Paper proposal and preliminary analysis

2024-01-03

Central question

1. How has the functional trait composition and distribution of the Northeast U.S. Shelf changed in the past 50 years?
2. How are these changes related to key drivers, including harvest pressure and ocean temperatures?

Methods

Species List

To generate the species list, I utilized the Northeast Trawl Survey dataset cleaned and prepared by Adam Kemberling (https://github.com/adamkemberling/nefsc_trawl). I then just found all unique species observed in the trawl survey data set (~450) and filtered the list to only fish species (removed all invertebrates, “unknown” classes, and other misc. classes like empty clam shells). This resulted in 334 unique species and 58 unique higher level taxa (e.g. individuals identified to genus, family, class) *(n = 392)*.

Literature review of commonly used traits

Initially, I conducted a non-systematic review of the literature searching Google Scholar for with the terms “fish” AND “functional” AND “trait”. I scanned the abstracts of the first ~100 entries and then extracted the traits used in these analyses for 17 papers that seemed particularly pertinent. Most papers used traits from FishBase (https://fishbase.mnhn.fr/search.php) or derivatives of FishBase. Therefore, there wasn’t considerable diversity in the traits represented (Fig. 1).

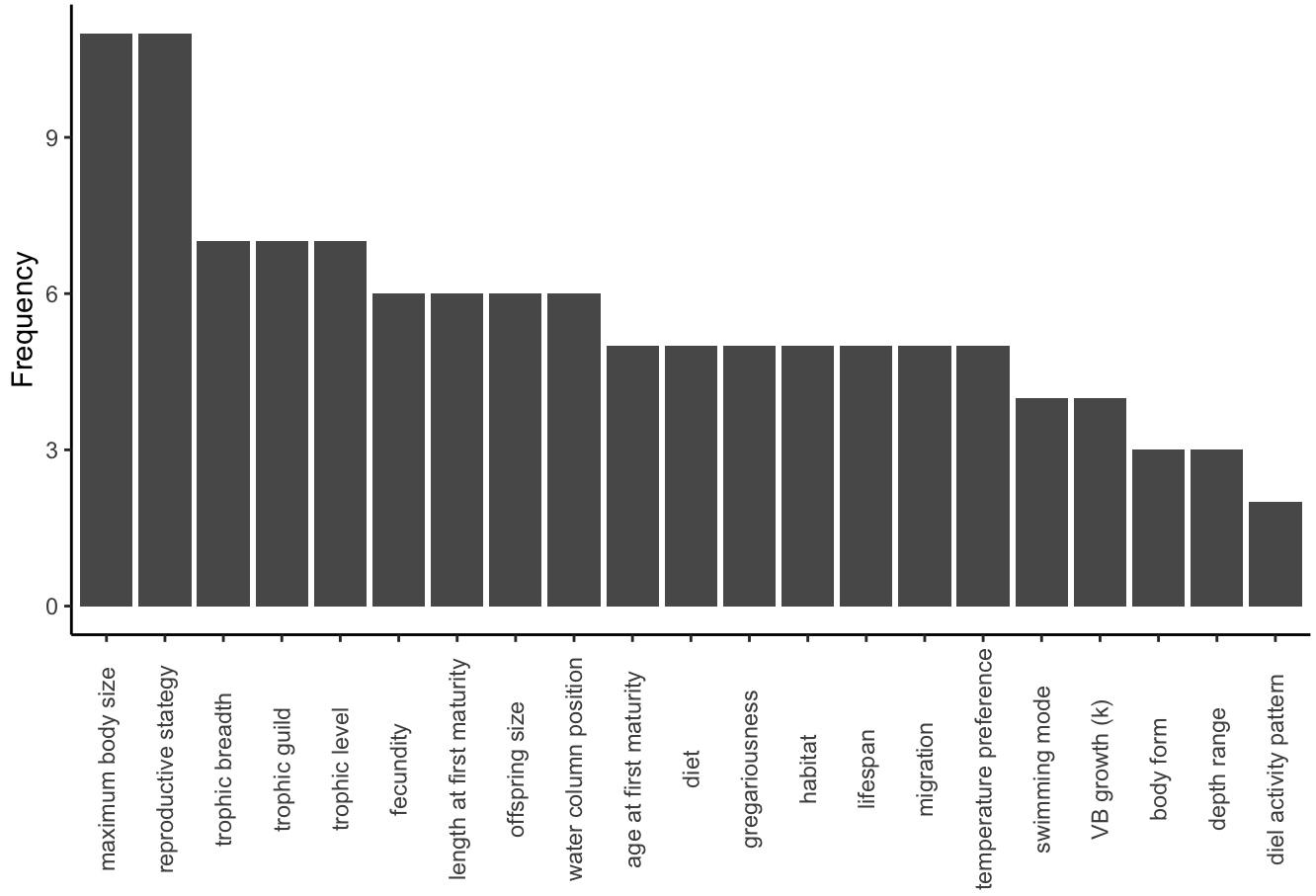

Figure 1. Frequency of different traits used in 17 papers that assessed marine fish function traits.

Developing trait database

To develop the trait database I utilized two primary sources: Beukhof et al. 2019 (https://doi.pangaea.de/10.1594/PANGAEA.900866) and Thorsen et al. (2023) FishLife database (https://github.com/James-Thorson-NOAA/FishLife). Both data sources are based on FishBase (https://fishbase.mnhn.fr/search.php), however provide different levels of detail.

Beukhof et al. (2019) queried FishBase for a species list that encompassed >80% of the species that we observe in the Northeast U.S. Shelf ecosystem. They extracted fish traits from fishbase at different levels of organization and supplemented trait values from FishBase with values from the literature. The narrowest scale of their estimate was at the species level within the large marine ecosystem (LME) designations. When trait values were not available at this level they decreased the geographic and taxonomic specificity until they reached a value for the trait. Thus most of their trait values were estimated at taxonomic scales coarser than the species level (e.g. the fecundity of species X, was estimated the average fecundity of specie in the family to which species X belongs).

Thorson et al. (2023) used a modeling approach to *impute* trait values for all fish species in FishBase based on taxonomic relatedness. Thus, if a trait for a particular species was unknown, the model generated a prediction for that trait based on taxonomic relatedness and a structural equation model.

Initially, I attempted to generate my own trait database from FishBase using the rfishbase package (https://github.com/ropensci/rfishbase). However, after building the database and comparing to Beukhof et al. (2019), I decided that Beukhof et al. (2019) had filled many of the gaps with literature based values. Therefore, I took as the default Beukhof et al. (2019), and then replaced all trait values that were estimated from taxonomic

levels > species (e.g. genus, family, class, order) and replaced those values with imputed estimates from FishLife. Finally, the NEFSC trawl survey often does not identify individuals to species. For all designations > species in the data set, I extracted estimates for the trait values from Thorson et al. (2023) at the higher levels of organization.

Thorson et al. (2023) and Beukhof et al. (2019) had considerable overlap in the species traits they collected. However, each also collected traits that were not represented in the other. The final list of traits that I collected and am interested in analyzing is summarized in Table 1.

Table 1. Traits included in the final data base.

Traits	**Datastructure**
Feeding mode	categorical
Trophic level	continuous
Offspring size	continuous
Spawning type	categorical
Age at maturity	continuous
Fecundity	continuous
Length inf.	continuous
Growth coefficient	continuous
Maximum observed length	continuous
Maximum age	continuous
Weight inf.	continuous
Natural mortality	continuous
Length at maturity	continuous

Community weighted means

There is an enormous body of literature describing how to relate species abundances, traits, and environmental covariates. Techniques range from simple regression based approaches (Grime et al. 1998), to multivarate methods (Kleyer et al. 2012), to complex spatiotemporal models that estimate the impact of traits on species communities in a hierarchical framework (HMSC, Ovaskainen et al. 2017). As an initial approach, I choose to model fish communities using community weighted mean (CWM) trait values (e.g. Grime et al. 1998, Lavorel et al. 2008, Frainer et al. 2017). This approach has been criticized (Peres-Neto et al. 2017), due to its inability to resolve the “4th corner problem” (e.g. the trait X environment matrix). More importantly, CWM can only provide insight into how community level traits are changing, not how changes in specific species abundances/biomass are driving that change. Despite these deficiencies, the CWM approach allows for a straight forward method to condense the trait and species matrices into single values for each observation, which can then be used as responses in regression based models (e.g. Frainer et al. 2017).

Following Lavorel et al. (2008), I estimated the community weighted mean (CWM) for each trait j as

$$\mathrm{CWM}_j = \frac{\sum_{i=1}^{n} b_i t_{i,j}}{\sum_{i=1}^{n} b_i}$$

where b_i is the biomass of species i, and $t_{i,j}$ is the value of trait j for species i. I estimated the CWM for each trait at each sampling location in each year. For the initial analysis, I only estimated the CWM for continuous traits (see Table 1). To estimate the CWM_j for each trait at each unique tow k I used matrix multiplication, such that

$$\mathbf{CWM} = \frac{\mathbf{B} \cdot \mathbf{T}}{\mathbf{W}}$$

where $\mathbf{CWM}$ represents a $k \times j$ matrix, $\mathbf{B}$ is a $k \times i$ matrix, $\mathbf{T}$ is a $i \times j$ matrix, and $\mathbf{W}$ is a $k \times 1$ vector containing the total biomass of all species in each unique tow, k. I confirmed that the estimates for $\mathbf{CWM}$ matched those calculated from the "functcomp" function in the FD package (Laliberté et al. 2014).

Spatio-temporal modeling of fish functional traits

The goal of our project is to determine a) if traits have changed and b) what drivers are correlated with any changes. There are many statistical approaches that could be used to address these questions. For instance, an initial approach might be to model the community weighted trait distributions for each ecological production unit (EPU) (or the related large marine ecosystems (LMEs)) through time. However, this approach required aggregating the raw data across space (e.g. the average value of trait, j, in a particular year in the Gulf of Maine), and losing finer scale variation in potential driver variables (temperature, harvest) or covariates (bathymetry, rugosity(?)).

Recent advances in spatio-temporal modeling provide an opportunity to directly model traits at the spatial and temporal scales at which they are sampled. While these approaches are traditionally used to model species distributions, I propose using these methods to model spatio-temporal variation in community weighted trait distributions.

Current software for fitting spatiotemporal models to large data sets include sdmTMB (https://pbs-assess.github.io/sdmTMB/index.html#citation) (Anderson et al. 2022) or VAST (https://github.com/James-Thorson-NOAA/VAST) (Thorson et al. 2019), or building models directly in software/packages such as INLA (https://www.r-inla.org/home), STAN (https://mc-stan.org/), or TMB (https://cran.r-project.org/web/packages/TMB/index.html). Of these approaches sdmTMB offers an user-friendly approach that is accessible to those (like me) who are familiar with the syntax of simpler generalized linear mixed effect model (GLMM) implementations (lme4, glmmTMB).

sdmTMB fits GLMMs to spatially and temporally explicit data and can be used to directly assess the effect of driver variables (fixed effects) on a response variable while explicitly accounting for spatial and temporal intercept random fields or random intercept effects. Spatiotemporal models can be modeling as IID, random walks, or first order autoregressive (AR(1)) processes.

As an initial first pass at modeling the CWM data, I implemented simple intercept-only sdmTMB models for each trait, and fit predictions from these models to a buffered spatial grid extending over the Northeast U.S. shelf. The code snippet below outlines the general syntax for an sdmTMB model.

```
m1 <- sdmTMB(
  data = df, # CWM data set
  formula = trophic_level ~ 1, # intercept only model with NO fixed effects here is where we would include spatially and temporally explicit driver variables (temp, harvest), and covariates (depth, temp at time of sampling) or random effects e.g. (1 | survey_area).
  # example formula = y ~ average SST + harvest + ... + (1|survey_area) -- We will need to consider the spatial scale of the harvest data and how to account for the fact that landings is aggregated at broader spatial scales
  mesh = mesh,
  family = gaussian(link = "identity"), # this can be adjusted based on the trait
  spatial = "on",
  time = "est_year",
  spatiotemporal = "IID" # other options include "ar1" or "rw"... "ar1" significantly slows down fitting
)
```

Preliminary results

Data visualization and potential “simple” models

Utilizing the CWM matrix I was able to generate time series for each of the traits in each region (Fig. 2). You can see that at least the mean (averaged across spatial domains) trend does vary through time. Some of this temporal variation seems to be linear/directional, while other traits appear to fluctuate since 1970.

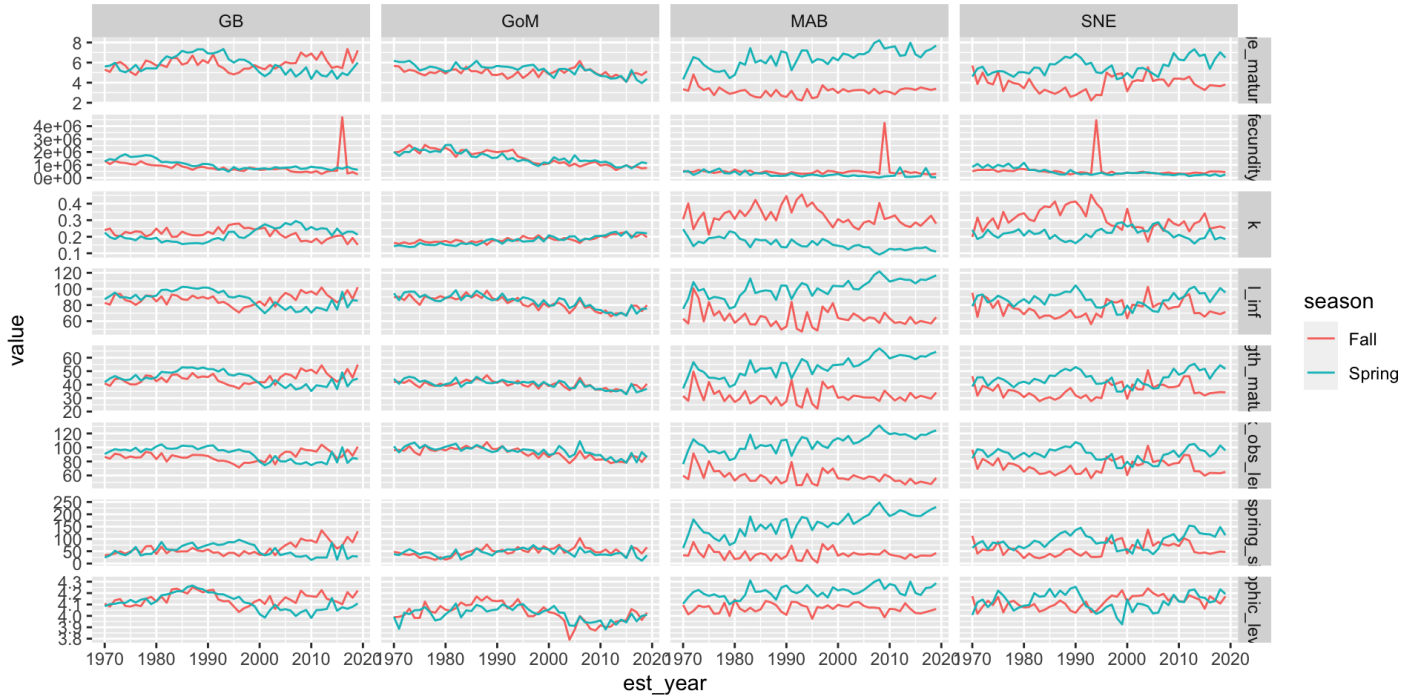

Figure 2. Time series of the community weighted mean trait value for each trait average across space in each ecosystem in each season.

One approach would be to assume that observations (tows) are independent and identically distributed (IID) within each EPU, and model the “raw” (e.g. at the level of tows) CWM trait values as a function of driver variables and spatial designations and seasons. For example,

```
nlme::gls(trophic_level~ est_year*survey_area*season, data = df,
                  correlation = nlme::corAR1(form = ~ est_year | ts_id)) # ts_id is a gr
ouping variable for each unique ts.
```

is a simple regression model with an ar(1) correlation structure that would test the hypothesis that the average trophic level has changed with time in each season and survey area. As as example, Fig. 3 depicts the predicted linear change in trophic level as a function of time in each season and area. All parameters are significant at $\alpha < 0.05$, which isn't very surprising considering the sample size. However, it is unclear if these shifts are biologically relevant. For instance, Georges Bank in the spring survey shows a decline in CWM trophic level by ~0.2 units over the 50 year time series. Is this a relevant change?

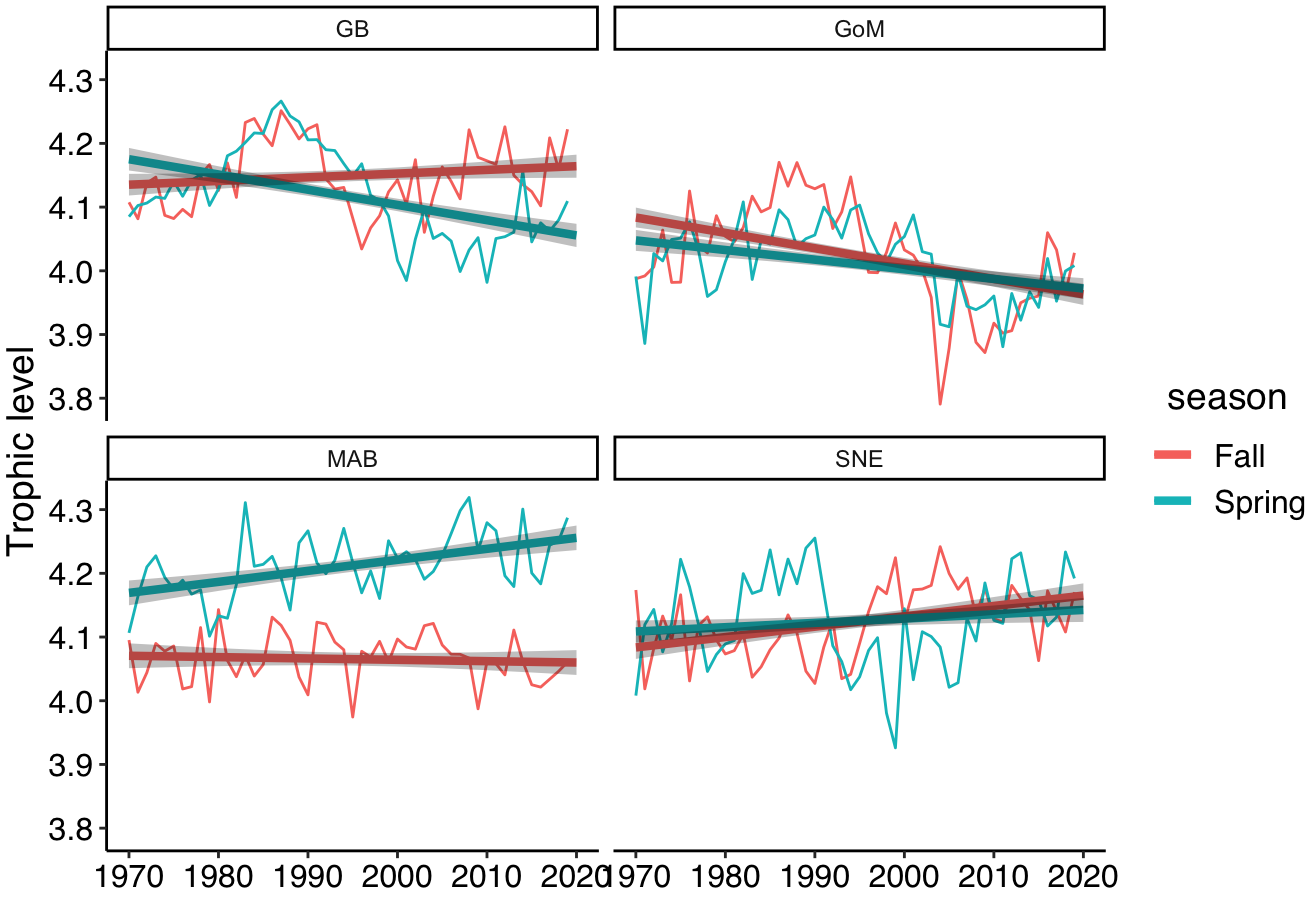

Figure 3. Time series of average CWM trophic level with associated predictions from a simple regression model with ar(1) correlated error structure.

As an example of a trait that has show more significant change, Fig. 4 shows changes in CWM length at maturity. We can see strong increases in length at maturity in the southern regions (SNE and MAB) particularly in the spring surveys. Interestingly, length of maturity has appeared to decline in the Gulf of Maine (GOM) and in only the spring survey on Georges (GB) despite increases in CWM length at maturity in the fall survey on George's.

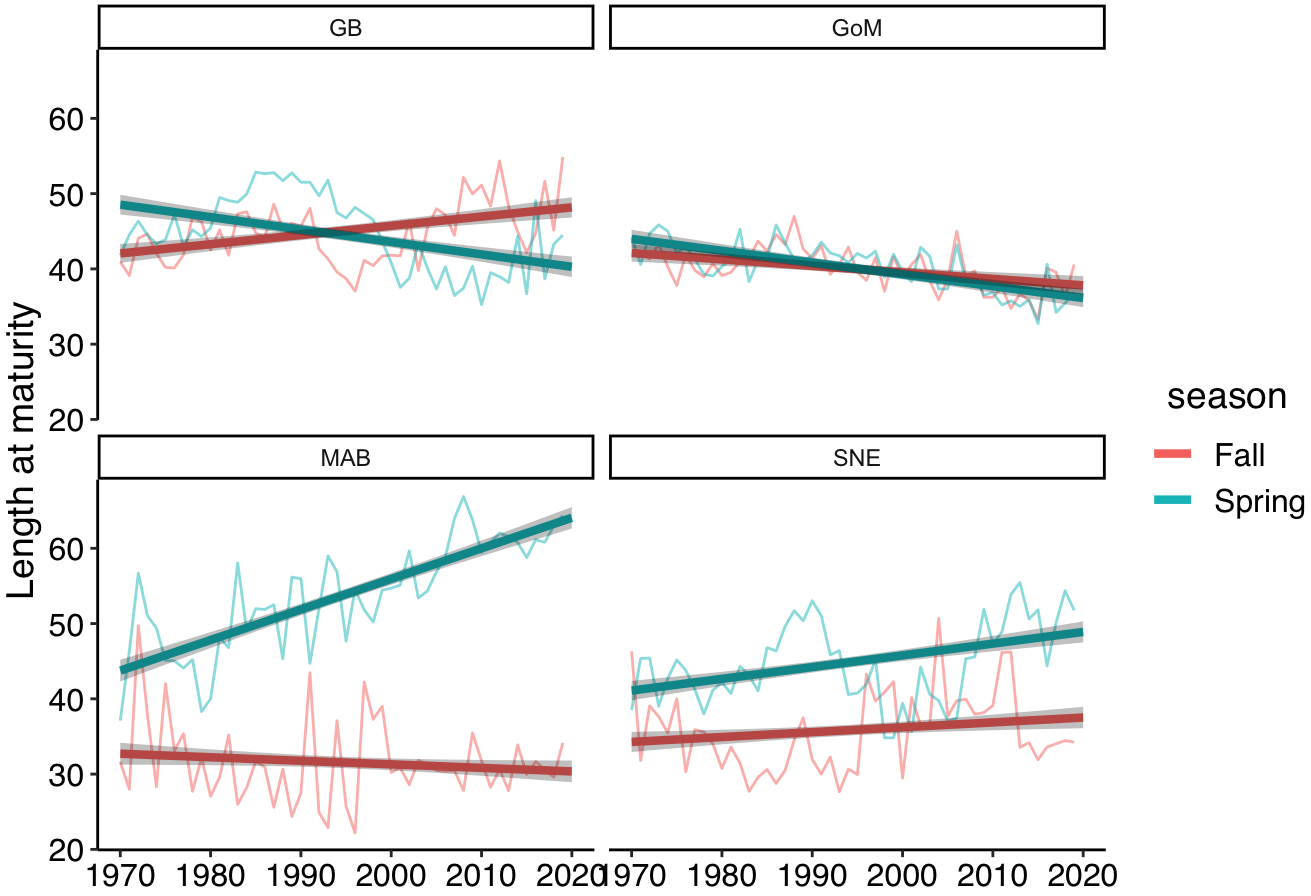

Figure 4. Time series of average CWM length at maturity with associated predictions from a linear regression model with ar(1) correlated errors.

Here, I've just included some simple models to spur conversation. However, this approach could easily be extended to include GAM's (considering the fluctuation in mean CWM values over time) or implemented in Bayesian approaches. Furthermore, we could decide to explicitly model it was a time series model (e.g. $\mathrm{CWM}_{y} \sim \mathrm{CWM}_{y-1} + \ldots$), where driver variables influenced how the CWM values change between time steps.

Spatiotemporal model results

To relax the assumption that there was not spatial structure to the data within each region, I also fit preliminary geospatial GLMMs (e.g. sdmTMB) to the CWM values for each trait. To date (1/04/24), I have not constructed the data set of driver variables. Therefore, each of these models represent an intercept only model with either no fixed effects or a single fixed effect of season (see methods), no random intercept effects, and spatiotemporal intercept random fields.

As an example, Fig. 5 shows some of the predicted output from one of these models fit to the length at maturity CWM data.

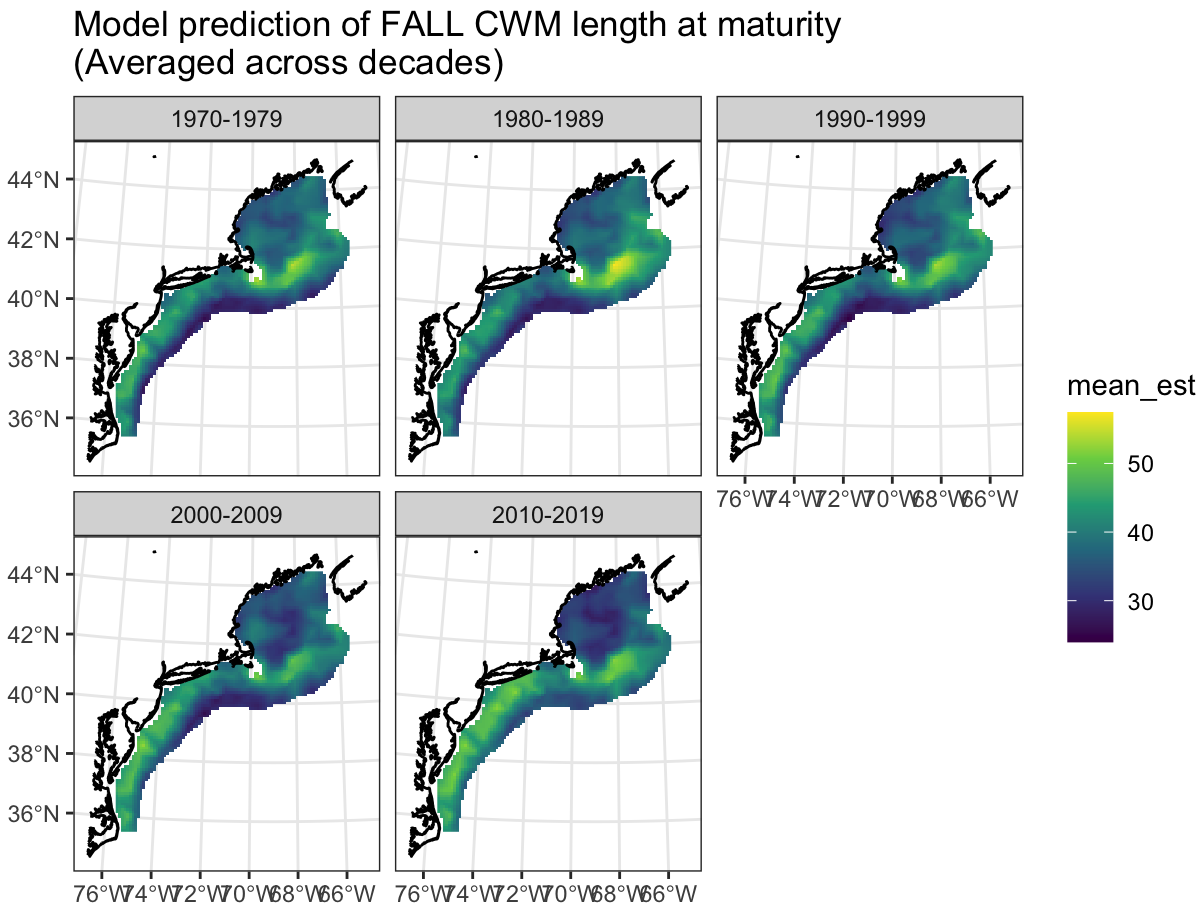

Figure 5. Predicted CWM length at maturity across the Northeast U.S. shelf for the fall season. Predictions are averaged across decades for visualization purposes. However, the model generates predictions for the entire grid in each season in each year.

What is interesting about this approach is that we can examine the model output from many different perspectives. We could generate predictions across different spatial domains. For example, the EPUs, and then extract time series of average trends across the area. Alternatively we could extract the center of gravity for traits to see how it was changed through time. Fig. 6 is an example of this for length at maturity. Please note that while this figure is compelling, the scale is tiny. These changes to the north east only reflect a shift of <100 km in 50 years.

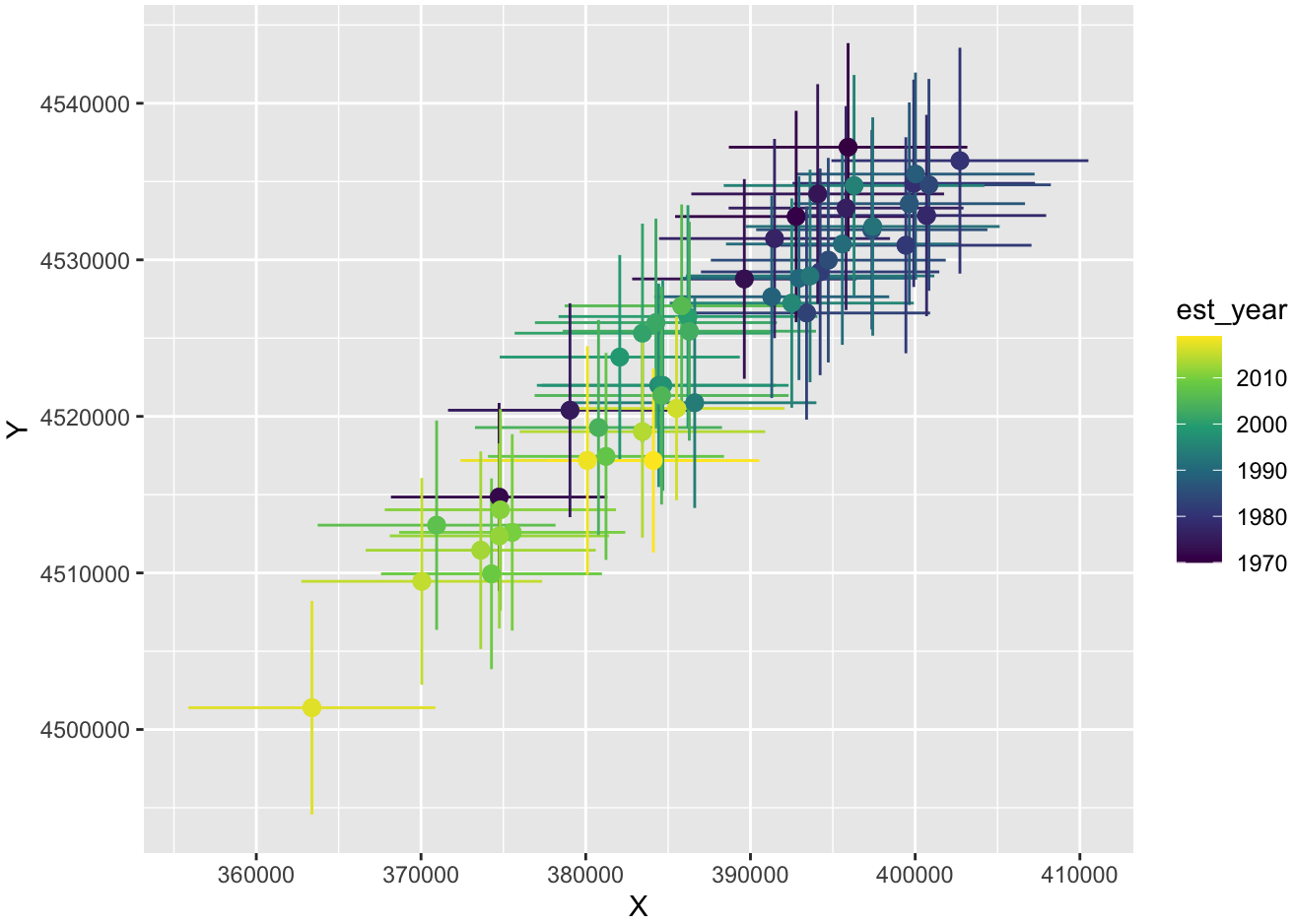

Figure 6. Center of gravity estimates in each year for CWM length at maturity.

Data availability

If you are interested in following along with any of these analyses you can find the code at (https://github.com/bartdifiore/Fish-Functional-Traits (https://github.com/bartdifiore/Fish-Functional-Traits)). The repo utilizes git large file storage, so to run the code locally you need to install git lfs (https://git-lfs.com/) and initialize git for the repository once you clone.