampled water volumes. However, this approach avoids extrapolating catch density information into habitats that are not sampled by the SKT survey.

The design-based approach stratified the waterways most commonly occupied by delta smelt into 27 subregions (SM Fig. S3). The subregion, year and month specific catch densities were expanded by subregion-specific water volumes and summed to obtain year and month-specific abundance estimates. Assuming the abundance estimates were lognormally distributed, the 2.5 and 97.5 percentiles of this distribution were used to construct design-based prediction intervals. Section 4.1 of the SM provides details on obtaining the parameters for these cohort specific distributions.

In contrast to the design-based approach, the model-based approach does not require spatial stratification of the habitat and allows predictions to be contingent on specific environmental conditions thought to affect catchability. Based on model selection results, model 9 was used to make model-based total abundance estimates as follows. We used 984 points distributed within the spatial limits of the survey and the areas of water with at least 2 meters depth (SM. Fig S1) as the spatial locations for predictions. At each one of these locations, the density per $10000m^3$ of water was predicted on February 15th (specifying a day is necessary for the Julian day effect) of each year, the tide set equal to the flood factor level, and the Secchi and conductivity values fixed at a month, year, and location specific value (described below). These densities were averaged within each subregion, multiplied by the subregion water volume down to 2m, and summed to produce overall abundance estimates (see SM Sec. 4.2 for details). Because direct observations on Secchi depth and conductivity at the point locations used in making predictions were not always available, spatially smoothed GAMs were used to predict both of these variables during the February survey periods of each year. The GAMs were fit using the available survey data on Secchi depth and conductivity and had the form $y_{t,c,l}$ ~ Normal($\mu = \beta_{m,c} + s_{m,c}, \sigma$), where $y_{t,c,l}$ was either the z-transformed Secchi depth or conductivity measurements from the SKT survey. The fits were generally quite good: the models of Secchi depth and conductivity described at least 88% and 94% of the null deviance for 80% of the months, respectively. Abundance prediction intervals were estimated using a parametric bootstrapping approach that included uncertainty in model parameters, covariate predictions, and observations (see SM Sec. 4.2).

RESULTS

Table 1 shows model summary statistics. There was clear support for including both a regional spatial adjustment and local environmental conditions in the expected catch models. The best model identified by AIC included a separate spatial distribution for each month (model 10), while the negative log-marginal-likelihood identified a model with a constant spatial smooth over time as the best (model 9). Residual diagnostics for models without a regional smooth adjustment term were poor as measured by distributional checks of residuals. In contrast, models including a regional spatial term had residual qq-plots and semivariograms that suggested no systematic bias in predictions due to the spatial variability in the distribution. Simpler models had higher dispersion parameters, reflecting larger prediction error when the mean structure was less flexible.

Models including a smooth term to capture regional variation in catch identified several density hotspots (Figs. 3 and S4; see also Fig. 2 for empirical densities): the waterways surrounding Grizzly Island, channels at the confluence of the Sacramento and San Joaquin rivers, the Cache Slough complex, and the Sacramento deep water shipping channel. These density hotspots were fairly consistent between cohorts, with the Cache Slough complex and Sacramento deep water shipping channel the most persistently high. We focused on model 9 for making predictions because the differences in month-specific predictions in model 10 are dominated by disappearance of density hotspots in April and May (likely reflecting post-spawning mortality) rather than a spatial shift in the locations of hotspots (SM Fig. S4).

The local environmental covariates tide and conductivity explained very little (<2%) of the null deviance beyond that of model 1, but Secchi depth explained an additional 21.3% of the null deviance when no regional spatial adjustment was made (Table 1, models 2-4 and 13-15). The effect size on the linear scale of Secchi was approximately double that of conductivity, but both local covariates could translate into substantially larger expected changes in density predictions over the range of observed turbidity and salinity indices (Table 2). Catch density was higher on flood and low slack tide levels in comparison with ebb tide (the increase on low slack tide was the highest, but surveys during this tide stage account for <3% of samples), and not significantly different for high slack tide conditions (Table 2).

Figure 4 shows the total abundance estimates and prediction uncertainty for February 15th of each year (see SM Table S1 for values) for the design- and model-based estimates. The geometric mean annual growth rate over the 13 years was 0.88 and 0.87 for the design- and model-based approaches, respectively, and the percentage decline from 2002 to 2014 was 82% and 79% for the design- and model-based approaches, respectively. Note that the results about declines do not depend on the tide factor level choice used in making total abundance estimates. Despite the general agreement between design- and model-based estimates of trend, the two approaches showed the same annual growth rate in only 6 of the 13 years, and differed in magnitude especially in 2003 and 2012 (Fig. 4 and SM Table 1). The differences in abundance magnitude did depend on the model chosen, with the most complicated model showing predictions very similar to the design-based approach (SM Fig. S5).

DISCUSSION

For small, elusive, and rare pelagic fish species such as delta smelt, often the only source of information from the wild is catch density from trawls or other types of nets (e.g., beach seining), along with additional measurements of local environmental conditions. Given such data, at a minimum we would like to quantify the variability in distribution and abundance. Ideally, we could go further to identify causal factors that explain the variability at different scales, or rule out those that do not, and to assess the extent to which findings from theoretical and laboratory work are identifiable in the wild.

The spatial distributions quantified here are similar to the descriptive reports by Merz et al. (2011) and Murphy and Hamilton (2013) in their general depiction. By constructing statistical models, we were able to test hypotheses about the variability of this spatial structure. At a regional scale, our models indicated that the distribution of adult delta smelt was fairly consistent across months and years, with the dominant within-year change being disappearance of hotspots likely due to post-spawn mortality as the spawning season progresses for this annual species. This suggests that the majority of regional movement from juvenile and sub-adult rearing locations to spawning areas has already happened by the time the SKT survey is conducted, that spawning habitat locations are relatively constant within and between years, and that no substantial further restructuring of the population at regional scales occurs afterwards.

What leads to the emergence of density hotspots remains to be determined. A recent pairing of the sub-adult delta smelt catch data used by Feyrer et al. (2011) with a three-dimensional hydrodynamic model suggests that density hotspots may reflect the interplay of local water quality conditions with tidal velocity differences that exist between shoals and deeper shipping channels (Bever et al. 2016). Other possible explanations for adult and spawning delta smelt spatial variation include distributions of prey or spawning habitats, or areas more suited for survival during spawning. Why no density hotspots emerge and persist upstream of the Jersey Point (located near the arrow tip showing the San Joaquin River in Fig. 1) area remains to be determined, but likely factors include inhospitable habitat and advection of fish into water export facilities (Kimmerer 2008; Kimmerer 2011).

At local spatial scales there continues to be high variability in the spatial distribution (which necessitated the use of a negative binomial catch distribution model), some of which is likely related to spawning-related aggregations of delta smelt and some of which is related to changes in local salinity (movement away from) and turbidity (movement towards) conditions. Our view is that the best interpretation of the categorical covariate tide is that it affects changes in fish availability to the gear, a component of catchability $q_{t,c,l}$, with the direction of the effects found here being consistent with Feyrer et al. (2013). In general it appears that, due to its relatively coarse spatial and temporal resolution, the SKT survey cannot distinguish between very local, site level movement, up to movement between adjacent locations, and changes in catchability related to local environmental conditions. The infrequent yet extremely large catches point to highly localized and ephemeral aggregations of fish but, similar to questions about the existence of regional density hotspots, the relative contributions of social cues vs. habitat cues vs. hydrodynamics leading to the formation of these aggregations remains to be determined.

Previous analyses of the sub-adult life-stage have found local environmental covariates to be statistically significant predictors of delta smelt distribution, with Feyrer et al. (2011) remarking that “specific conductance and Secchi depth accounted for a meaningful reduction of null deviance.” In contrast, we found that these covariates explained very little of the variation in adult catch when a regional spatial adjustment to density was included. The comparatively large amount of deviance explained by Secchi depth when no spatial smooths were included in the model (model 2) suggests that water clarity has some influence on both local and regional distributions, although from a statistical perspective any models not containing a spatial adjustment beyond what is made by the local environmental covariates were very inferior. While suitable local environmental conditions are necessary to explain the distribution and abundance of delta smelt, they are far from sufficient. We suggest that to better understand both the regional and local changes in densities, an understanding of the characteristics leading to ideal spawning habitat features is needed, along with assessments of the variability of these characteristics in space and time.

At the decadal time scale delta smelt are currently in a severe state of population decline, with suspected causes including removal of water from the system and alien species (Moyle et al. 2016). Here we used the best available survey data to quantify this decline more precisely. Design- and model-based approaches closely agreed in the rate and amount of overall decline from 2003 to 2014.

Despite the general agreement in long-term trends between the two approaches for abundance estimation, there were also differences. In 2003 the design-based estimates showed a decline in abundance compared to 2002, while model-based estimates showed an increase. During this year the frequency of sampling on the flood tide was only 8%, and this may have led to the qualitative mismatch in year-over-year abundance change between the design- and model-based methods. It seems likely that the design-based approach is negatively biased when compared with the model-based approaches due to the failure to account for the effect of tide cycle on catchability $q_{t,c,l}$. Another difference was in prediction intervals, with model-based ones being notably wider likely related to the more complete inclusion of the different sources of uncertainty in the model-based approach which is accounting for spatial and tow-specific sources of uncertainty. Finally, the magnitude of the estimates also differed, with model-based estimates generally being substantially higher, although models with more complicated smooths had estimates that increasingly approached the design-based ones (SM Fig. S5). This closer agreement of the models with the most complicated smooths and the design-based approach is likely due in part to overfitting, whereby the expected model predictions are able to more closely track zero catch data. Other surveys making multiple tows per site visit have found that although the frequency of zero catch was similarly high on any given tow, nonzero catch usually occurred at least once

(Polansky et al. 2014). Thus, we suspect that the models with simpler spatial smooth terms are more reflective of actual distributions because they are drawing on information across time, and hence less informed by zero catch when in fact fish may be locally in the area. Whether using design- or model-based approaches to construct abundance estimates, information about false zero catches as well as abundances in shoal habitats as well as the vertical density gradients in channel and open-water habitats are needed to reduce abundance estimate bias and uncertainty.

Pinpointing the relative contributions of anthropogenic vs. natural sources to the population decline will continue to be challenging, and will likely best be done in a complete life-cycle analysis framework that integrates survey data from all life stages. Absolute abundance estimates will first be needed from each source in order to integrate information from different life stages, and catch level models such as applied here can help achieve this. The importance of tide, found here and elsewhere (Bennett et al. 2002; Feyrer et al. 2013; Bennett and Burau 2015), emphasizes a need to consider accounting for this covariate analyses where organism detection might be driven by tidal conditions (see also Arreguín-Sánchez 1996) to control for its effect on catch density. None of the previous population dynamics models using annual abundance indices (Mac Nally et al. 2010; Thomson et al. 2010; Maunder and Deriso 2011) attempted to standardize catch data when making these indices, which could mean that abundance and covariate relationships have not been described accurately.

Acknowledgments

Two reviewers and Associate Editor Dr. W. J. Kimmerer provided extensive comments that substantially improved earlier versions of this manuscript. The viewpoints expressed are those of

the authors and do not necessarily reflect the opinions of the U.S. Department of the Interior or the U.S. Fish and Wildlife Service.

Supplementary Material

Supplementary material with additional details, figures and code is provided. Data and code are available from the U.S. Fish and Wildlife Service.

REFERENCES

Arreguín-Sánchez, Francisco. 1996. Catchability: a key parameter for fish stock assessment. *Reviews in Fish Biology and Fisheries* 6: 221–242.

Augustin, Nicole H., Verena M. Trenkel, Simon N. Wood, and Pascal Lorance. 2013. Space-time modelling of blue ling for fisheries stock management. *Environmetrics* 24: 109–119.

Bennett, William A., and Jon R. Burau. 2015. Riders on the storm: selective tidal movements facilitate the spawning migration of threatened delta smelt in the San Francisco Estuary. *Estuaries and Coasts* 38: 826–835.

Bennett, William A. 2005. Critical assessment of the delta smelt population in the San Francisco Estuary, California. *San Francisco Estuary and Watershed Science* 3: 1–71.

Bennett, William A., Wim J. Kimmerer, and Jon R. Burau. 2002. Plasticity in vertical migration by native and exotic estuarine fishes in a dynamic low-salinity zone. *Limnology and Oceanography* 47: 1496–1507.

Bever, Aaron J., Michael L. MacWilliams, Bruce Herbold, Larry R. Brown, and Frederick V. Feyrer. 2016. Linking hydrodynamic complexity to delta smelt (*Hypomesus* *transpacificus*) distribution in the San Francisco Estuary, USA. *San Francisco Estuary and Watershed Science* 14: 1–25.

Brown, Larry R., Wim J. Kimmerer, and Randall Brown. 2009. Managing water to protect fish: a review of California’s environmental water account, 2001-2005. *Environmental Management* 43: 357–368.

Burnham, Kenneth P., and David Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer.

Clark, James S. 2007. Models for Ecological Data: An Introduction. Princeton, NJ: Princeton University Press.

Cloern, James E., and Alan D. Jassby. 2012. Drivers of change in estuarine-coastal ecosystems: Discoveries from four decades of study in San Francisco Bay. *Reviews of Geophysics* 50.

Feyrer, Frederick, Ken B. Newman, Matthew Nobriga, and Ted Sommer. 2011. Modeling the effects of future outflow on the abiotic habitat of an imperiled estuarine fish. *Estuaries and Coasts* 34: 120–128.

Feyrer, Frederick, Donald Portz, Darren Odum, Ken B. Newman, Ted Sommer, Dave Contreras, Randall Baxter, Steven B. Slater, Deanna Sereno, and Erwin Van Nieuwenhuyse. 2013. SmeltCam: underwater video codend for trawled nets with an application to the distribution of the imperiled delta smelt. *PLoS ONE* 8: e67829.

Hastie, Trevor, and Robert Tibshirani. 1986. Generalized additive models (with discussion). *Statistical Science* 1: 297–318.

Kimmerer, Wim J. 2008. Losses of Sacramento River Chinook Salmon and Delta Smelt to Entrainment in Water Diversions in the Sacramento–San Joaquin Delta. *San Francisco Estuary and Watershed Science* 6.

Kimmerer, Wim J. 2011. Modeling Delta Smelt Losses at the South Delta Export Facilities. *San Francisco Estuary and Watershed Science* 9.

Kimmerer, Wim J., and A. D. McKinnon. 1987. Zooplankton in a marine bay. II. Vertical migration to maintain horizontal distributions. *Marine Ecology Progress Series* 41: 53–60.

Komoroske, L. M., R. E. Connon, J. Lindberg, B. S. Cheng, G. Castillo, M. Hasenbein, and N. A. Fangue. 2014. Ontogeny influences sensitivity to climate change stressors in an endangered fish. *Conservation Physiology* 2:1-13.

Komoroske, Lisa M., Ken M. Jeffries, Richard E. Connon, Jason Dexter, Matthias Hasenbein, Christine Verhille, and Nann A. Fangue. 2016. Sublethal salinity stress contributes to habitat limitation in an endangered estuarine fish. *Evolutionary Applications* 9: 963–981.

Krebs, Charles J. 1994. *Ecology: the experimental analysis of distribution and abundance*. 4th ed. New York, NY: HarperCollins College Publishers.

Mac Nally, Ralph, James R. Thomson, Wim J. Kimmerer, Frederick Feyrer, Ken B. Newman, Andy Sih, William A. Bennett, et al. 2010. Analysis of pelagic species decline in the upper San Francisco Estuary using multivariate autoregressive modeling (MAR). *Ecological Applications* 20: 1417–1430.

Maunder, Mark N., and Richard B. Deriso. 2011. A state–space multistage life cycle model to evaluate population impacts in the presence of density dependence: illustrated with application to delta smelt (*Hyposmesus transpacificus*). *Canadian Journal of Fisheries and Aquatic Sciences* 68: 1285–1306.

Maunder, M.N., P.R. Crone, J.L. Valero, and B.X. Semmens. 2014. Selectivity: Theory, estimation, and application in fishery stock assessment models. *Fisheries Research* 158: 1–4.

Merz, Joseph E., Scott Hamilton, Paul S. Bergman, and Bradley Cavallo. 2011. Spatial perspective for delta smelt: a summary of contemporary survey data. *California Fish and Game* 97: 164–189.

Moyle, Peter B., Larry R. Brown, John R. Durand, and James A. Hobbs. 2016. Delta smelt: life history and decline of a once-abundant species in the San Francisco Estuary. *San Francisco Estuary and Watershed Science* 14: 1–30.

Murphy, Dennis D., and Scott A. Hamilton. 2013. Eastward migration or marshward dispersal: exercising survey data to elicit an understanding of seasonal movement of delta smelt. *San Francisco Estuary and Watershed Science* 11: 1–21.

Nobriga, Matthew L., Ted R. Sommer, Frederick Feyrer, and Kevin Fleming. 2008. Long-term trends in summertime habitat suitability for delta smelt (*Hypomesus transpacificus*). *San Francisco Estuary and Watershed Science* 6: 1–13.

Odum, William E. 1988. Comparative Ecology of Tidal Freshwater and Salt Marshes. *Annual Review of Ecology and Systematics* 19: 147–176.

Peebles, Ernst B., Scott E. Burghart, and David J. Hollander. 2007. Causes of Interestuarine Variability in Bay Anchovy (*Anchoa mitchilli*) Salinity at Capture. *Estuaries and Coasts* 30: 1060–1074.

Polansky, L., Matt Nobriga, Ken Newman, Matt Dekar, Kim Webb, and Mike Chotkowski. 2014. Delta smelt movement during and extreme drought: Intensive Kodiak trawling at Jersey Point. *Interagency Ecological Newsletter* 4:5-13.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing* (version 3.3.0). Vienna, Austria: R Foundation for Statistical Computing.

Reum, Jonathan C. P., Timothy E. Essington, Correigh M. Greene, Casimir A. Rice, and Kurt L. Fresh. 2011. Multiscale influence of climate on estuarine populations of forage fish: the role of coastal upwelling, freshwater flow and temperature. *Marine Ecology Progress Series* 425: 203–215.

Rose, Kenneth A., Wim J. Kimmerer, Karen P. Edwards, and William A. Bennett. 2013. Individual-based modeling of delta smelt population dynamics in the upper San Francisco Estuary: I. model description and baseline results. *Transactions of the American Fisheries Society* 142: 1238–1259.

Swanson, Christina, Turid Reid, Patricia S. Young, and Joseph J. Cech Jr. 2000. Comparative environmental tolerances of threatened delta smelt (*Hypomesus transpacificus*) and introduced wakasagi (*H. nipponensis*) in an altered California estuary. *Oecologia* 123: 384–390.

Sweetnam, Dale A. 1999. Status of delta smelt in the Sacramento-San Joaquin Estuary 85: 22–27.

Thomson, James R., Wim J. Kimmerer, Larry R. Brown, Ken B. Newman, Ralph Mac Nally, William A. Bennett, Frederick Feyrer, and Erica Fleishman. 2010. Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco Estuary. *Ecological Applications* 20: 1431–1448.

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S. Fourth Edition. New York, NY: Springer.

Wood, Simon N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99: 673–686.

Wood, Simon N. 2006. Generalized additive models: an introduction with R. Boca Raton, FL: Chapman & Hall.

Wood, Simon N. 2008. Soap film smoothing. *Journal of the Royal Statistical Society B* 70: 931–955.

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Society: Series B* 73: 3–36.

Tables

Table 1- Overview of catch models showing the (effective) degrees of freedom (df), information theoretic measures of model goodness of fit (AIC and ΔAIC), the negative log-marginal-likelihood (NLML- smaller values are better), and percent of the null deviance explained (% dev. expl.). Global is defined by Eqn. (2) and Global + regional is defined by Eqn. (3), while local terms are Secchi (Sec), conductivity (Cond), and Tide- see Eqn. (4). Regional spatial smooth terms were either constant across months and years (single), different by month but not year (monthly), different by year but not month (yearly), or different for each month and year. Models 13-15 had fixed smooth term parameters using the estimates from model 9 and were used for estimating the % dev. expl. by each of the three individual local terms in model 9.

Model	Density model	df	AIC	ΔAIC	NLML	% dev. expl.	θ
1	Global	27	6717.2	1178.2	-	12.9	0.1
2	Global + Sec	28	6337.8	798.9	-	34.2	0.2
3	Global + Cond	28	6692.8	1153.8	-	14.5	0.1
4	Global + Tide	30	6701.4	1162.5	-	14.2	0.1
5	Global + regional (single)	49.9	5643.0	104.0	2821.8	63.6	0.4
6	Global + regional (monthly)	118.1	5638.3	99.3	2853.3	67.4	0.4
7	Global + regional (yearly)	199.0	5603.3	64.4	2831.6	72.2	0.5
8	Global + regional (month and year)	632.5	5888.7	349.7	2933.1	83.8	0.8
9	Global + regional (single) + Sec + Cond + Tide	54.7	5548.2	9.3	2769.2	66.9	0.4
10	Global + regional (monthly) + Sec + Cond + Tide	128.1	5538.9	0.0	2789.3	70.6	0.5
11	Global + regional (yearly) + Sec + Cond + Tide	198.6	5572.3	33.4	2798.6	72.9	0.5
12	Global + regional (month and year) + Sec + Cond + Tide	506.3	5726.6	187.7	2819.6	82.4	0.7
13	Global + regional (single, fixed) + Cond + Tide	52.7	5606.3	67.4	2801.1	65.0	0.4
14	Global + regional (single, fixed) + Sec + Tide	52.7	5566.9	28.0	2780.2	66.1	0.4
15	Global + regional (single, fixed) + Sec + Cond	50.7	5557.9	18.9	2778.0	66.4	0.4

Table 2- Parameter estimates and bootstrapped estimates of uncertainty for the parameters associated with the local environmental covariates for model 9 (see Table 1) on the $\log_e$ scale. Lower and upper columns show the 2.5 and 97.5 percentiles from 1000 samples from a multivariate normal distribution parameterized by the mean and covariance matrix from the fitted model 9. The final columns show density prediction differences on the response scale given the described local environmental change, where the changes are based on changing from the 2.5 to the 97.5 percentile for the continuous covariate observations, and in comparison with an ebb tide.

Covariate	Estimate	Lower	Upper	Density factor change on response scale	
Secchi depth	-0.880	-1.112	-0.670	Decrease in turbidity	0.415
Conductivity	-0.403	-0.583	-0.232	Increase in salinity	0.669
Flood	0.338	0.113	0.552	From ebb to flood	1.398
High slack	-0.093	-0.658	0.476	From ebb to high slack	0.910
Low slack	0.962	0.389	1.571	From ebb to low slack	2.622

Figures

Figure 1- Overview of the inland portion of the San Francisco Estuary where adult delta smelt are most commonly found. Black x's denote the regular monthly Spring Kodiak Trawl survey locations.

Figure 2- Mean catch per unit volume at each sampling location for each month (averaged over 2002-2014). Units are per 10000m^3 of water.

Figure 3- Density predictions at a flood tide per 10000m^3 of water based on model 9 on February 15th 2004 using the mean Secchi and conductivity values. By fixing the local covariates the figure emphasizes density variation due to intrinsic variability. For clarity catch densities above 10 fish/10000m^3 of water are colored the same. See SM Fig. S4 for month specific predictions using model 10.

Figure 4- Abundance estimates on February 15th of each year. Design-based abundance estimates are shown by the line with filled circles with vertical lines extending to the 2.5 and 97.5 percentiles of the lognormal distributions. Model-based predictions from model 9 are shown as a solid line with dashed lines drawn at the 2.5 and 97.5 prediction percentiles based on 1000 bootstrapped samples. Inset numbers show the percentage of samples in each February that were done on a flood tide to illustrate the variability in sample conditions, which the model-based estimates account for. See SM Fig. S5 predictions using models 8 and 12.

Figure 1

Figure 1

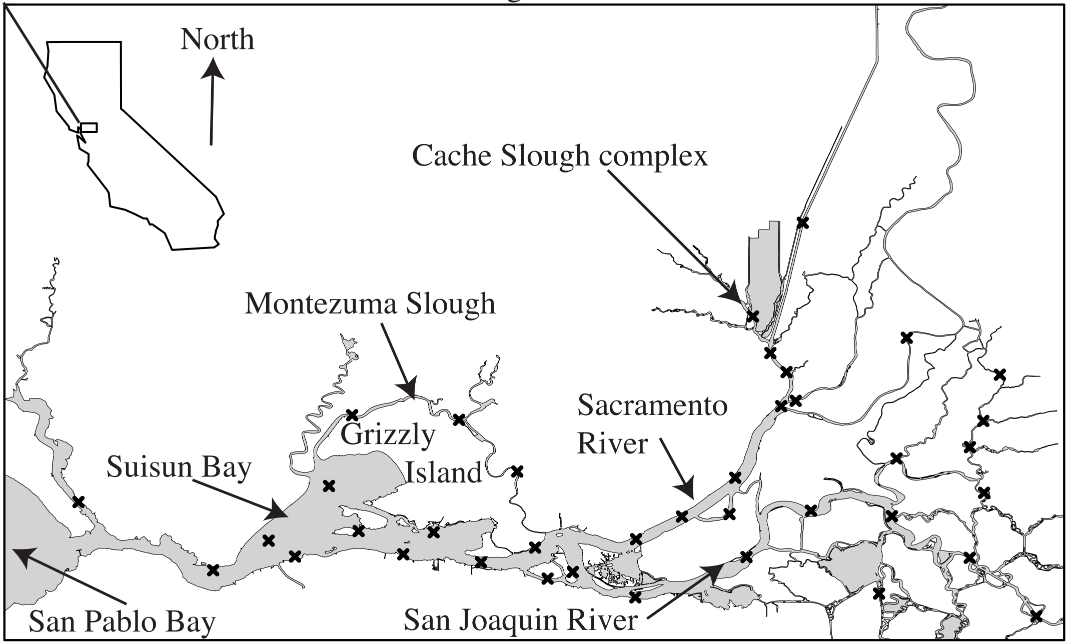


Figure 2

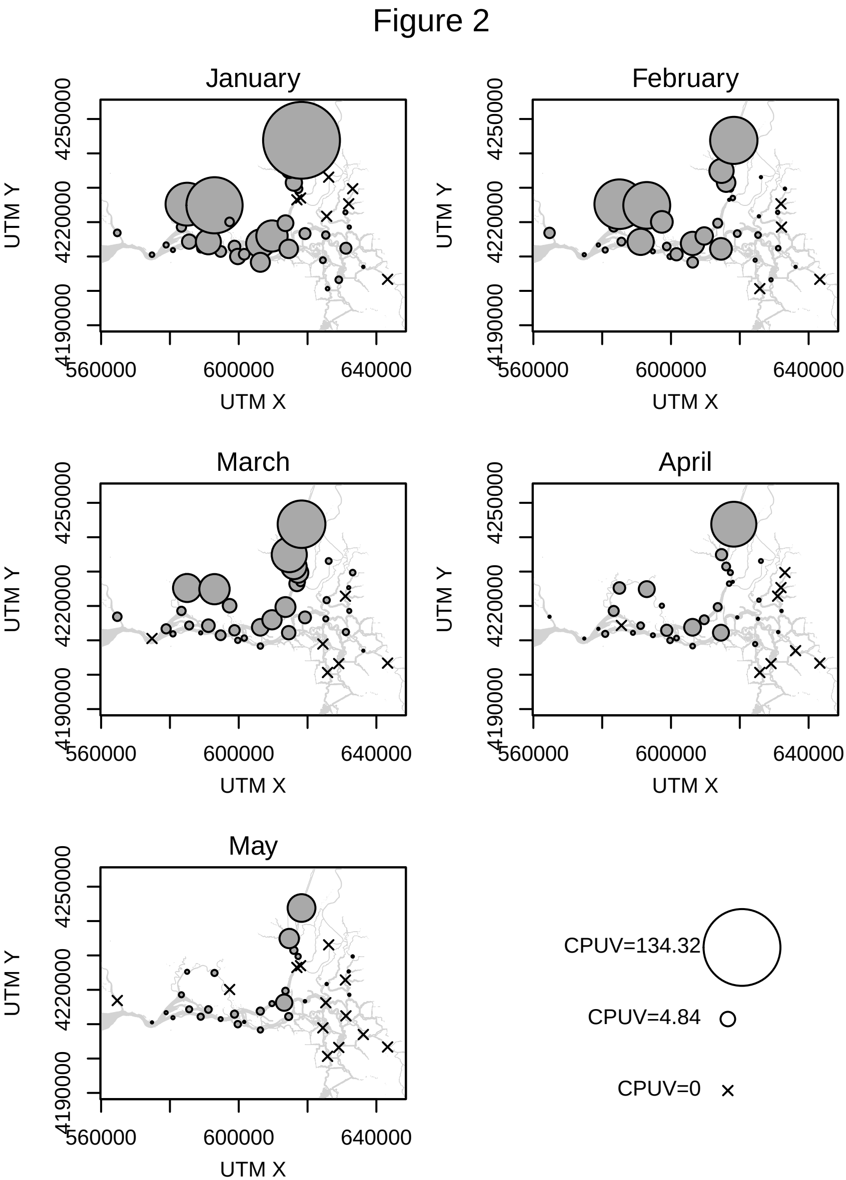


Figure 2
January
February
March
April
May
UTM Y
UTM X
4190000
4220000
4250000
560000
600000
640000
CPUV=134.32
CPUV=4.84
CPUV=0

Figure 3

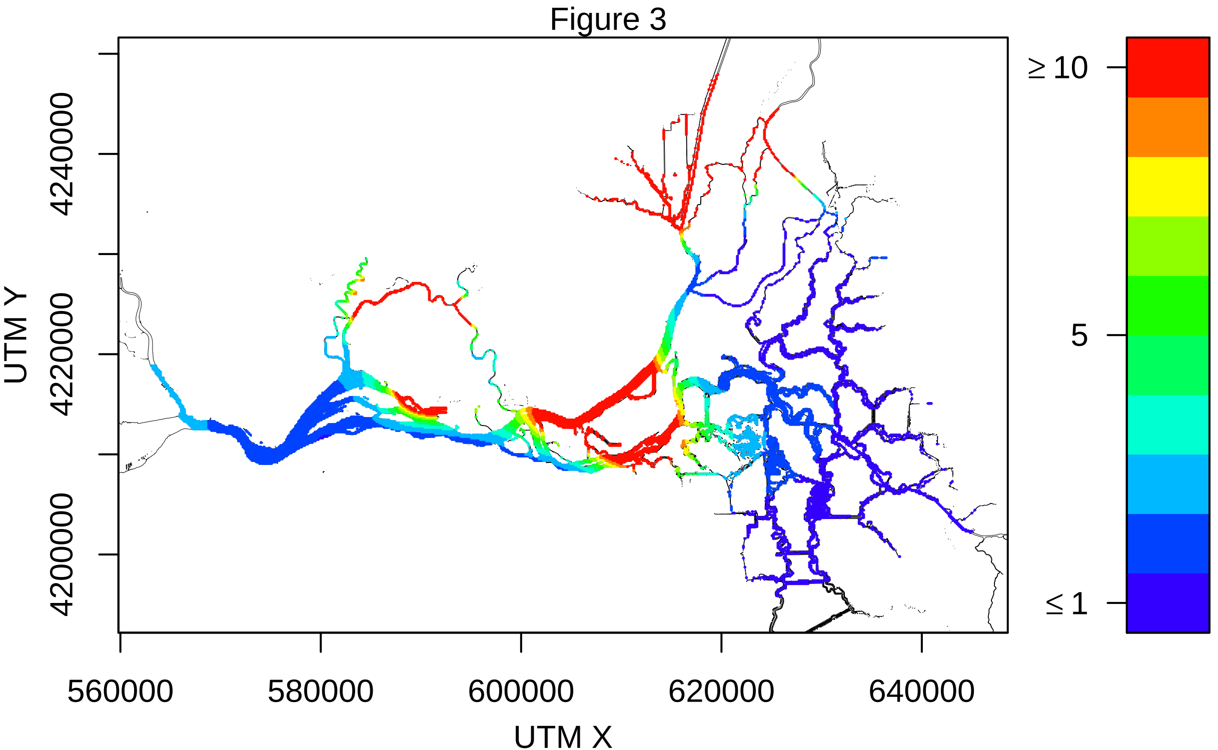


Figure 3
UTM Y
4200000
4220000
4240000
UTM X
560000
580000
600000
620000
640000
≥ 10
5
≤ 1

Figure 4

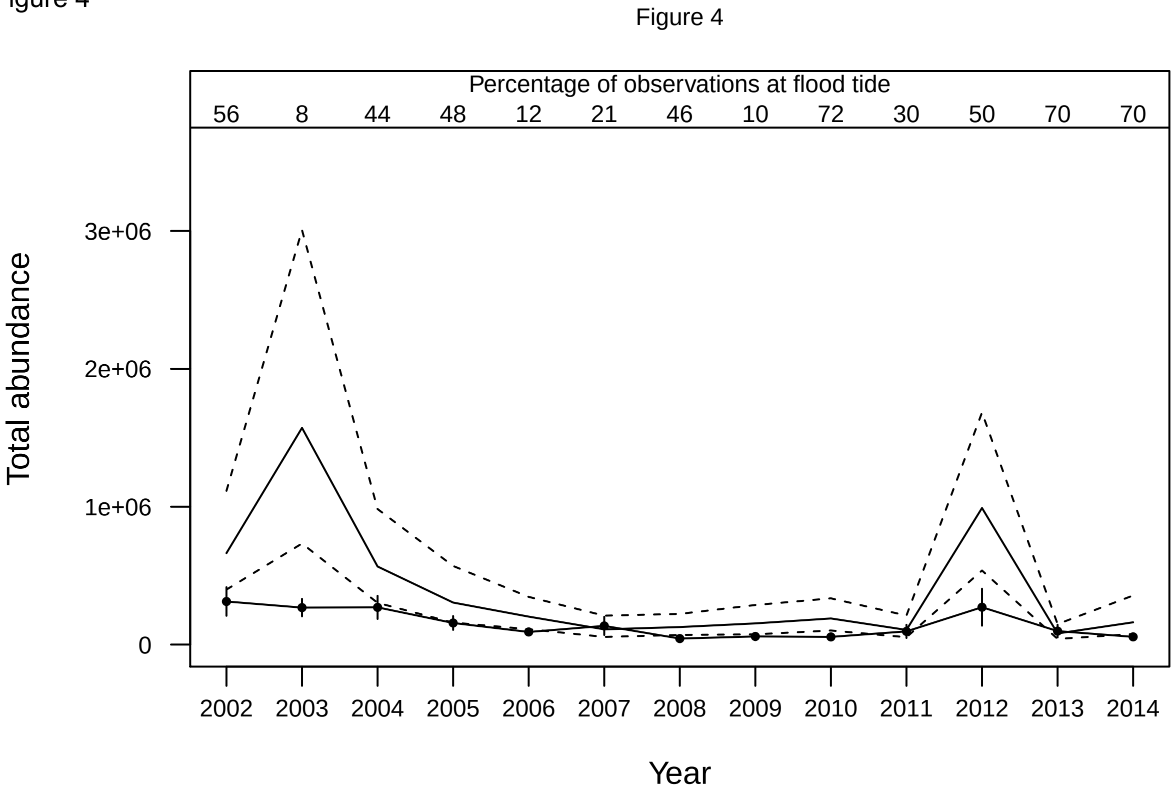


Figure 4
Percentage of observations at flood tide
56
8
44
48
12
21
46
10
72
30
50
70
70
3e+06
2e+06
1e+06
0
Total abundance
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year

Spatiotemporal models of an estuarine fish species to identify patterns and factors impacting their distribution and abundance

Leo Polansky, Ken B. Newman, Matthew L. Nobriga, Lara Mitchell

Supplementary Material

1 Survey locations, knots for smoothing basis, boundaries, and density prediction locations

Figure S1: Model smooth boundaries, knot locations, and prediction locations used in the analysis.

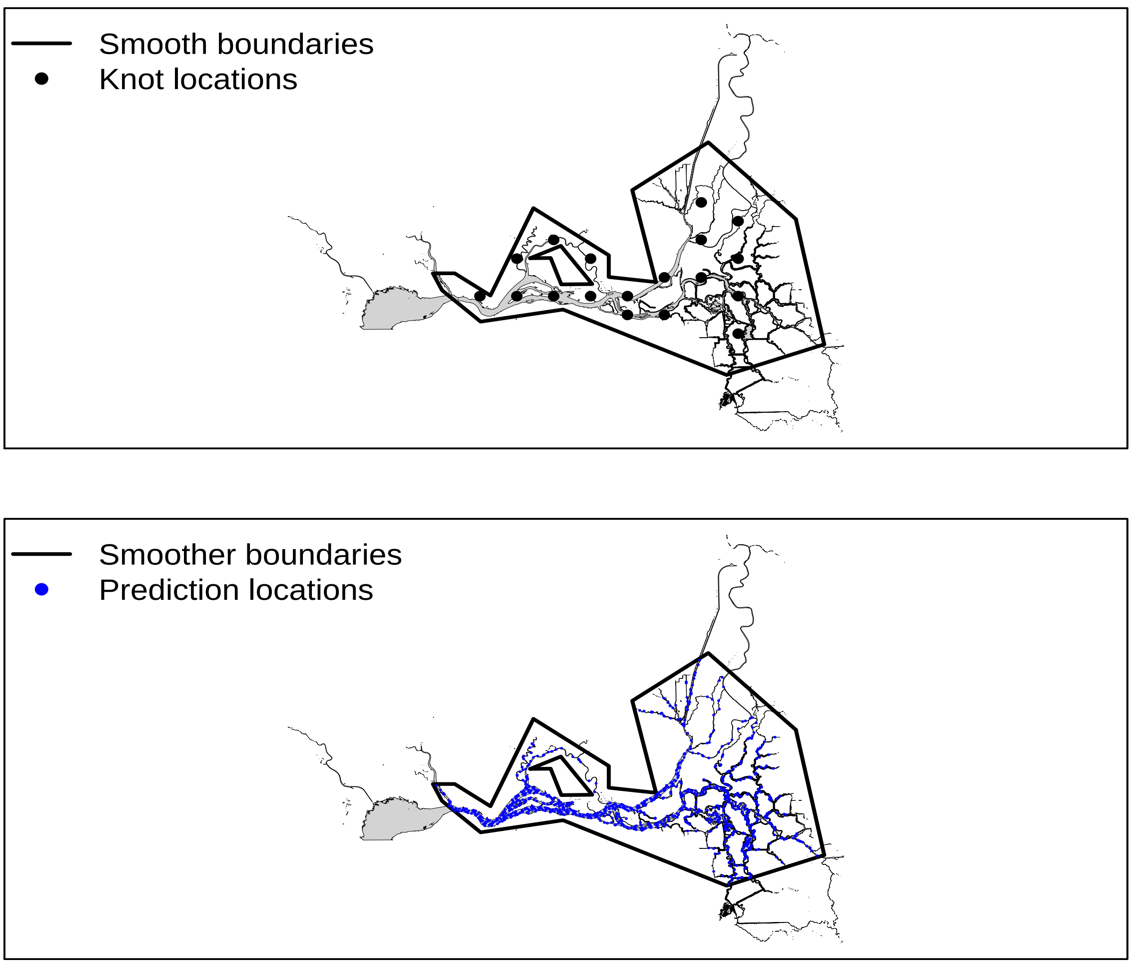
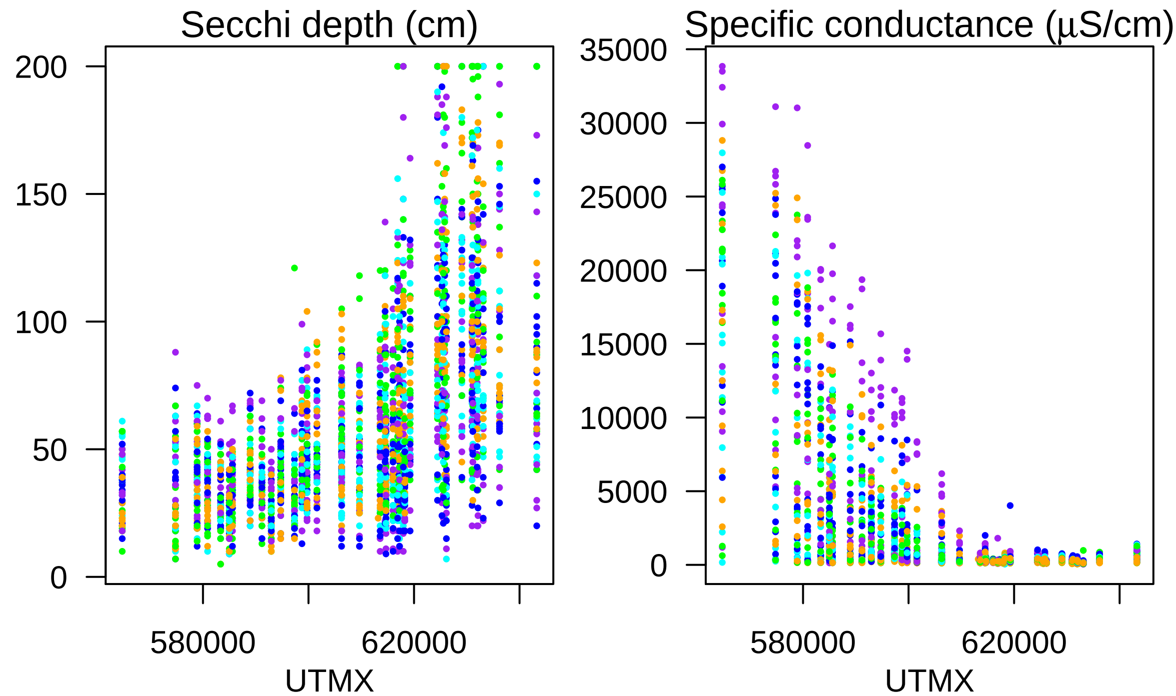


Figure S2: Tow specific values of Secchi and electrical conductance vs. UTMX. Points are colored by the month during which they were recorded: purple-January; blue-February; cyan-March; April-green; May-orange.



2 Locally measured covariates

A visual display of how turbidity and salinity vary in the UTMX direction, which corresponds approximately to an up and down estuary change, is shown in Figure S2.

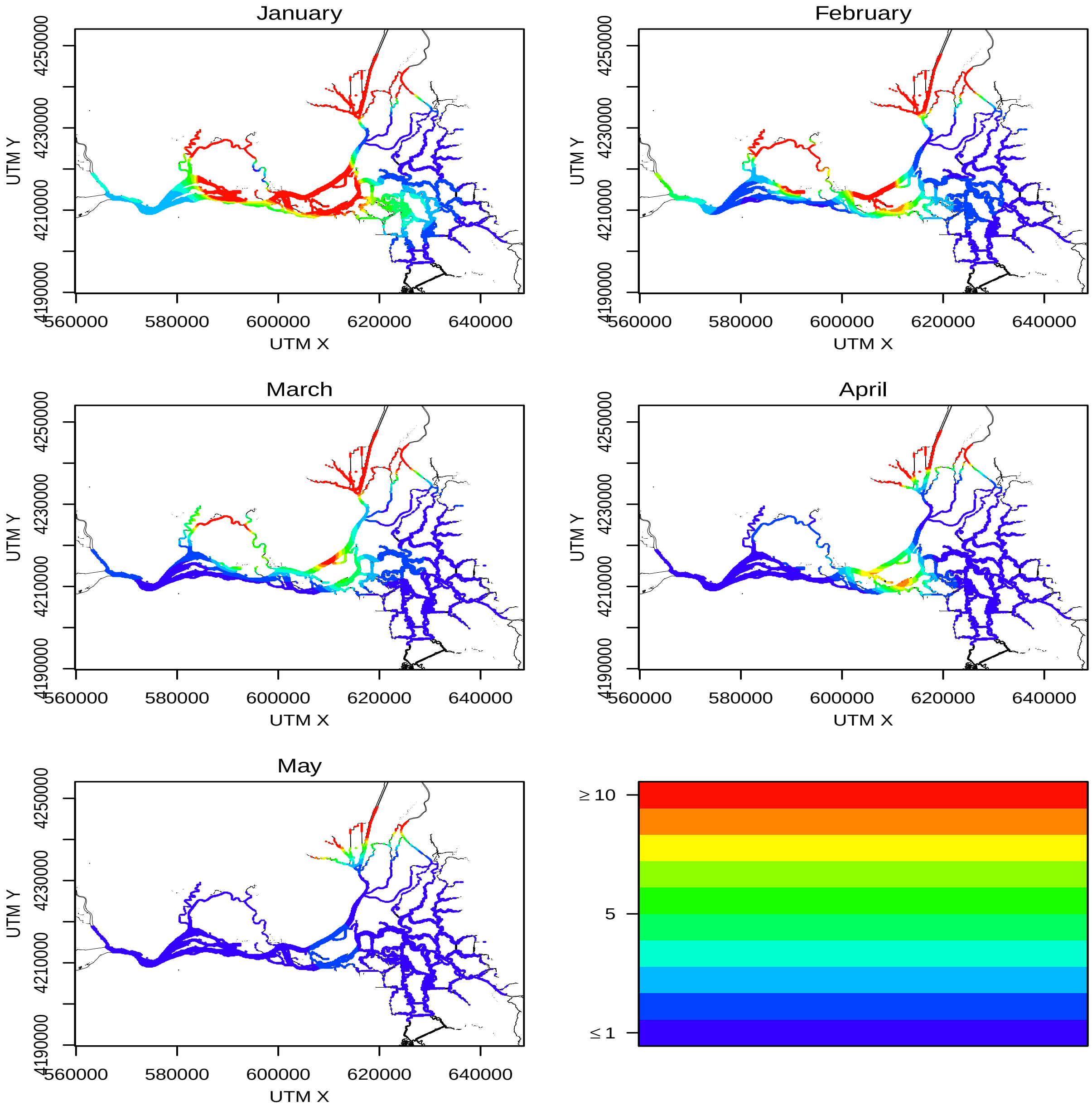
3 Intra-cohort distribution changes

Figure S3 shows density predictions using model 10 of the main text, which has a different spatial smooth for each month.

4 Total abundance estimates

Two distinctly different perspectives in sampling theory on making inferences from samples to populations are design-based inference and model-based inference (Thompson 2002). Design-based inference views the values on sampling units as fixed, non-random quantities, and the only randomness present is that induced by the sample selection pro-

Figure S3: Month specific density predictions based on model 10 (Table 1 of the main text) at a flood tide per $10000m^3$ using the mean Secchi and conductivity values; compare with figure 3 of the main text. Turbidity and salinity at each prediction point are set at their mean value, the Julian day is 15, 45, 74, 105, and 135 for the months of January, February, March, April, and May, respectively. As such these density maps emphasize the changes in density due to spatial and temporal changes. Color value map is shown in figure 3 of the main text.



cess. For example, assume in a body of water there are N cubic meter water volumes "plots", with plot i having y_i individual fish, $i = 1, ..., N$, and the inference objective is to estimate total population, $\tau_y = \sum_{i=1}^{N} y_i$. A simple random sample of size $n < N$ is drawn without replacement and τ_y is estimated by multiplying the sample average of y by N. The total population estimate is thus a random variable where the randomness arises solely from the random selection process.

In contrast, model-based inference views the values on sample units as realizations from some underlying random natural process. When the sample units are partitions of a spatial domain the random process often induces spatial correlation in the attributes defined on the units, e.g., adjacent plots are more likely to have similar values than more spatially separated plots. Inference is directed at estimating parameters that characterize the underlying random natural process, e.g., a mean value (μ), variance (σ^2), and covariance between plots i and j ($\sigma_{i,j}$). Realized population characteristics, e.g., τ_y, can still be estimated using estimates of the parameters of the random process, e.g. $\hat{\tau} = N\hat{\mu}$.

We note that strictly speaking, from a model-based inference perspective, the sample units do not need to be randomly selected for inference. However, it is our view that such additional human-induced randomization is advisable as it allows for comparison between model-based and design-based inference, and assessment of the sensitivity of assumptions made about the random process.

4.1 Design-based total abundance estimates

Design-based estimates of total monthly abundance N_{tot} (indices for month and year are suppressed for clarity) were calculated with historical SKT data by dividing the delta into 27 subregions (see Fig. S5) and carrying out volume expansions of average delta smelt catch densities at the subregion level. The average density in each subregion was calculated as the total catch divided by the total water volume sampled

$$\hat{\delta_h} = \frac{\sum_{j=1}^{m_h} Catch_{h,j}}{\sum_{j=1}^{m_h} Vol_{h,j}}$$

is the average density calculated over the m_h sampling locations in the subregion, $Catch_{h,j}$ is the catch in a single tow j in subregion h, and $Vol_{h,j}$ is the associated tow volume. The total abundance was calculated by expanding the subregion specific catch densities by the water volume in areas at least 2 meters deep down to 2 meters depth in subregion h, Vol_h, and then summing across all subregions the subregion specific totals $\hat{N}_h$,

$$\hat{N}_{tot} = \sum_{h=1}^{27} \hat{N}_h = \sum_{h=1}^{27} \hat{\delta_h} Vol_h$$

In some months not all 27 subregions were sampled by the SKT. In cases where subregion density estimates were missing due to lack of sampling, an estimate from a neighboring subregion was used for imputation.

For a given year, variance estimates for the total abundances are given by

$$Var\left(\hat{N}_{tot}\right) = \sum_{h=1}^{h=29} \left(\frac{Vol_h^2 \; s_h^2}{\left(\frac{1}{m_h} \sum_{j=1}^{n_h} Vol_{h,j}\right)^2 m_h} \right)$$

where

$$s_h^2 = \frac{\sum_{j=1}^{m_h} \left(Catch_{h,j} - \hat{\delta}_h Vol_{h,j}\right)^2}{m_h - 1}$$

is the variance contribution from each subregion. Some values of s_h^2 were missing because no sampling was done in a subregion or because a single site was sampled (in which case $m_h = 1$). In these cases, the median value of s_h^2, calculated over all available values, was used in place of missing values.

Suppressing time specific indices, confidence intervals were calculated for these abundance estimates by assuming the abundances $\hat{N}_{tot}$ were log-normally distributed. For a sample point estimate $\hat{N}_{tot}$ and variance of $Var\left(\hat{N}_{tot}\right)$, an estimate of the coefficient of variation is

$$\text{CV} = \frac{\sqrt{Var(\hat{N}_{tot})}}{\hat{N}_{tot}}$$

the location and scale parameters are

$$\mu = \log_e \left(\frac{\hat{N}_{tot}}{\sqrt{1 + \text{CV}^2}} \right)$$

and

$$\sigma = \sqrt{\log_e \left(1 + \text{CV}^2\right)}$$

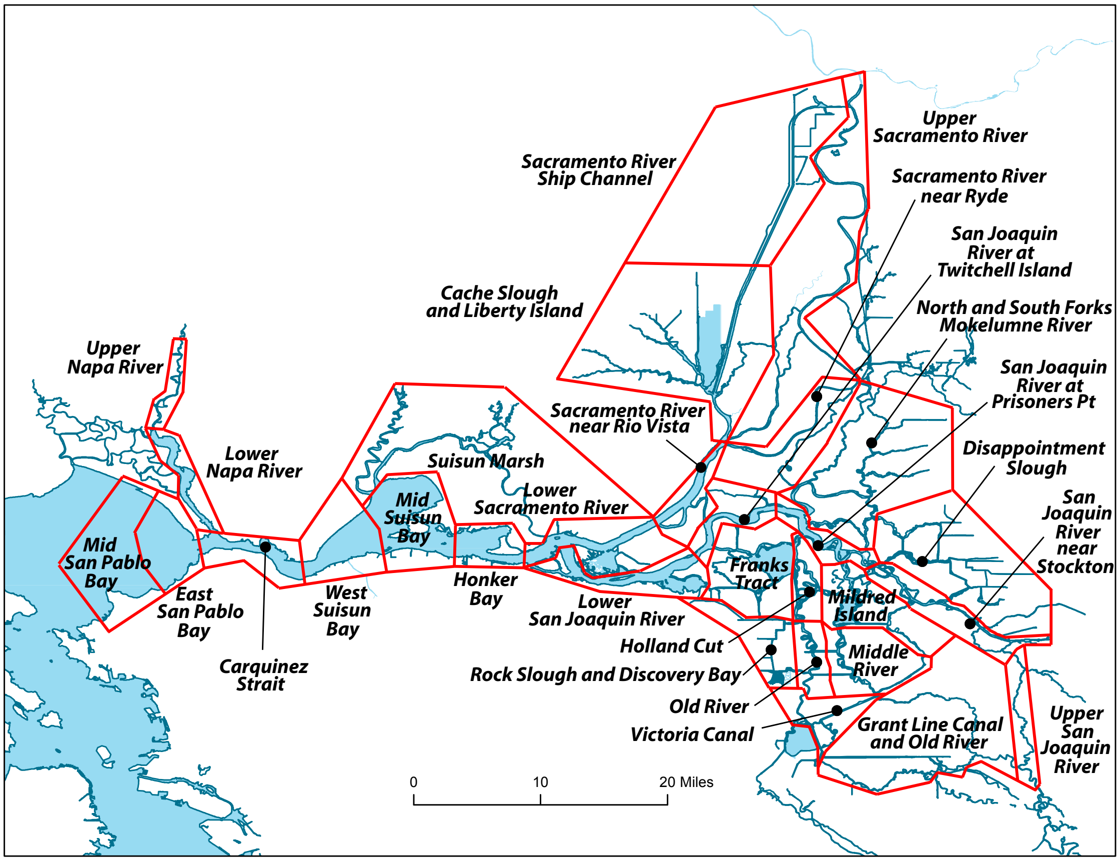
respectively. The natural log transformed abundance $z = \log_e(\hat{N}_{tot})$ is normally distributed with mean $\mu - \sigma^2/2$ and variance σ^2, where the mean is bias corrected so that the expected value of $\exp(z)$ is $\hat{N}_{tot}$. The 2.5 and 97.5 percentiles of z were exponentiated to estimate a 95% confidence interval for $\hat{N}_{tot}$.

4.2 Model-based total abundance estimates

Similar to the design-based approach using subregion specific average catch densities as the starting point for constructing a total abundance estimate, the model-based approach uses a model averaged density estimate per subregion to expand by subregion specific water volumes en route to obtaining a total abundance estimate. The month and year indices are suppressed for clarity. Denote the parameter vector of coefficients (including the coefficients for the smooth terms) from the fitted model by

$$\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_0, \ldots, \hat{\beta}_n\right]^\top$$

Figure S4: Spatial stratification of the Delta used for subregion based density expansions in the design-based estimates of abundance. The Mid and East San Pablo Bay subregions, along with Franks Tract, were excluded in total abundance calculations because these areas are not surveyed by the SKT.



and denote by $\mathbf{X}_i$ the design vector of values at location i from one of the lcoations shown in Fig S1. The model estimated abundance on the log scale at location i is

$$\hat{Y}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}} + \log_e (Vol_p)$$

where Vol_p is the prediction volume. The estimated number of fish $\hat{y}_i$ (per $Vol_p \text{m}^3$ of water) at location i on the response scale is

$$\hat{y}_i = \exp\left(\hat{Y}_i\right)$$

The mean density in subregion h is

$$\hat{\bar{\delta}}_h = \frac{\sum_{k=1}^{K_h} \hat{y}_i}{K_h Vol_p}$$

where the sum is over the K_h locations in subregion h. The estimate of the total abundance in subregion h is

$$\hat{N}_h = \hat{\bar{\delta}}_h Vol_h$$

where Vol_h is the total water volume in areas at least 2 meters deep down to 2 meters depth in subregion h. Again the total abundance estiamte is simply the sum of these subregion level estimates over all subregions

$$\hat{N}_{tot} = \sum_{h=1}^{27} \hat{N}_h$$

Prediction intervals for total abundance were obtained by parametric bootstrap and posterior simulation of GAM model coefficients, conditional on the smoothing parameter (Sections 4.8 and 5.4.2 in Wood (2006) and the scale parameters (the $\hat{\theta}$ from the catch model GAM and the $\hat{\sigma}$'s from the Secchi and conductivity GAMs). (Obtaining posterior distributions unconditional on the smoothing parameter, as outlined in Section 4.9.3 in Wood (2006) which involves wrapping the entire steps described next into a simulation-refitting process was not possible due to computational time.) Because estimates of total abundance are based on predicted values of Secchi depth and conductivity at each point location i, the first step is to simulate predictions of these covariate values at each location, and then, given these values, simulate location specific observations from the catch model.

For $b = 1, \ldots, B$, compute a bootstrapped predicted total abundance $N_{tot}^{(b)}$ by adding up the predicted catches $y_k^{(b)}$ at each location $k = 1, \ldots, K$ as follows

1. Prediction of covariate values. The fitted covariate models for Secchi and conductivity are of the form $z_k \sim N(\hat{\mu}_k, \hat{\sigma}^2)$, where $\hat{\mu}_k = \mathbf{W}_k \hat{\boldsymbol{\beta}}$. $\mathbf{W}_k$ is a $1 \times J$ row vector of the design matrix with values corresponding to the intercept and soap film smooth basis of the latitude and longitude dimensions at location k, $\hat{\boldsymbol{\beta}}$ is a $J \times 1$ column vector of the estimated GAM parameter vector with $J \times J$ covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$, and $\hat{\sigma}^2$ is the estimated observation variance. For each covariate, obtain a $K \times 1$ column vector $\mathbf{z_b}$ of values as follows:

   (a) Simulate a $\boldsymbol{\beta}^{(b)} \sim N(\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}})$.

   (b) Set $\mu_k^{(b)} = \mathbf{W}_k \boldsymbol{\beta^{(b)}}$

   (c) For $k = 1, \ldots, K$, simulate $z_k^{(b)} \sim N(\mu_k^{(b)}, \hat{\sigma}^2)$.

2. Construct the simulated covariate based design matrix $\mathbf{X}^{(b)}$ using the $\mathbf{z^{(b)}}$ values from step 1 for the Secchi and conductivity columns.

3. Prediction of catch given covariates. The catch model is of the form $y_k \sim NB(\hat{\lambda}_k, \hat{\theta})$, where $\log_e(\hat{\lambda}_k) = \mathbf{X}_k \hat{\boldsymbol{\beta}} + \log_e(Volume)$, Vol_p is the volume sampled, $\hat{\theta}$ is the estimated dispersion parameter of the negative binomial distribution, $\mathbf{X}_k$ is the design matrix, $\hat{\boldsymbol{\beta}}$ is the estimated model parameter vector and $\Sigma_{\hat{\boldsymbol{\beta}}}$ is its covariance matrix. Given $\mathbf{X}^{(b)}$, $\log_e(\lambda_k, b)$ depends only on the value of a realization of $\boldsymbol{\beta}^{(b)} \sim N(\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}})$. Viewed this way, $\log_e(\lambda_k^{(b)})$ are iid normal random variables with mean $\mathbf{X}_k \hat{\boldsymbol{\beta}} + \log_e(Volume)$ and variance

$$\tau^2_{\mu_k^{(b)}} = \sum_{j=1}^{J} (x_{k,j}^{(b)})^2 \mathrm{Var}(\hat{\beta}_j) + \sum_{1 \leq j < l \leq J} 2 x_{k,j}^{(b)} x_{k,l}^{(b)} \mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_l)$$

where $x_{k,j}^{(b)} = \mathbf{X}_{k,j}^{(b)}$. To simulate catch values per Vol_p of water,

(a) Simulate $\boldsymbol{\beta}^{(b)} \sim N(\hat{\boldsymbol{\beta}}, \Sigma_{\hat{\boldsymbol{\beta}}})$ and set $\mu_k^{(b)} = \mathbf{X}_k^{(b)} \boldsymbol{\beta}^{(b)} + \log_e(Vol_p)$.

(b) Compute the bias adjusted value $\mu_k^{(b,adj)} = \exp(\log_e(\mu_k^{(b)}) - \tau^2_{\mu_k^{(b)}}/2)$,

(c) For $k = 1, \ldots, K$, simulate $y_k^{(b)} \sim NB(\mu_k^{(b,adj)}, \hat{\theta})$

4. Compute subregion mean density, subregion abundance, and total abundance as

(a) The mean catch density in subregion h is

$$\bar{\delta_h}^{(b)} = \frac{\sum_{k \in h} y_k^{(b)} / K_h}{Vol_p}$$

where K_h is the number of prediction locations in h.

(b) The total predicted number of fish per subregion is $\hat{N}_h^{(b)} = \bar{\delta_h}^{(b)} Vol_h$

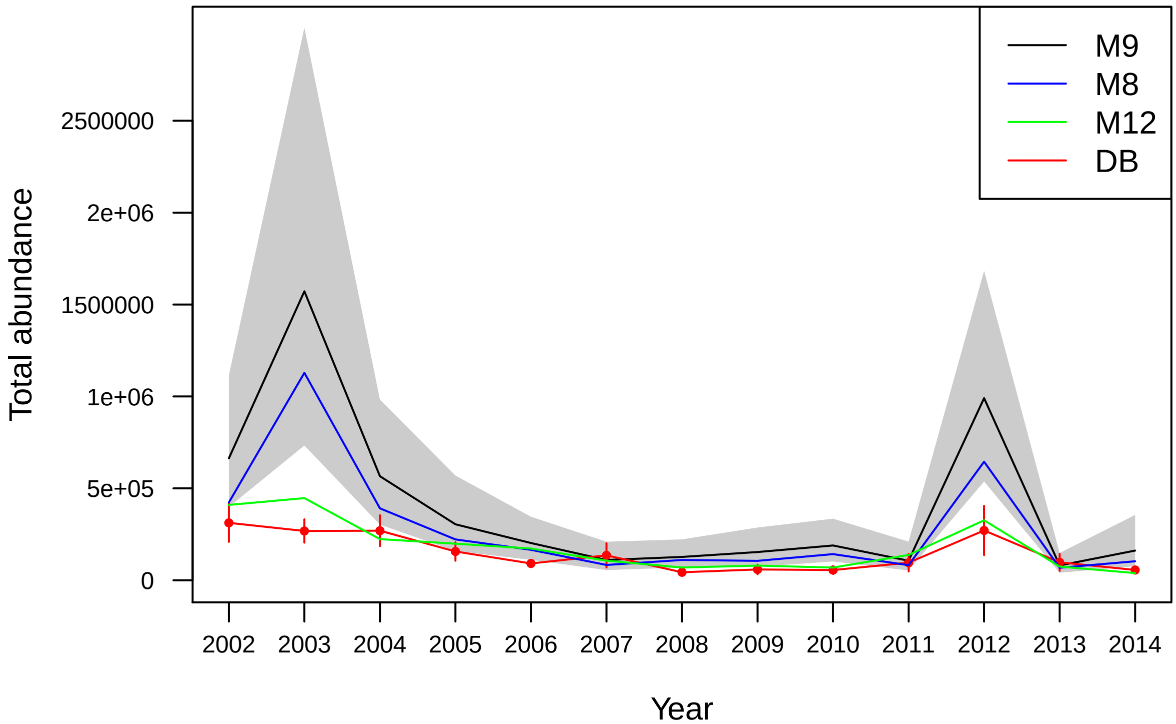
(c) The total predicted number of fish is

$$N_{tot}^{(b)} = \sum_{h=1}^{27} \hat{N}_h^{(b)}$$

Table S1: Abundance estimates and annual growth rates for February using design- and model-based approaches. Model based estimates are from Model 9 (Table 1 of the main text). Growth rates are year-over-year ratios.

	Model-based		Design-based	
Year	Abundance	Growth rate	Abundance	Growth rate
2002	312488		649028	
2003	268157	0.86	1479322	2.28
2004	269777	1.01	545347	0.37
2005	156633	0.58	291774	0.54
2006	91509	0.58	191509	0.66
2007	135563	1.48	100526	0.52
2008	43603	0.32	125398	1.25
2009	58877	1.35	142908	1.14
2010	55650	0.95	173163	1.21
2011	95682	1.72	86463	0.50
2012	271020	2.83	891304	10.31
2013	97707	0.36	74772	0.08
2014	56027	0.57	136596	1.83

Figure S5: Abundance estimates from model-based approaches using models M8, M9, and M12, and design-based. Grey shading shows the central 95% prediction interval for the M9 based predictions.



References

Thompson, Steven K. 2002. Sampling, Second Edition. John Wiley & Sons, Inc. New York, NY

Wood, Simon W. 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall, Boca Raton, FL

Appendix

R code used to fit the models and compute predictions and prediction intervals. Complete R code and input data available on request.

```
rm(list=ls())
library(MASS)
library(mgcv)
library(maptools)
library(proj4)
library(rgdal)
library(xtable)
library(rgeos)
library(car)
library(ncf)
library(geoR)
if(Sys.info()['sysname'] == 'Darwin'){
  library(parallel)
  }else{library(parallelsugar)}


data.root <- '~/smelt/gam-analyses/SKT-gam-analyses/Data/'

load(paste0(data.root,'SKT-2002-2014-gam-analysis-data-prep-v7.RData'))


# Time consuming model fits
load.M8.M12 <- TRUE
if(load(load.M8.M12)){
  load(file=paste0(data.root,'SKT-2002-2014-gam-analysis-soap-v9-M8.RData'))
  load(file=paste0(data.root,'SKT-2002-2014-gam-analysis-soap-v12-M8.RData'))
}

# M12
fit.M12 <- FALSE
if(fit.M12){
  m.regional.local.by.month.year.formula <- as.formula("smelt~offset(logVol)+
            fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fCohort.year.month)
                                                          +Secchi.z+Cond.z+Tide")
  m.regional.local.by.month.year <- gam(m.regional.local.by.month.year.formula,
                                        family=nb(),
                                        knots=knots,data=ds,method="ML")
  save(list=ls(),
       file=paste0(data.root,'SKT-2002-2014-gam-analysis-soap-v9-M12.RData')
```

```
}

# M8
fit.M8 <- FALSE
if(fit.M8){
  m.regional.by.month.year.formula <- as.formula("smelt~offset(logVol)+
        fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fCohort.year.month)")
  m.regional.by.month.year <- gam(m.regional.by.month.year.formula,family=nb(),
                                  knots=knots,data=ds,method="ML")
  save(list=ls(),
       file=paste0(data.root,'SKT-2002-2014-gam-analysis-soap-v9-M8.RData')
}

# Remaining models
#VIF model
r.vif <- glm.nb(smelt~offset(logVol)+Secchi.z+Cond.z+Tide+Month+SubRegion,
                data=ds)
vif(r.vif)

r.vif <- glm.nb(smelt~offset(logVol)+Secchi.z+Cond.z+Tide+Month+Lon.z+Lat.z,
                data=ds)
vif(r.vif)

# 1) Global: no regional (smooth), no local
m.global <- glm.nb(smelt~offset(logVol)+fCohort.year*JD,data=ds)

# Global + one local to estimate the best a particular local covariate can do
m.global.plus.Secchi <- glm.nb(smelt~offset(logVol)+fCohort.year*JD+Secchi.z,data=ds)
m.global.plus.cond <- glm.nb(smelt~offset(logVol)+fCohort.year*JD+Cond.z,data=ds)
m.global.plus.tide <- glm.nb(smelt~offset(logVol)+fCohort.year*JD+Tide,data=ds)

# 2) Global x regional: no by in smooth
m.regional.formula <- as.formula("smelt~offset(logVol)+
                              fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb))")
t1 <- Sys.time()
m.regional <- gam(m.regional.formula,family=nb(),
                  knots=knots,data=ds,method="ML")
t2 <- Sys.time()
difftime(t2,t1)

# 3) Global x regional: by month
m.regional.by.month.formula <- as.formula("smelt~offset(logVol)+
                    fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fmonth)")

t1 <- Sys.time()
```

```
m.regional.by.month <- gam(m.regional.by.month.formula,family=nb(),
                           knots=knots,data=ds,method="ML")
t2 <- Sys.time()
difftime(t2,t1)


# 4) Global x regional: by year (cohort)
m.regional.by.year.formula <- as.formula("smelt~offset(logVol)+
              fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fCohort.year)")

t1 <- Sys.time()
m.regional.by.year <- gam(m.regional.by.year.formula,family=nb(),
                          knots=knots,data=ds,method="ML")
t2 <- Sys.time()
difftime(t2,t1)

# 6) Global x regional x local: No by
m.regional.local.formula <- as.formula("smelt~offset(logVol)+
          fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb))+Secchi.z+Cond.z+Tide")

t1 <- Sys.time()
m.regional.local <- gam(m.regional.local.formula,family=nb(),
                        knots=knots,data=ds,method="ML")
t2 <- Sys.time()
difftime(t2,t1)

# 7) Global x regional x local: by month
m.regional.local.by.month.formula <- as.formula("smelt~offset(logVol)+
                      fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fmonth)
                      +Secchi.z+Cond.z+Tide")

t1 <- Sys.time()
m.regional.local.by.month <- gam(m.regional.local.by.month.formula,
                                 family=nb(),knots=knots,data=ds,method="ML")
t2 <- Sys.time()
difftime(t2,t1)

# 8) Global x regional x local: by year (cohort)
m.regional.local.by.year.formula <- as.formula("smelt~offset(logVol)+
                fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb),by=fCohort.year)
                +Secchi.z+Cond.z+Tide")

t1 <- Sys.time()
m.regional.local.by.year <- gam(m.regional.local.by.year.formula,family=nb(),
                                knots=knots,data=ds,method="ML")
```

```
t2 <- Sys.time()
difftime(t2,t1)


m <- m.regional.local

# Drop one local cov at a time to look at
# proportion deviance explained by adding this cov to a global X regional model
m.regional.local.minus.Secchi <- gam(smelt~offset(logVol)+
                                fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb))+
                                Cond.z+Tide,family=nb(),
                                sp=m$sp,knots=knots,data=ds,method="ML")

m.regional.local.minus.cond <- gam(smelt~offset(logVol)+
                                fCohort.year*JD+s(x,y,bs='so',xt=list(bnd=fsb))+
                                Secchi.z+Tide,family=nb(),
                                sp=m$sp,knots=knots,data=ds,method="ML")

m.regional.local.minus.tide <- gam(smelt~offset(logVol)+fCohort.year*JD+
                                     s(x,y,bs='so',xt=list(bnd=fsb))+
                                     Secchi.z+Cond.z,family=nb(),
                                   sp=m$sp,knots=knots,data=ds,method="ML")

##########--------- Make predictions of abundance in Feb ---------##########
# Make predictions at each grid location on Feb 15th
# for flood and ebb tides and bootstrap prediction intervals
ucym <- as.character(unique(ds$fCohort.year.month))
index.Feb <- which(as.numeric(unlist(lapply(as.character(ucym),
                        FUN=function(x){y=strsplit(x,split="-")[[1]][3]})))==2)
ucym.Feb <- ucym[index.Feb]

# Make predictions on grid-
# Why doesn't crs(DSLCM.SubRegions) or CRS(DSLCM.SubRegions) work here?
# Why does crs work on a Windows PC? Or does it?
spatial.grid.predict <- SpatialPoints(grid.predict,
                          proj4string=attributes(DSLCM.SubRegions)$proj4string)
grid.predict.with.subregions <- cbind(grid.predict,
                                  over(spatial.grid.predict,DSLCM.SubRegions))

# Prediction volume
vol.p <- 10000

# Don't fix boundary at 0 for covariates
fsb.cov <- vector("list",1)
fsb.cov[[1]]$x <- region.boundary[,"x"]
```

```
fsb.cov[[1]]$y <- region.boundary[,"y"]

grid.cov.gam.func <- function(ucym,dat,cov){
  # Get subregion averages of a given covariate for a dataset dt
  dt <- subset(dat,fCohort.year.month==ucym)
  names(dt)[which(names(dt)=="UTMX")] <- "x"
  names(dt)[which(names(dt)=="UTMY")] <- "y"
  cov.temp <- dt[,cov]
  if(nrow(dt)>27){
    cov.gam <- try(gam(cov.temp~s(x,y,bs="so",xt=list(bnd=fsb.cov)),
                       knots=knots,data=dt,method="ML"))
  }else{
    cov.gam <- try(gam(cov.temp~s(x,y,k=25),data=dt,method="ML"))
  }
  return(cov.gam)
}

Secchi.z.models <- lapply(as.character(ucym),
                          FUN=grid.cov.gam.func,dat=ds,cov='Secchi.z')
Cond.z.models <- lapply(as.character(ucym),
                        FUN=grid.cov.gam.func,dat=ds,cov='Cond.z')

Secchi.z.gam.gof <- unlist(lapply(Secchi.z.models,
                                  FUN=function(x){summary(x)$dev.expl}))
Cond.z.gam.gof <- unlist(lapply(Cond.z.models,
                                FUN=function(x){summary(x)$dev.expl}))

range(Secchi.z.gam.gof[index.Feb])
median(Secchi.z.gam.gof[index.Feb])
quantile(Secchi.z.gam.gof[index.Feb],probs=seq(.1,1,by=.1))
range(Cond.z.gam.gof[index.Feb])
median(Cond.z.gam.gof[index.Feb])
quantile(Cond.z.gam.gof[index.Feb],probs=seq(.1,1,by=.1))

grid.cov.gam.pred.func <- function(m,gcv.est=TRUE){
  # Sample a prediction from a fitted GAM model m
  # Returns a prediction at each location of the grid
  # If gcv.est=F, prediction includes uncertainty in the
  #    model coefficients and observation error
  # Prediction does not include uncertainty in the smoothing paramter
  data.new <- data.frame(x=grid.predict$x,y=grid.predict$y)
  if(gcv.est){
    y <- predict(m,newdata=data.new,type='response')
  }else{
    beta <- coef(m)
```

```
    Vb <- m$Vc
    Cv <- chol(Vb)
    n.rep=1
    nb <- length(beta)
    br <- t(Cv) %*% matrix(rnorm(n.rep*nb),nb,n.rep) + beta
    Xp <- predict(m,newdata=data.new,type="lpmatrix")
    lp <- Xp %*% br
    y <- rnorm(length(lp),mean=lp,sd=sqrt(m$sig2))
  }
  return(y)
}

cpue.newdata.grid.func <- function(fcym,tide.set,gcv.est){
  index.temp <- which(ucym==fcym)
  Secchi.z.temp <- grid.cov.gam.pred.func(m=Secchi.z.models[[index.temp]],
                                          gcv.est=gcv.est)
  Cond.z.temp <- grid.cov.gam.pred.func(m=Cond.z.models[[index.temp]],
                                        gcv.est=gcv.est)

  # Sets up a data frame of new data for making CPUE predictions
  Month <- as.numeric(strsplit(as.character(fcym),split="-")[[1]][3])
  if(Month==1){JD=15}
  if(Month==2){JD=45}
  if(Month==3){JD=74}
  if(Month==4){JD=105}
  if(Month==5){JD=135}

  dn <- data.frame(
    logVol=log(vol.p),
    fCohort.year=factor(paste(
      strsplit(as.character(fcym),split="-")[[1]][1:2],collapse="-"),
      levels=levels(ds$fCohort.year)),
    fCohort.year.month=factor(fcym,levels=levels(ds$fCohort.year.month)),
    JD=JD,
    fmonth=strsplit(fcym,split="-")[[1]][3],
    Secchi.z=Secchi.z.temp,
    Cond.z=Cond.z.temp,
    Tide=tide.set,
    x=grid.predict$x,
    y=grid.predict$y
  )
  return(dn)
}

beta.param.vect.sample.from.gam <- function(m,b){
```

```
  # m a fitted GAM, returns a J x b column vector of
  #  a samples of beta, J=length(beta)
  beta <- coef(m)
  Vb <- m$Vc
  Cv <- chol(Vb)
  n.rep <- b
  nb <- length(beta)
  br <- t(Cv) %*% matrix(rnorm(n.rep*nb),nb,n.rep) + beta
  return(br)
}

var.sum.func <- function(a,x,Sigma){
  # a, x- vectors of same length; Sigma- covariance matrix of x
  # Let X=(a_1*x_1,...,a_n*x_n)
  # Computes the variance of the of sum of the elements of X
  # Var(sum(X))=sum_i a_i^2*Var(x_i)+2*sum_1<=i<j<=n a_i*a_j*Cov(x_i,x_j)

  ai.aj <- combn(a,m=2,prod) # Get all a_i*a_j products for 1<=i<j<=n

  # Get index of cov(x_i,x_j) values in same order as ai.aj vector
  off.diag.index <- combn(1:ncol(Sigma),m=2) #All possible combinations of 1,...,n
  off.diag.index <- cbind(off.diag.index[1,],off.diag.index[2,])

  # Get the Cov(x_i,x_j) terms
  off.diag.var.cov <- Sigma[off.diag.index]

  # Compute variance of the sum
  r <- sum(a^2*diag(Sigma))+sum(2*ai.aj*off.diag.var.cov)
  return(r)
}

boot.pred.func <- function(m,fcym,tide.set,boot){
  # ****Get covariance matrix of beta from m
  # For bias correcting mu_k samples in the boot loop.
  #   See **** Do this here for speed.
  # Actual variance will depend on Xp.boot so need to wait till boot
  #    loop to finish computing
  off.diag.index <- combn(1:ncol(m$Vc),m=2) #All possible combinations of 1,..,n
  off.diag.index <- cbind(off.diag.index[1,],off.diag.index[2,])
  off.diag.var.cov <- m$Vc[off.diag.index]
  v.m <- diag(m$Vc)

  # Predictions at estimated parameters
  data.new <- cpue.newdata.grid.func(fcym=fcym,tide.set=tide.set,gcv.est=T)
  Xp <- predict(m,newdata=data.new,type="lpmatrix")
```

```
mu.pred.linear <- Xp %*% coef(m)+log(vol.p)
#GLM models don't bias correct when making prediction from log link models
mu.pred <- exp(mu.pred.linear)

mu.pred.mean.sr <- tapply(mu.pred,grid.predict.with.subregions$SubRegion,mean)

index.match <- match(wv$SubRegion,names(mu.pred.mean.sr))

tot.pop.size <- sum(mu.pred.mean.sr[index.match]*wv$twom,na.rm=T)/vol.p

if(boot==0){
  return(list(
    tot.pop.size=tot.pop.size,
    mean.pop.boot=NA,
    tot.pop.pred.boot.interval=NA
  ))
}else{
  boot.tot.pop.size <- rep(NA,boot)
  theta.est <- m$family$getTheta(TRUE)
  for(i in 1:boot){
    # Step 1- simulate from covariate data models Xp_boot
    boot.data.new <- cpue.newdata.grid.func(fcym=fcym,tide.set=tide.set,
                                            gcv.est=F)
    Xp.boot <- predict(m,newdata=boot.data.new,type="lpmatrix")

    # Sample a beta_b from N(hat(beta),Sigma_hat(beta))
    beta.samp <- beta.param.vect.sample.from.gam(m=m,b=1)

    # Make linear predictor using sample beta and sample covariates
    boot.mu.pred.linear <- Xp.boot %*% beta.samp+log(vol.p)

    # View log(tau_b)=boot.mu.pred.linear as a normally distributed variable
    # tau_b ~ LN(logmean=Xp.boot %*% hat(beta)+log(vol.p),
    #    varlog= Var(Xp.boot %*% hat(beta)+log(vol.p))=Var(Xp.boot %*% hat(beta))
    # Then bias correct exp(log(tau_b))
    # Bias correct assuming boot.mu ~ LN with mean=boot.mu.pred,
    #    variance=sigma^2
    # mu_i=1*beta_0+Xp1[i,1]*beta1+...+Xp[i,n]*beta_n
    # mu=beta_0+x1*beta_1+x2*beta_2+...+xn*beta_n
    # Var(mu)=sum_over_i x_i^2*var(beta_i)+
    #     sum_over_i*sum_over_j x_i*x_j*Cov(beta_i,beta_j) covariance of sums formula
    # Var(mu)=
    #  sum_over_i x_i^2*var(beta_i)+2*sum_1<=i<j<=N x_i*x_j*cov(beta_i,beta_j)
    # ****Have covariance matrix of beta
    sig2 <- rep(NA,length(boot.mu.pred.linear))
```

```
      for(k in 1:length(boot.mu.pred.linear)){
        ai.aj <- combn(Xp.boot[k,],m=2,prod) #All a_i*a_j products for 1<=i<j<=n
        sig2[k] <- sum(Xp.boot[k,]^2*v.m)+sum(2*ai.aj*off.diag.var.cov)
        #sig2[k] <- var.sum.func(a=Xp.boot[k,],x=coef(m),Sigma=m$Vc)
      }
      # End bias correct

      boot.mu.pred <- exp(boot.mu.pred.linear-sig2/2)
      boot.pred <- rnegbin(boot.mu.pred,theta=theta.est)
      boot.dens.pred <- boot.pred/vol.p

      boot.mean.dens.sr <- tapply(
        boot.pred,grid.predict.with.subregions$SubRegion,mean)

      boot.tot.pop.size[i] <- sum(
        boot.mean.dens.sr[index.match]*wv$twom,na.rm=T)/vol.p
    }
    return(list(
      tot.pop.size=tot.pop.size,
      mean.pop.boot=mean(boot.tot.pop.size),
      tot.pop.pred.boot.interval=quantile(
        boot.tot.pop.size,probs=c(.025,.25,.5,.75,.975))
    ))
  }
}

# t1 <- Sys.time()
# e=boot.pred.func(m=m,fcym=ucym[2],tide.set="Flood",boot=2)
# t2 <- Sys.time()
# difftime(t2,t1)

# Check on understanding p1 should equal p1.alt
# t1 <- Sys.time()
# i=1
# fcym.temp <- as.character(ucym[i])
# data.new <- cpue.newdata.grid.func(fcym=fcym,tide.set=tide.set,gcv.est=T)
# p1 <- predict(m,newdata=data.new,type="response")
# Xp.alt <- predict(m,newdata=data.new,type="lpmatrix")
# p1.alt <- exp(Xp.alt %*% (coef(m))+log(vol.p))
# max(abs(p1-as.numeric(p1.alt)))
# t2 <- Sys.time()
# difftime(t2,t1)

# Point estimates and uncertainty using model 9
pop.estimate.posterior.sim.feb.func <- function(X,tide,boot){
```

```
  r <- boot.pred.func(m=m,fcym=X,tide.set=tide,boot=boot)
  return(r)
}

boot.set <- 4 #1000
cor.set <- 4

t1 <- Sys.time()
p.store.list.ebb.feb <- mclapply(X=ucym.Feb,
                               FUN=pop.estimate.posterior.sim.feb.func,
                               tide="Ebb",boot=boot.set,mc.cores=cor.set)
t2 <- Sys.time()
difftime(t2,t1)

boot.set <- 1000
t1 <- Sys.time()
p.store.list.flood.feb <- mclapply(X=ucym.Feb,
                                 FUN=pop.estimate.posterior.sim.feb.func,
                                 tide="Flood",boot=boot.set,mc.cores=cor.set)
t2 <- Sys.time()
difftime(t2,t1)

save(list=ls(),file=paste0(data.root,'SKT-2002-2014-gam-analysis-soap-v10.RData'))

# Point estimates using model 10
pop.estimate.posterior.sim.feb.func.M10 <- function(X,tide,boot){
  r <- boot.pred.func(m=m.regional.local.by.month,fcym=X,tide.set=tide,boot=boot)
  return(r)
}
p.store.list.flood.feb.M10 <- lapply(X=ucym.Feb,
                                     FUN=pop.estimate.posterior.sim.feb.func.M10,
                                     tide="Flood",boot=0)

# Point estimates using model 12
pop.estimate.posterior.sim.feb.func.M12 <- function(X,tide,boot){
  r <- boot.pred.func(m=m.regional.local.by.month.year,fcym=X,tide.set=tide,
                      boot=boot)
  return(r)
}
p.store.list.flood.feb.M12 <- lapply(X=ucym.Feb,
                                     FUN=pop.estimate.posterior.sim.feb.func.M12,
                                     tide="Flood",boot=0)

pop.est.ebb.feb <- data.frame(
  est=unlist(lapply(p.store.list.ebb.feb,FUN=function(x){x$tot.pop.size})),
```

```
  mean=unlist(lapply(p.store.list.ebb.feb,FUN=function(x){x$mean.pop.boot})),
  median=unlist(lapply(p.store.list.ebb.feb,FUN=function(x){
    x$tot.pop.pred.boot.interval[’50%’]})),
  lower=unlist(lapply(p.store.list.ebb.feb,FUN=function(x){
    x$tot.pop.pred.boot.interval[’2.5%’]})),
  upper=unlist(lapply(p.store.list.ebb.feb,FUN=function(x){
    x$tot.pop.pred.boot.interval[’97.5%’]})))
pop.est.flood.feb <- data.frame(
  est=unlist(lapply(p.store.list.flood.feb,FUN=function(x){x$tot.pop.size})),
  mean=unlist(lapply(p.store.list.flood.feb,FUN=function(x){x$mean.pop.boot})),
  median=unlist(lapply(p.store.list.flood.feb,FUN=function(x){x$mean.pop.boot})),
  lower=unlist(lapply(p.store.list.flood.feb,FUN=function(x){
    x$tot.pop.pred.boot.interval[’2.5%’]})),
  upper=unlist(lapply(p.store.list.flood.feb,FUN=function(x){
    x$tot.pop.pred.boot.interval[’97.5%’]})))

pop.point.estimate.ebb.feb <- unlist(lapply(p.store.list.ebb.feb,
                                            FUN=function(x){x$tot.pop.size}))
pop.point.estimate.flood.feb <- unlist(lapply(p.store.list.flood.feb,
                                              FUN=function(x){x$tot.pop.size}))
pop.point.estimate.flood.feb.M10 <- unlist(lapply(p.store.list.flood.feb.M10,
                                            FUN=function(x){x$tot.pop.size}))
pop.point.estimate.flood.feb.M12 <- unlist(lapply(p.store.list.flood.feb.M12,
                                              FUN=function(x){x$tot.pop.size}))

plot(pop.point.estimate.flood.feb,
     pop.point.estimate.flood.feb.M10,type=’n’,xlab=’’,ylab=’’)
abline(a=0,b=1)
title(xlab=’Model 9 (single smooth for all months)’,line=2.2)
title(ylab=’Model 10 (month specific smooth)’,line=2.2)
text(pop.point.estimate.flood.feb,pop.point.estimate.flood.feb.M10,
     labels=sapply(uy,FUN=function(x){substr(x,start=3,stop=4)}))
cor(pop.point.estimate.flood.feb,pop.point.estimate.flood.feb.M10)
round(100*(pop.point.estimate.flood.feb.M10-pop.point.estimate.flood.feb)/
        pop.point.estimate.flood.feb,2)

ab=data.frame(Year=uy,M9=pop.point.estimate.flood.feb,
              M10=pop.point.estimate.flood.feb.M10,
              M12=pop.point.estimate.flood.feb.M12)
print(ab,row.names=F)

ab=data.frame(
  Coef=names(coef(m)[1:14]),
  M9=coef(m)[1:14],
  M10=coef(m.regional.local.by.month)[1:14],
```

```
  M12=coef(m.regional.local.by.month.year)[1:14]
)
print(ab,row.names=F)

p.f <- function(x,p.t,col.set){
  y.lim <- c(-10,3)
  plot(x,type='n',ylim=y.lim,xaxt='n',xlab='',ylab='')
  axis(side=1,at=x,labels=rownames(p.t),las=2,cex.axis=.85)

  for(i in 1:nrow(p.t)){
    points(x[i],p.t[i,'Estimate'],col=col.set,pch=20)
    lines(rep(x[i],2),c(p.t[i,'Estimate']-p.t[i,'Std. Error'],
                        p.t[i,'Estimate']+p.t[i,'Std. Error']),col=col.set)
  }
}
p.f2 <- function(x,p.t,col.set){
  for(i in 1:nrow(p.t)){
    points(x[i],p.t[i,'Estimate'],col=col.set,pch=20)
    lines(rep(x[i],2),c(p.t[i,'Estimate']-p.t[i,'Std. Error'],
                        p.t[i,'Estimate']+p.t[i,'Std. Error']),col=col.set)
  }
}

par(mar=c(10,3,2,1))
p.f(x=seq(1,31,by=1),p.t=summary(m.regional.local)$p.table,col.set='blue')
p.f2(x=seq(1.1,31.1,by=1),p.t=summary(m.regional.local.by.month)$p.table,
     col.set='red')
p.f2(x=seq(1.2,31.2,by=1),p.t=summary(m.regional.local.by.month.year)$p.table,
     col.set='green')
legend('bottomright',legend=c('M9','M10','M12'),
       col=c('blue','red','green'),pch=20)

p.t.M9=summary(m)$p.table
p.t.M12=summary(m.regional.local.by.month.year)$p.table

delta0.M9 <- exp(c(p.t.M9[1,'Estimate'],
                   p.t.M9[1,'Estimate']+p.t.M9[2:13,'Estimate']))
delta0.M12 <- exp(c(p.t.M12[1,'Estimate'],
                    p.t.M12[1,'Estimate']+p.t.M12[2:13,'Estimate']))

print(data.frame(cohort=cohort,M9.delta0=delta0.M9,
                 M12.delta0=delta0.M12,ratio=delta0.M12/delta0.M9),row.names=F)

data.frame(cohort=cohort,M9.delta0=delta0.M9,
           M12.delta0=delta0.M12,ratio=delta0.M12/delta0.M9)
```

```
s.t.M9=summary(m)$s.table
s.t.M12=summary(m.regional.local.by.month.year)$s.table

a=names(coef(m.regional.local.by.month.year))


# Summaries across models
# Theta estimates
theta.est <- c(m.global$theta,
m.global.plus.Secchi$theta,
m.global.plus.cond$theta,
m.global.plus.tide$theta,
unlist(lapply(list(m.regional,
   m.regional.by.month,
   m.regional.by.year,
   m.regional.by.month.year,
               m.regional.local,
               m.regional.local.by.month,
               m.regional.local.by.year,
         m.regional.local.by.month.year,
               m.regional.local.minus.Secchi,
               m.regional.local.minus.cond,
               m.regional.local.minus.tide),FUN=function(x){
                 return(x$family$getTheta(TRUE))}))
)
theta.est
range(theta.est)
m$family$getTheta(TRUE)

# Model comparison
prop.dev.func <- function(a){
return((a$null.deviance-a$deviance)/a$null.deviance)
}

AIC.set <- AIC(m.global,
   m.global.plus.Secchi,
   m.global.plus.cond,
   m.global.plus.tide,
   m.regional,
   m.regional.by.month,
   m.regional.by.year,
   m.regional.by.month.year,
               m.regional.local,
               m.regional.local.by.month,
               m.regional.local.by.year,
```

```
          m.regional.local.by.month.year,
               m.regional.local.minus.Secchi,
               m.regional.local.minus.cond,
               m.regional.local.minus.tide
               )

index.temp <- which.min(AIC.set$AIC)
delta.AIC.set <- AIC.set$AIC-AIC.set$AIC[index.temp]

ML.score.set <- c(NA,NA,NA,NA,unlist(lapply(list(
   m.regional,
   m.regional.by.month,
   m.regional.by.year,
   m.regional.by.month.year,
               m.regional.local,
               m.regional.local.by.month,
               m.regional.local.by.year,
         m.regional.local.by.month.year,
               m.regional.local.minus.Secchi,
               m.regional.local.minus.cond,
               m.regional.local.minus.tide),FUN=function(x){x$gcv.ubre})))

Percent.dev.exl.set <- unlist(lapply(list(
   m.global,
   m.global.plus.Secchi,
   m.global.plus.cond,
   m.global.plus.tide,
   m.regional,
   m.regional.by.month,
   m.regional.by.year,
   m.regional.by.month.year,
               m.regional.local,
               m.regional.local.by.month,
               m.regional.local.by.year,
         m.regional.local.by.month.year,
               m.regional.local.minus.Secchi,
               m.regional.local.minus.cond,
               m.regional.local.minus.tide),FUN=function(x){prop.dev.func(x)}))

dev.expl.table <- data.frame(
Model=c('Global',
'Global + Secchi',
'Global + cond',
'Global + tide',
'Global + regional (single)',
```

```
'Global + regional (month)',
'Global + regional (year)',
'Global + regional (month year)',
        'Global + regional (single) + local',
        'Global + regional (month) + local',
        'Global + regional (year) + local',
    'Global + regional (year) + local',
        'Global + regional (single) + local - Secchi',
        'Global + regional (single) + local - cond',
        'Global + regional (single) + local - Tide'),
Df=AIC.set$df,
AIC=AIC.set$AIC,
delta.AIC=delta.AIC.set,
SSC=ML.score.set,
'Dev. exp.'=100*Percent.dev.exl.set,
'Theta'=theta.est
)
dev.expl.table <- data.frame(M=1:nrow(AIC.set),Model=dev.expl.table$Model,round(dev.expl
dev.expl.table
write.csv(dev.expl.table,
          file='~/smelt/gam-analyses/SKT-gam-analyses/dev-explained-table-v10.csv',
          row.names=F)
print.xtable(xtable(dev.expl.table,digits=1),include.rownames=F)
```