Bluefish VAST working paper

Abigail Tyrell
December 2022

Introduction

This working paper documents how Vector Autoregressive Spatiotemporal (VAST) models (Thorson 2019) were explored to standardize bluefish survey indices in the 2022 bluefish research track stock assessment.

Data

Bluefish count and biomass data was acquired from the NMFS bottom trawl. Depth data was taken from the NMFS bottom trawl. Gridded mean September temperature for the Northeast U.S. shelf was calculated from the OISST dataset available from the Physical Sciences Laboratory.

VAST model parameterization

All VAST models had the same general structure. Spatial and spatiotemporal effects were modeled with random effects. The number of factors used was determined from the number of size groups in the input data (either one or three). Temporal and spatiotemporal autocorrelation were not included, because temporal autocorrelation is not desirable for index calculation (Thorson 2019). Years with zero catch of any of the size groups were dropped from the analyses because the model could not accommodate these years without including temporal autocorrelation.

Comparing VAST index of abundance with the canonical stratified swept area

Methods

VAST

To compare the VAST biomass index to the mean swept area biomass index, which has historically been used in the assessment, a “stratified-swept-area-like” VAST model was constructed.

A univariate VAST model was fit using depth as a density covariate. Depth was included as a density covariate to approximate the strata information that was used to calculate the stratified mean swept biomass. Bluefish biomass was modeled with a Poisson-link delta model.

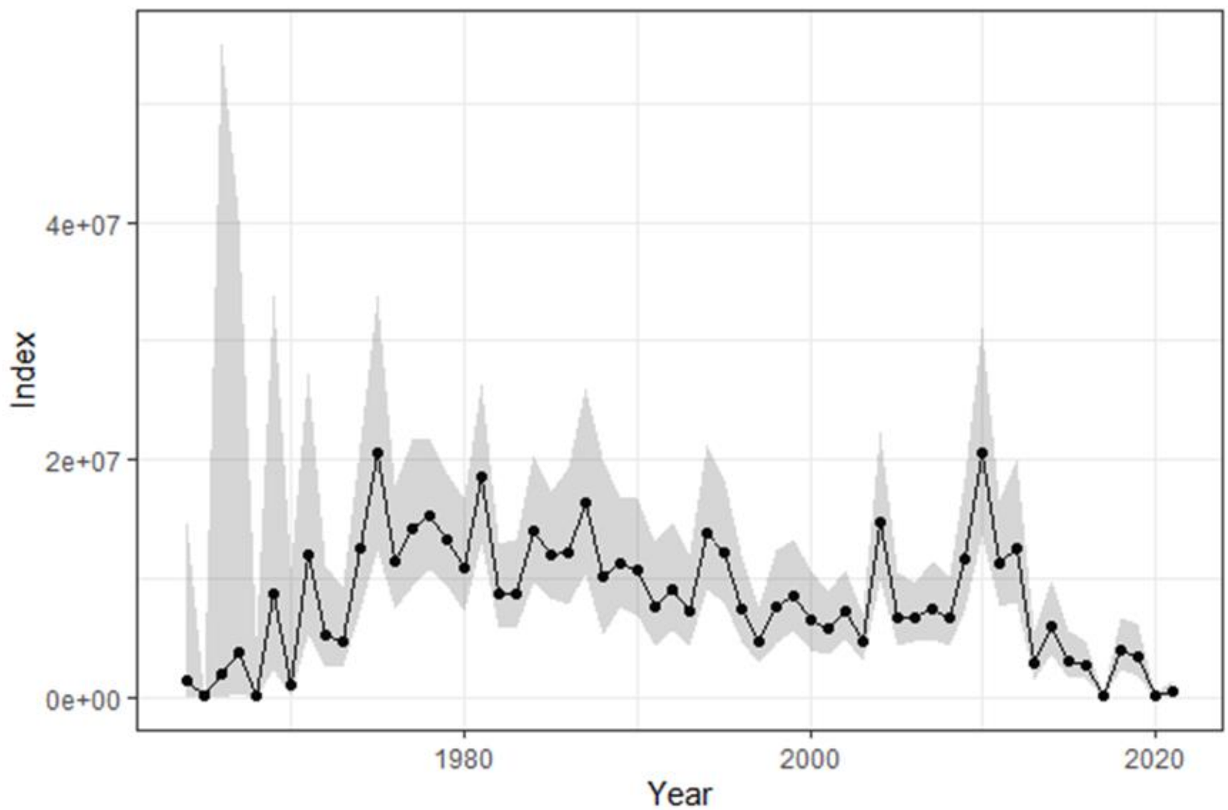
VAST R code is available at the end of this document.

Swept mean biomass

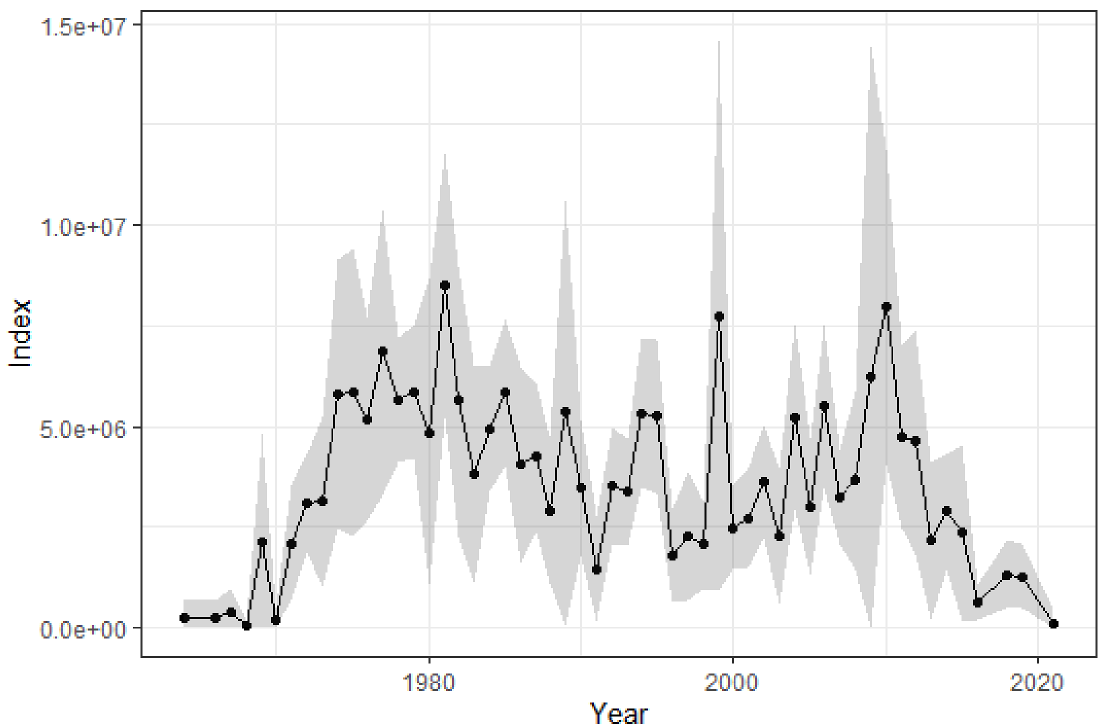
Swept mean biomass was calculated using the {survdat} R package.

Results

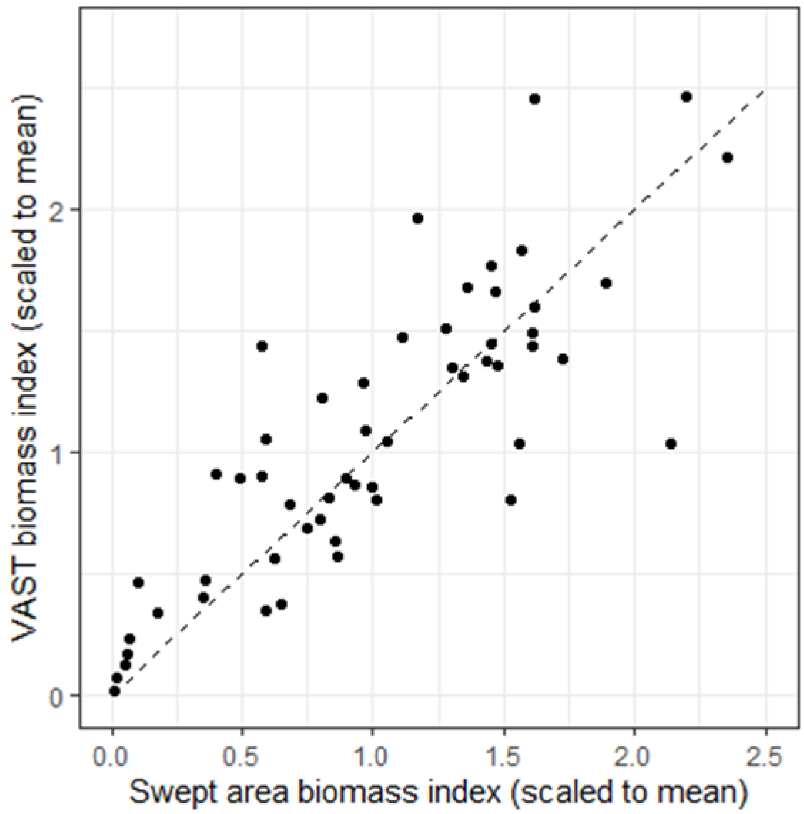
VAST



Swept area biomass index



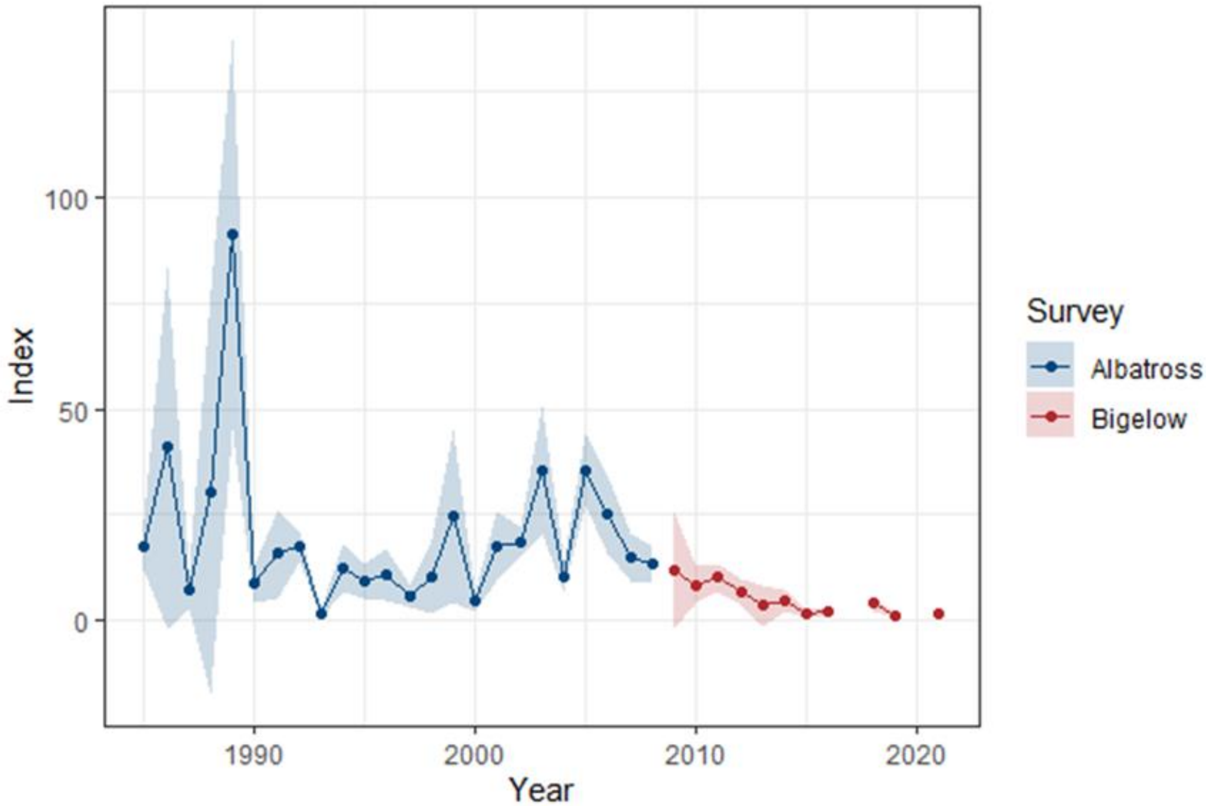
Comparison between VAST and swept area biomass



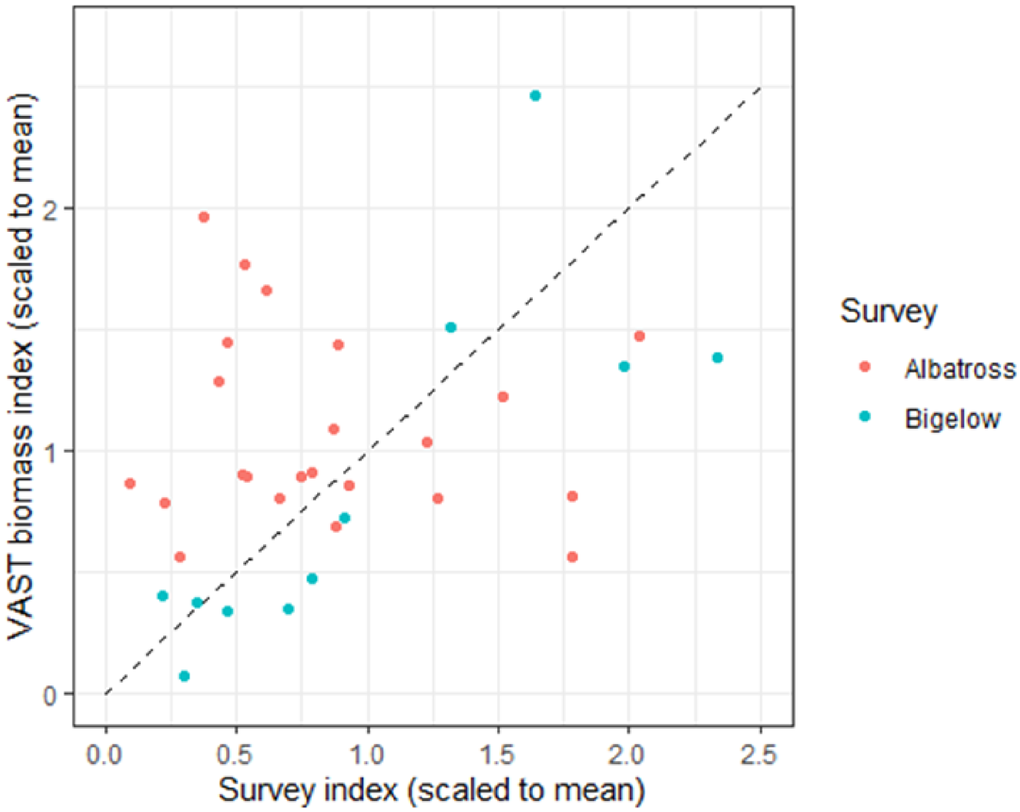
Correlation coefficient:

```
## [1] 0.8307785
```

Survey indices used in assessment model



Comparison between VAST and survey indices used in the assessment model



Correlation coefficient:

```
## [1] 0.3135219
```

The correlation between the survey indices used in the assessment model and the VAST index is lower than the correlation between the swept area biomass and the VAST index. This is to be expected for two reasons; first, the VAST model and swept area biomass calculation both use the entire geographic range of the NMFS bottom trawl survey, while the indices used in the assessment model were based only on strata that were determined to be suitable for bluefish; even though the additional strata did not contain many bluefish, expanding even a small bluefish biomass over the increased area results in changes to the biomass estimates. Second, the survey indices used in the assessment model span a shorter time period because the assessment model begins in 1985, which reduces the amount of data informing the correlation.

Multivariate length groups

Because bluefish exhibit size-based habitat usage, we fit a multivariate VAST model with three length categories (small bluefish <=30.3cm, medium bluefish 30.3-50.0cm, and large bluefish >=50.0cm). The temperature dataset only goes back to 1982, so all models were restricted to 1982-2021. Years were dropped from the analysis if there were fewer than 300 total fall tows or if one or more groups had zero bluefish observations in that year. These criteria resulted in 1988, 2017, 2020, and 2021 being dropped.

Modeling

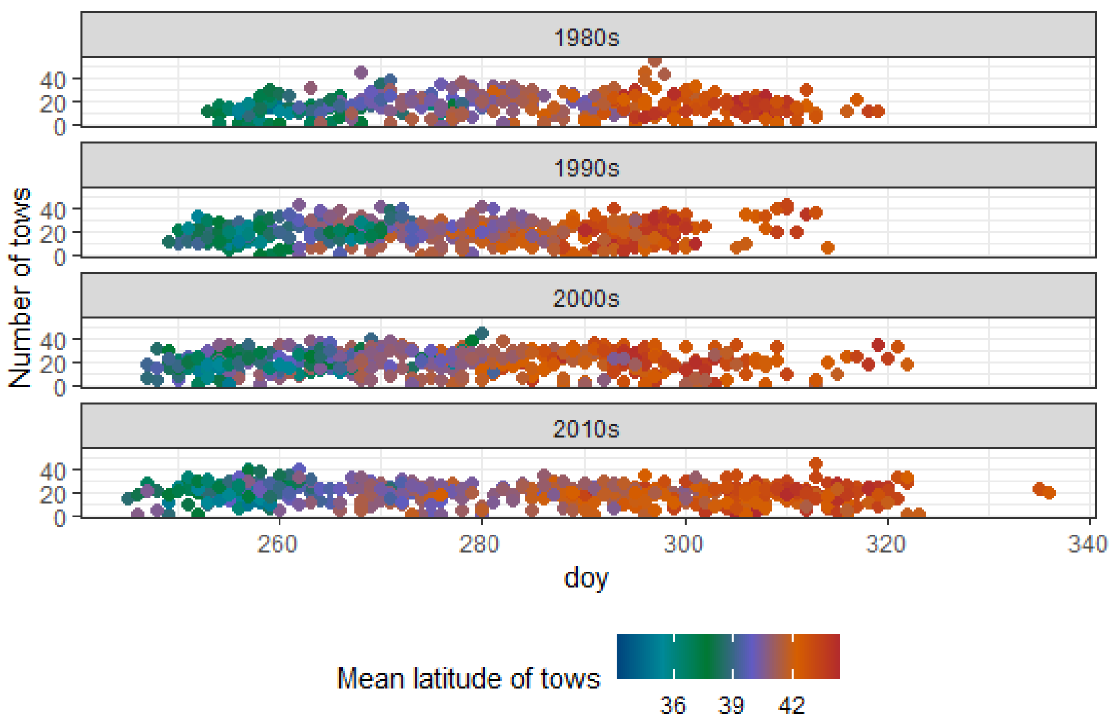
All VAST code is available at the end of this document. First, a base model with no covariates was fit. Depth, day of year, and temperature were added stepwise as covariates. In all models, bluefish counts were modeled with a Poisson-link delta model.

Depth

Depth was added as a density covariate because depth affects the distribution of most fish. Depth was log-transformed and normalized. To further allow for a non-linear relationship with depth, the effect of depth was modeled with a b-spline with three knots.

Day of year

Day of year was added as a catchability covariate to account for the day of year of sampling of each location varying over time. Day of year was modeled as a linear effect.



Temperature

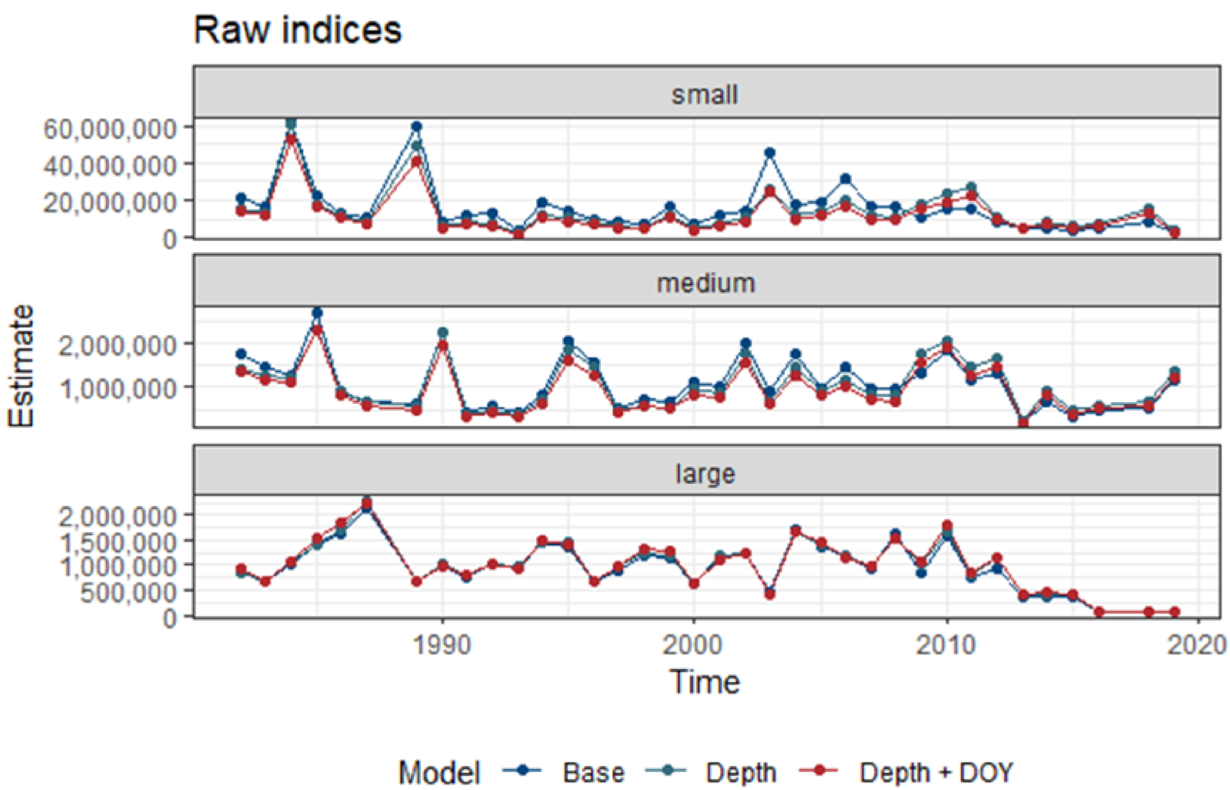
Normalized mean gridded sea surface temperature in September calculated from the daily OISST dataset from the Physical Sciences Laboratory and was used as a density covariate.

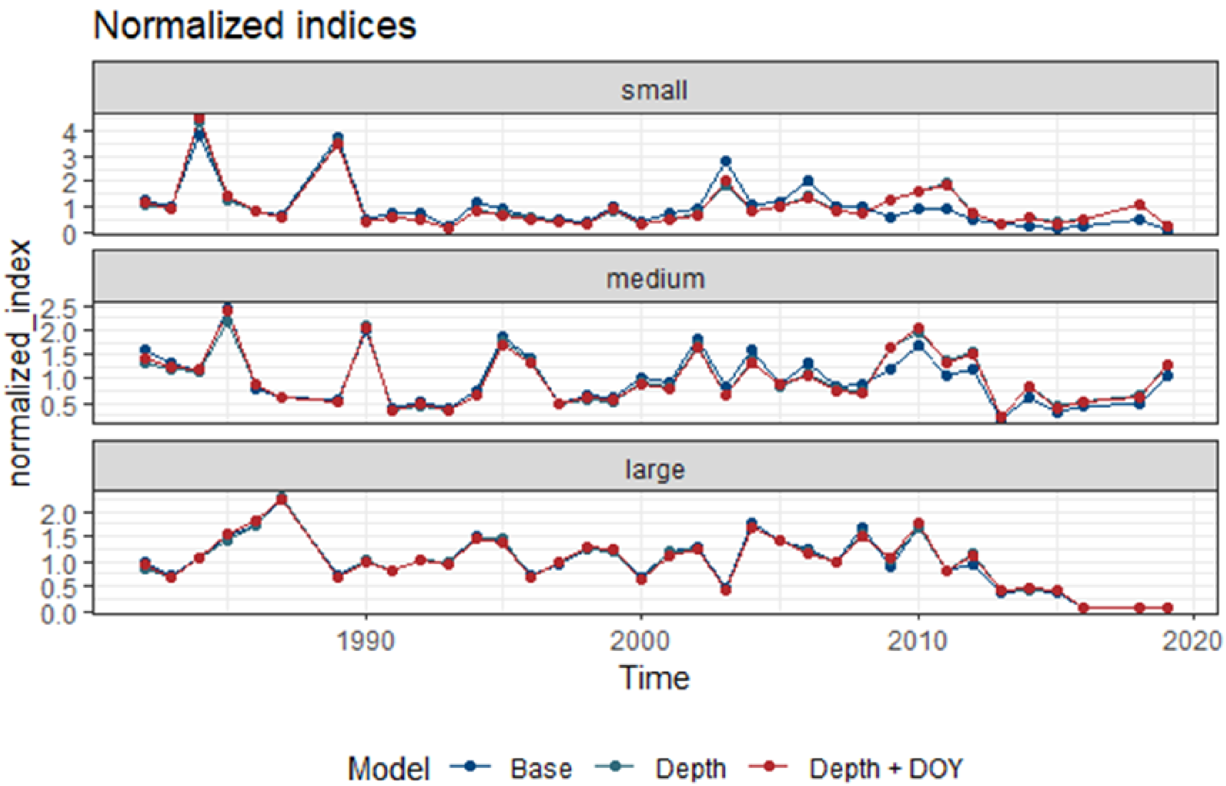
To further allow for a non-linear relationship with temperature, the effect of temperature was modeled with a b-spline with three knots.

Results

The base model and the models that included depth and depth and day of year all converged with good diagnostics. The model that included depth, day of year, and temperature as covariates converged with a high maximum gradient (0.02) and the Hessian was not positive definite. Therefore, this model was dropped from further consideration.

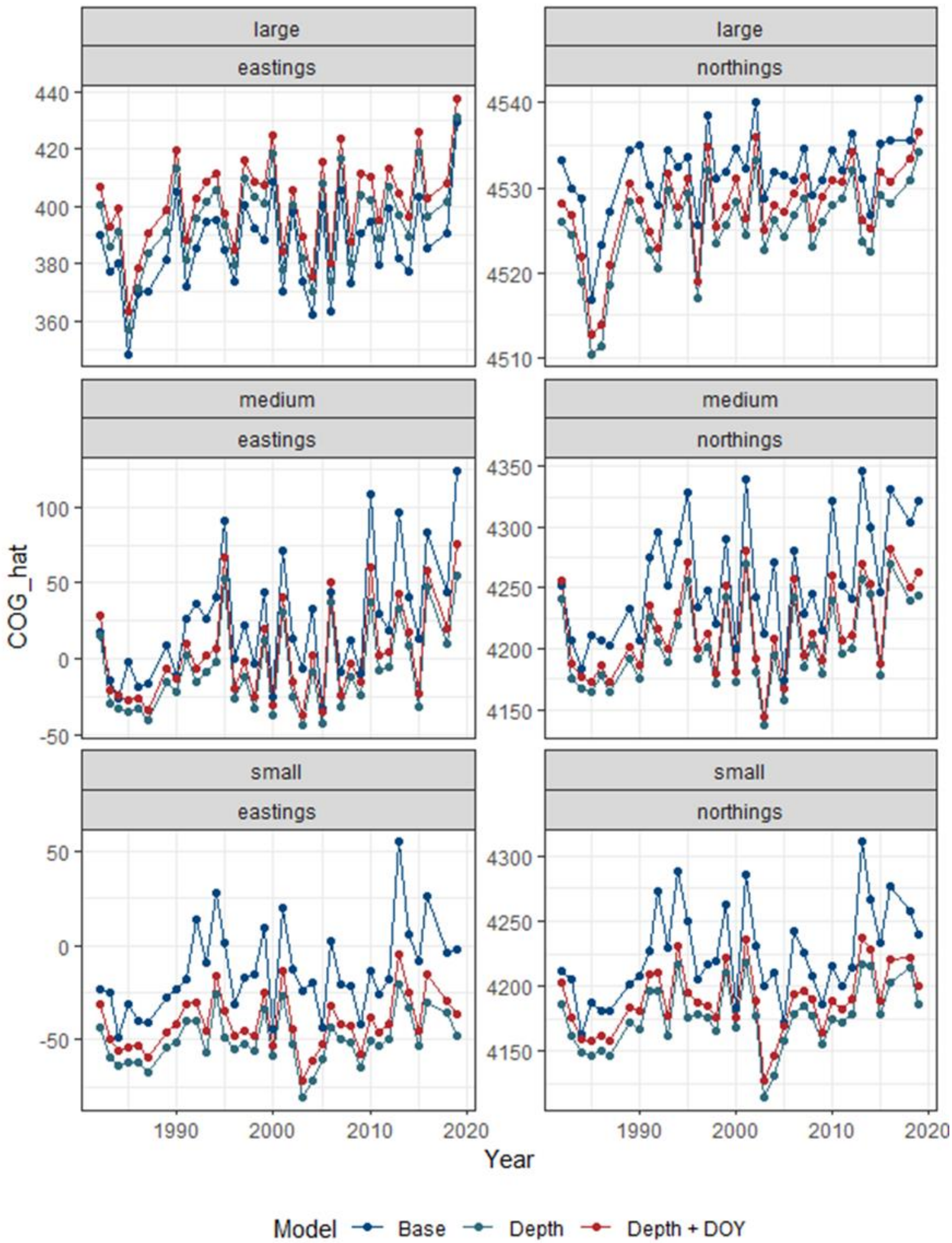
Indices





Normalized indices
small
medium
large
normalized_index
0
1
2
3
4
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
1990
2000
2010
2020
Time
Model
Base
Depth
Depth + DOY

Centers of gravity



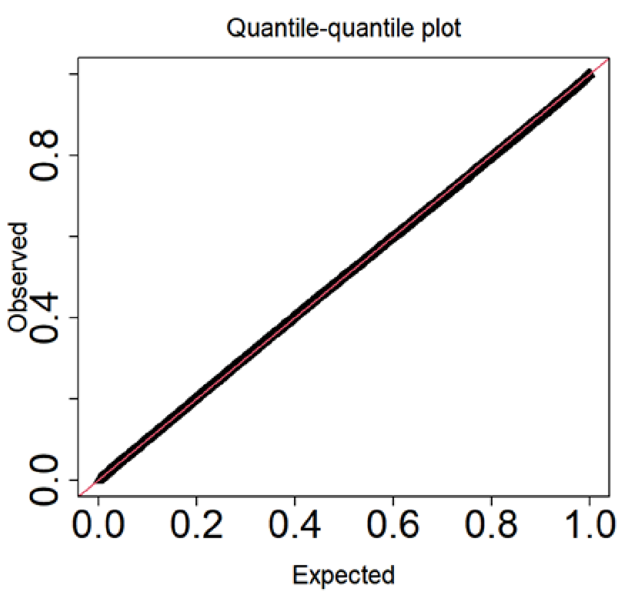
AIC and deviance

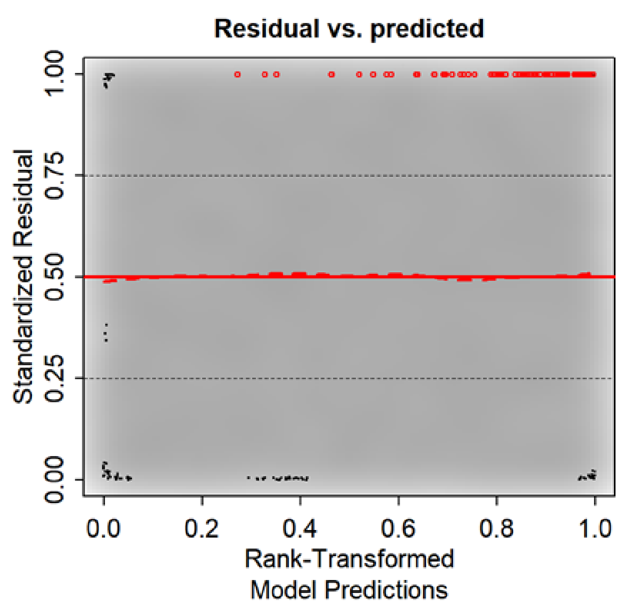
Model	AIC	Deviance
Base	29,883.52	15,348.43
Depth	28,513.39	13,872.31
Depth + DOY	28,514.19	13,881.60

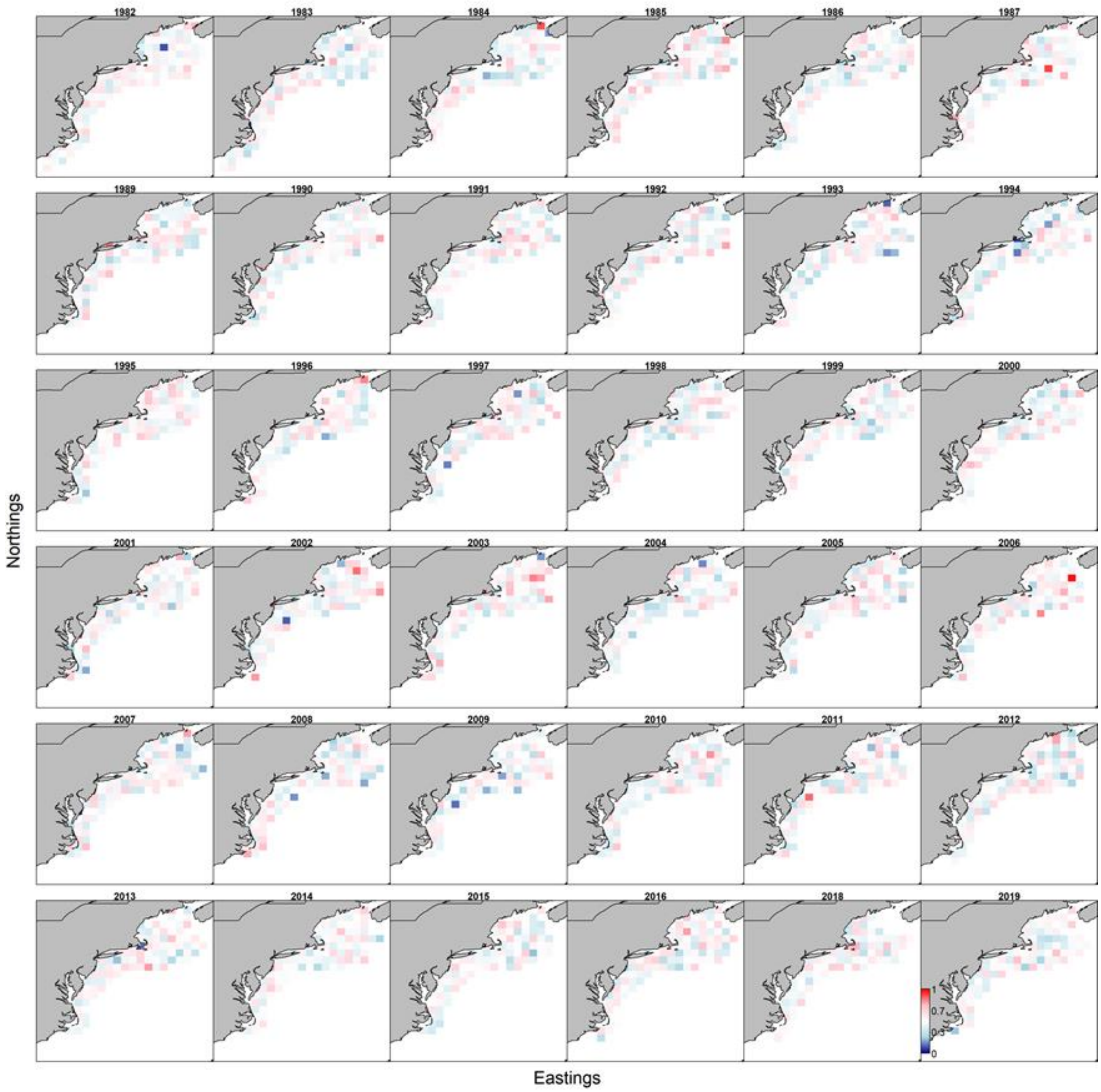
Preferred model

All models produced similar indices and centers of gravity for all three size groups of bluefish. Most notably, the inclusion of depth and day of year as covariates shifted the estimate of center of gravity of small bluefish to the west and to the south compared to the base model. Both the depth-only model and the model including depth and day of year performed better than the base model, according to AIC and deviance. These two models with covariates had nearly equivalent deviance and accounted for 9.6% of deviance in the base model. The model including depth and day of year was selected as the preferred model. Although this model had an AIC 0.8 units larger than the depth-only model, the day of year covariate was expected to account for within-season movements of bluefish and therefore was retained in order to produce more realistic outputs. Further diagnostic information and results are presented here for the depth and day of year model.

Model diagnostics



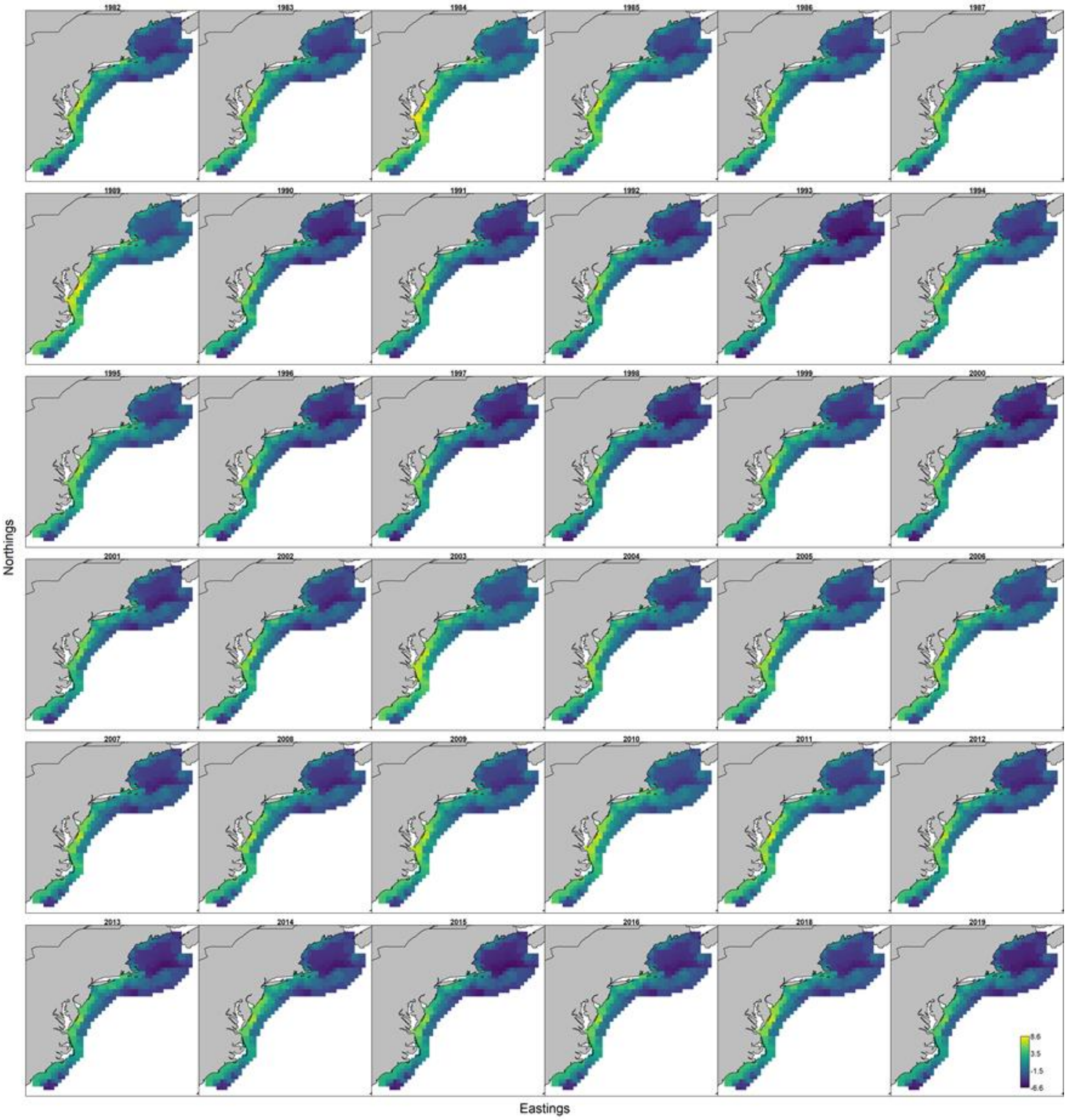




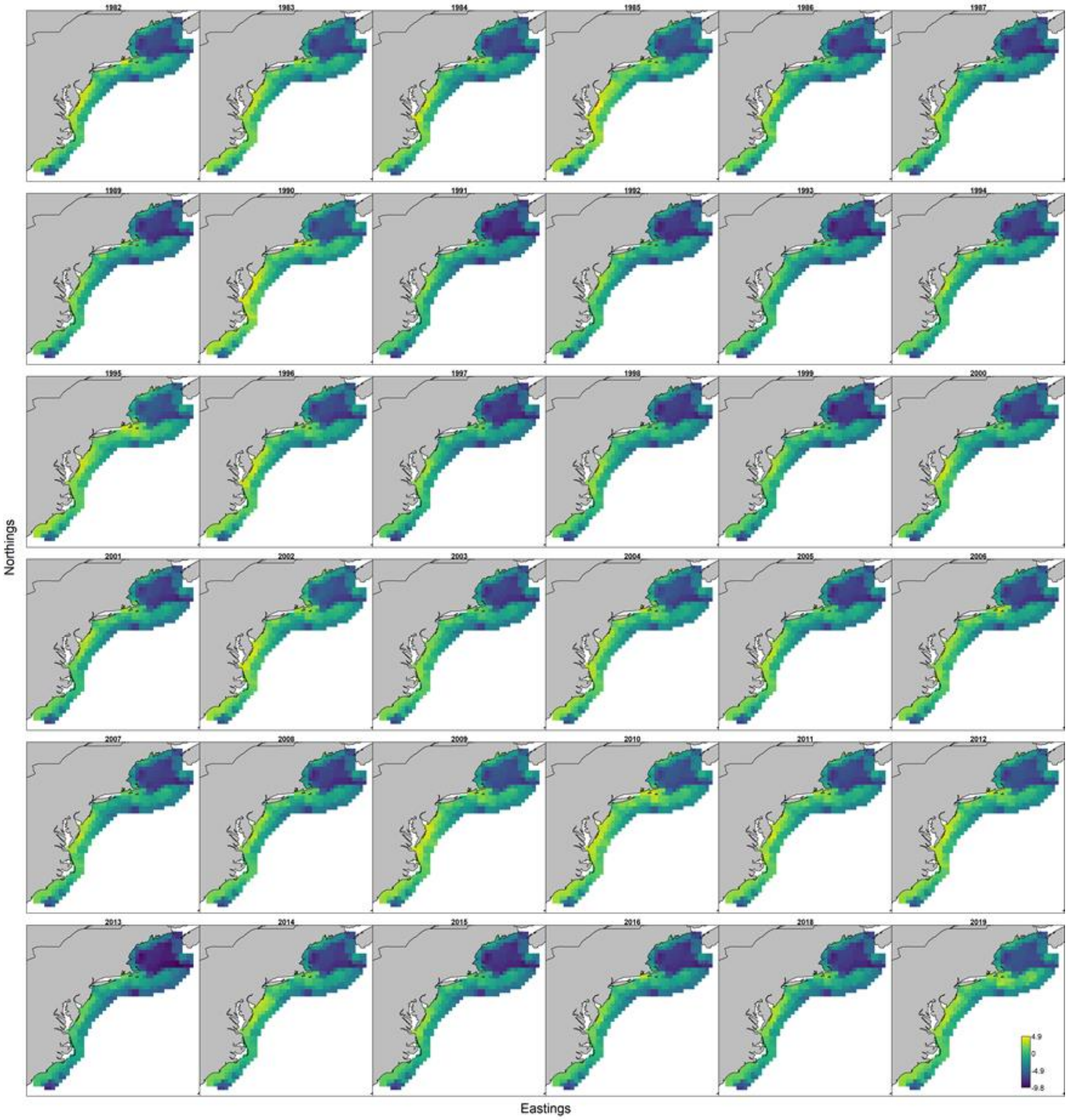
Model results

Note that the scales differ on the following three figures.

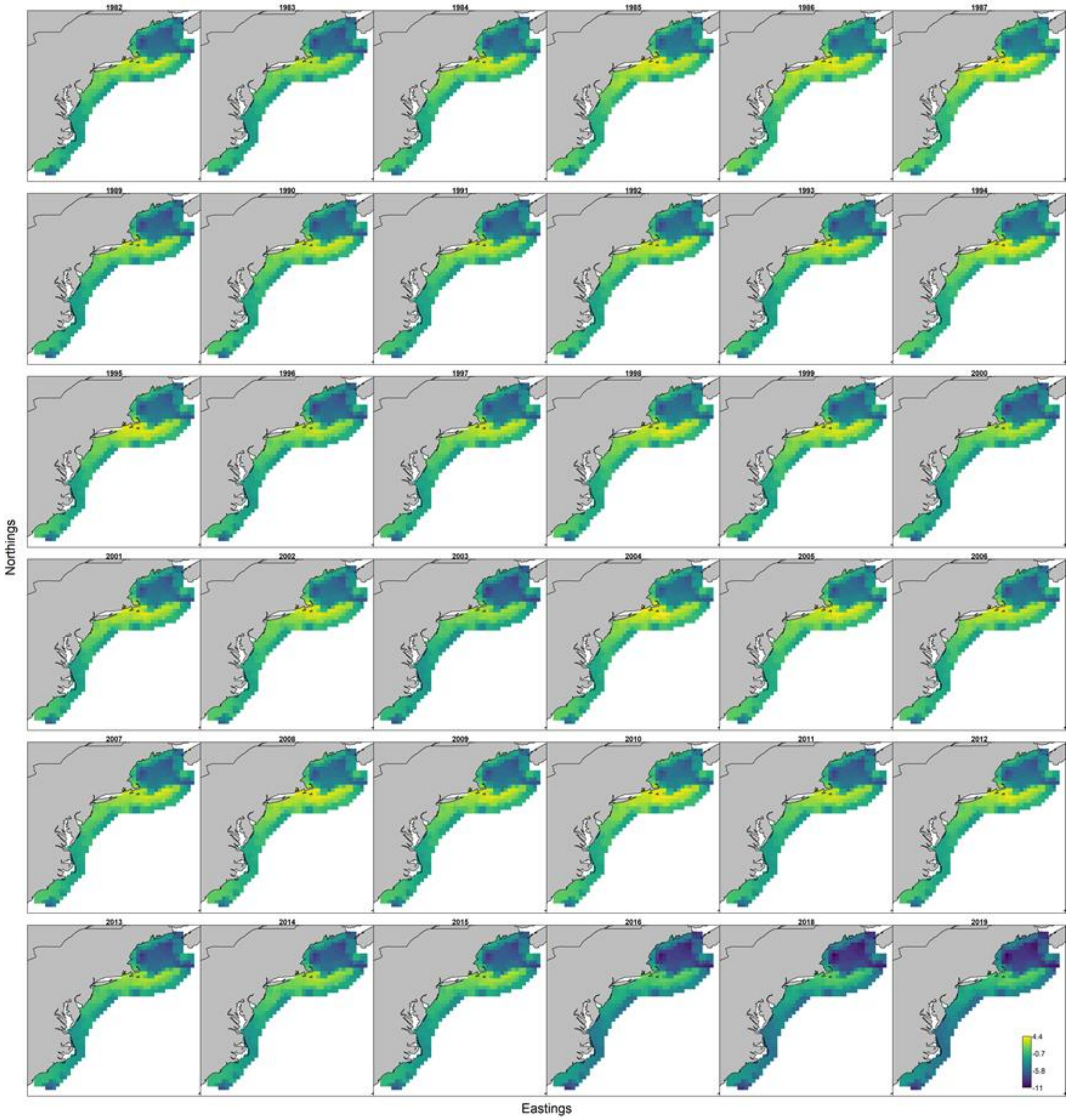
Small bluefish



Medium bluefish



Large bluefish



Appendix

Swept area R code

```r
# connect to database
channel <- dbutils::connect_to_database(
  server = "sole",
  uid = "atyrell"
```

```
)

# pull all trawl data
pull <- survdat::get_survdat_data(channel,
                                  getBio = FALSE)
saveRDS(pull, here::here("data-raw/share/2022-08-31_nmfs_pull.RDS"))

# calculate swept area
swept <- survdat::calc_swept_area(
  surveyData = pull$survdat,
  areaPolygon = bluefish_shape,
  areaDescription = "STRATA",
  filterByArea = "all",
  filterBySeason = "FALL",
  groupDescription = "SVSPP",
  filterByGroup = "all",
  mergesexFlag = TRUE,
  tidy = FALSE,
  # q = NULL, # default catchability is 1, can change
  q = data.frame(groupDescription = 135,
                 q = 0.013),
  a = 0.0384 # average swept area of trawl
) %>%
  dplyr::filter(SVSPP == 135) # filter to bluefish only

write.csv(swept, here::here("data-
raw/share/bluefish_nmfs_fall_swept_area.csv"))
```

VAST R code

```
`%>%` <- magrittr::`%>%`

library(VAST)
library(dplyr)
library(tidyr)

load(here::here("analysis/data/derived_data/bluefish_nmfs.rda"))
load(here::here("analysis/data/derived_data/all_nmfs.rda"))

depth <- all_nmfs %>%
  dplyr::select(Lat = LAT, Lon = LON, Depth = DEPTH) %>%
  dplyr::distinct() %>%
  tidyr::drop_na() %>%
  dplyr::full_join(tibble::tibble(Year = 1963:2021),
                   by = as.character())

covariates <- read.csv(here::here("density_covariates.csv"))

# univariate model ----
```

```
bfish <- bluefish_nmfs %>%
  # add zeros
  dplyr::full_join(all_nmfs %>%
                   dplyr::mutate(BIOMASS = 0)) %>%
  dplyr::filter(SEASON == "FALL"
  ) %>%
  dplyr::select(Year = YEAR,
                Lat = LAT,
                Lon = LON,
                Biomass = BIOMASS,
                EST_TOWDATE
  ) %>%
  dplyr::distinct() %>% # remove duplicate biomass entries
  dplyr::group_by(Year, Lat, Lon) %>%
  dplyr::summarise(Biomass = sum(Biomass),
                   AreaSwept_km2 = 0.0384,
                   doy = lubridate::yday(EST_TOWDATE),
                   Vessel = "NEFSC") %>%
  dplyr::ungroup() %>%
  dplyr::distinct() %>%
  na.omit() %>%
  as.data.frame()

dat <- bfish %>%
  dplyr::group_by(Year) %>%
  dplyr::mutate(total_biomass = sum(Biomass)) %>%
  dplyr::filter(total_biomass > 0) %>%
  dplyr::ungroup() %>%
  as.data.frame()

bluefish <- list(sampling_data = dat,
                 Region = "northwest_atlantic",
                 strata.limits = list('All_areas' = 1:1e5),# full area
                 covariate_data = with(depth, data.frame(Year = Year,
                                                         Lat = Lat,
                                                         Lon = Lon,
                                                         Depth =
(log(Depth) - mean(log(Depth))) / sd(log(Depth))
                 )),
                 catchability_data =  with(dat,
                                           data.frame(Year = Year,
                                                      Lat = Lat,
                                                      Lon = Lon,
                                                      doy = doy))
)

# only depth covariate

name <- "final/univariate2"
dir.create(here::here(name),
```

```
            recursive = TRUE)
setwd(here::here(name))

settings <- make_settings( n_x = 100,
                          Region = "northwest_atlantic",
                        strata.limits = bluefish$strata.limits,
                        purpose = "index2",
                        bias.correct = TRUE,
                        use_anisotropy = TRUE,
                        fine_scale = FALSE,
                        n_categories = 3,
                        FieldConfig = c("Omega1" = 1, "Epsilon1" = 1,
                                            "Omega2" = 1, "Epsilon2" = 1),
                        RhoConfig = c("Beta1" = 0,
                                          "Beta2" = 0,
                                          "Epsilon1" = 0,
                                          "Epsilon2" = 0),
                        OverdispersionConfig = c(0,0),
                        ObsModel = c(1,1),
                        Options = c("Calculate_Range" = TRUE,
                                        "Calculate_effective_area" = TRUE),
                        treat_nonencounter_as_zero = FALSE)

fit <- fit_model(settings = settings,
                  Lat_i = bluefish$sampling_data$Lat,
                  Lon_i = bluefish$sampling_data$Lon,
                  t_i = bluefish$sampling_data$Year,
                  c_i = rep(0, nrow(bluefish$sampling_data)),
                  b_i = as_units(bluefish$sampling_data$Biomass, "kg"),
                  a_i = as_units(bluefish$sampling_data$AreaSwept_km2,
"km^2"),

                  covariate_data = bluefish$covariate_data,

                  X1_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),
                  X2_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),

                  test_fit = FALSE,
                  run_model = TRUE,
                  getsd = TRUE,
              optimize_args = list("lower" = -Inf,
                                      "upper" = Inf)
)

results <- plot(fit,
                  working_dir = paste0(here::here(name), "/")) %>%
  try()
```

```
saveRDS(results, here::here(name, "results_100.RDS"))
saveRDS(fit, here::here(name, "fit_100.RDS"))

cog <- results$Range$COG_Table
write.csv(cog, here::here(name, "cog_100.csv"))


# multivariate model data ----
bfish <- bluefish_nmfs %>%
  dplyr::mutate(size = ifelse(LENGTH >= 50, 2, # large
                            ifelse(LENGTH <= 30.3,
                                   0, # small
                                   1 # medium
                               ))) %>%
  dplyr::full_join(all_nmfs %>%
                dplyr::mutate(NUMLEN = 0,
                              size = 0)) %>%
  dplyr::full_join(all_nmfs %>%
                dplyr::mutate(NUMLEN = 0,
                              size = 1)) %>%
  dplyr::full_join(all_nmfs %>%
                dplyr::mutate(NUMLEN = 0,
                              size = 2)) %>%
  dplyr::filter(SEASON == "FALL",
                YEAR > 1981) %>%
  dplyr::select(Year = YEAR,
                EST_TOWDATE,
                Lat = LAT,
                Lon = LON,
                N = NUMLEN,
                size
  ) %>%
  dplyr::group_by(Year, EST_TOWDATE, Lat, Lon, size) %>%
  dplyr::summarise(N = sum(N),
                doy = lubridate::yday(EST_TOWDATE),
                AreaSwept_km2 = 0.0384,
                Vessel = "NEFSC") %>%
  dplyr::ungroup() %>%
  dplyr::mutate(doy = (doy - mean(doy)) / sd(doy)) %>%
  dplyr::arrange(EST_TOWDATE) %>%
  dplyr::distinct() %>%
  na.omit() %>%
  # remove years when one or more group had 0 observations
  dplyr::group_by(Year, size) %>%
  dplyr::mutate(max_n = max(N)) %>%
  dplyr::ungroup() %>%
  dplyr::group_by(Year) %>%
  dplyr::mutate(group_with_0 = min(max_n) == 0) %>%
  dplyr::filter(group_with_0 == FALSE) %>%
  dplyr::ungroup() %>%
```

```
  dplyr::group_by(Year, size) %>%
  dplyr::mutate(pos_tows = sum(as.numeric(N > 0))) %>%
  as.data.frame()

dat <- bfish %>%
  dplyr::group_by(Year, size) %>%
  dplyr::mutate(n_tows = dplyr::n()) %>%
  dplyr::filter(n_tows > 300) %>%
  dplyr::ungroup() %>%
  as.data.frame()

bluefish <- list(sampling_data = dat,
                 Region = "northwest_atlantic",
                  strata.limits = list('All_areas' = 1:1e5),# full area
                  covariate_data = with(covariates,
                                        data.frame(Year = Year,
                                                   Lat = Lat,
                                                     Lon = Lon,
                                                     Depth = (log(Depth) -
mean(log(Depth))) / sd(log(Depth)),
                                                     Temperature =
((Temperature) - mean((Temperature))) / sd((Temperature))
                                     )),
                  catchability_data =  with(dat,
                                            data.frame(Year = Year,
                                                       Lat = Lat,
                                                       Lon = Lon,
                                                       doy = doy))
)

# base model ----
name <- "final/base3"
dir.create(here::here(name),
            recursive = TRUE)
setwd(here::here(name))

settings <- make_settings( n_x = 100,
                        Region = "northwest_atlantic",
                        strata.limits = bluefish$strata.limits,
                        purpose = "index2",
                        bias.correct = TRUE,
                        use_anisotropy = TRUE,
                        fine_scale = FALSE,
                        n_categories = 3,
                        FieldConfig = c("Omega1" = 3, "Epsilon1" = 3,
                                          "Omega2" = 3, "Epsilon2" = 3),
                        RhoConfig = c("Beta1" = 0,
                                       "Beta2" = 0,
                                       "Epsilon1" = 0,
                                       "Epsilon2" = 0),
```

```
                        OverdispersionConfig = c(0,0),
                        ObsModel = c(1,1),
                        Options = c("Calculate_Range" = TRUE,
                                      "Calculate_effective_area" = TRUE),
                           treat_nonencounter_as_zero = FALSE)

fit <- fit_model(settings = settings,
                  Lat_i = bluefish$sampling_data$Lat,
                  Lon_i = bluefish$sampling_data$Lon,
                  t_i = bluefish$sampling_data$Year,
                  c_i = bluefish$sampling_data$size,
                  b_i = as_units(bluefish$sampling_data$N, "count"),
                  a_i = as_units(bluefish$sampling_data$AreaSwept_km2,
"km^2"),

                  test_fit = FALSE,
                  run_model = TRUE,
                  getsd = TRUE,
                  optimize_args = list("lower" = -Inf,
                                        "upper" = Inf),
                  newtonsteps = 5
)

results <- plot(fit,
                  working_dir = paste0(here::here(name), "/")) %>%
  try()

saveRDS(results, here::here(name, "results_100.RDS"))
saveRDS(fit, here::here(name, "fit_100.RDS"))

cog <- results$Range$COG_Table
write.csv(cog, here::here(name, "cog_100.csv"))

# depth model ----
name <- "final/depth2"
dir.create(here::here(name),
            recursive = TRUE)
setwd(here::here(name))

settings <- make_settings( n_x = 100,
                        Region = "northwest_atlantic",
                        strata.limits = bluefish$strata.limits,
                        purpose = "index2",
                        bias.correct = TRUE,
                        use_anisotropy = TRUE,
                        fine_scale = FALSE,
                        n_categories = 3,
                        FieldConfig = c("Omega1" = 3, "Epsilon1" = 3,
                                         "Omega2" = 3, "Epsilon2" = 3),
                        RhoConfig = c("Beta1" = 0,
```

```
                                        "Beta2" = 0,
                                         "Epsilon1" = 0,
                                         "Epsilon2" = 0),
                        OverdispersionConfig = c(0,0),
                        ObsModel = c(1,1),
                        Options = c("Calculate_Range" = TRUE,
                                       "Calculate_effective_area" = TRUE),
                           treat_nonencounter_as_zero = FALSE)

fit <- fit_model(settings = settings,
                   Lat_i = bluefish$sampling_data$Lat,
                  Lon_i = bluefish$sampling_data$Lon,
                  t_i = bluefish$sampling_data$Year,
                  c_i = bluefish$sampling_data$size,
                  b_i = as_units(bluefish$sampling_data$N, "count"),
                  a_i = as_units(bluefish$sampling_data$AreaSwept_km2,
"km^2"),

                  covariate_data = bluefish$covariate_data,

                  X1_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),
                   X2_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),

                  test_fit = FALSE,
                  run_model = TRUE,
                  getsd = TRUE,
                  optimize_args = list("lower" = -Inf,
                                        "upper" = Inf)
)

results <- plot(fit,
                  working_dir = paste0(here::here(name), "/")) %>%
  try()

saveRDS(results, here::here(name, "results_100.RDS"))
saveRDS(fit, here::here(name, "fit_100.RDS"))

cog <- results$Range$COG_Table
write.csv(cog, here::here(name, "cog_100.csv"))

## day of year model ----
name <- "final/doy2"
dir.create(here::here(name))
setwd(here::here(name))

settings <- make_settings( n_x = 100,
                        Region = "northwest_atlantic",
                        strata.limits = bluefish$strata.limits,
```

```
                        purpose = "index2",
                        bias.correct = TRUE,
                        use_anisotropy = TRUE,
                        fine_scale = FALSE,
                        n_categories = 3,
                        FieldConfig = c("Omega1" = 3, "Epsilon1" = 3,
                                           "Omega2" = 3, "Epsilon2" = 3),
                        RhoConfig = c("Beta1" = 0,
                                        "Beta2" = 0,
                                         "Epsilon1" = 0,
                                         "Epsilon2" = 0),
                        OverdispersionConfig = c(0,0),
                        ObsModel = c(1,1),
                        Options = c("Calculate_Range" = TRUE,
                                       "Calculate_effective_area" = TRUE),
                           treat_nonencounter_as_zero = FALSE)

fit <- fit_model(settings = settings,
                   Lat_i = bluefish$sampling_data$Lat,
                  Lon_i = bluefish$sampling_data$Lon,
                  t_i = bluefish$sampling_data$Year,
                  c_i = bluefish$sampling_data$size,
                  b_i = as_units(bluefish$sampling_data$N, "count"),
                  a_i = as_units(bluefish$sampling_data$AreaSwept_km2,
"km^2"),

                  catchability_data = bluefish$sampling_data,
                  covariate_data = bluefish$covariate_data,

                  X1_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),
                  X2_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE),
                  Q1_formula = ~ doy,
                  Q2_formula = ~ doy,

                  test_fit = FALSE,
                  run_model = TRUE,
                  getsd = TRUE,
                  optimize_args = list("lower" = -Inf,
                                          "upper" = Inf)
)

results <- plot(fit,
                  working_dir = paste0(here::here(name), "/")) %>%
  try()

saveRDS(results, here::here(name, "results_100.RDS"))
saveRDS(fit, here::here(name, "fit_100.RDS"))
```

```
cog <- results$Range$COG_Table
write.csv(cog, here::here(name, "cog_100.csv"))


# temperature model ----
name <- "final/temp3"
dir.create(here::here(name))
setwd(here::here(name))

settings <- make_settings( n_x = 100,
                        Region = "northwest_atlantic",
                        strata.limits = bluefish$strata.limits,
                        purpose = "index2",
                        bias.correct = TRUE,
                        use_anisotropy = TRUE,
                        fine_scale = FALSE,
                        n_categories = 3,
                        FieldConfig = c("Omega1" = 3, "Epsilon1" = 3,
                                        "Omega2" = 3, "Epsilon2" = 3),
                        RhoConfig = c("Beta1" = 0,
                                      "Beta2" = 0,
                                      "Epsilon1" = 0,
                                        "Epsilon2" = 0),
                        OverdispersionConfig = c(0,0),
                        ObsModel = c(1,1),
                        Options = c("Calculate_Range" = TRUE,
                                    "Calculate_effective_area" = TRUE),
                           treat_nonencounter_as_zero = FALSE)

fit <- fit_model(settings = settings,
                  Lat_i = bluefish$sampling_data$Lat,
                  Lon_i = bluefish$sampling_data$Lon,
                 t_i = bluefish$sampling_data$Year,
                  c_i = bluefish$sampling_data$size,
                  b_i = as_units(bluefish$sampling_data$N, "count"),
                  a_i = as_units(bluefish$sampling_data$AreaSwept_km2,
"km^2"),

                   catchability_data = bluefish$sampling_data,
                  covariate_data = bluefish$covariate_data,

                  X1_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE) +
                  splines::bs(Temperature, knots = 3, intercept = FALSE),
                  X2_formula = ~ splines::bs(Depth, knots = 3, intercept =
FALSE) +
                  splines::bs(Temperature, knots = 3, intercept = FALSE),
                  Q1_formula = ~ doy,
                  Q2_formula = ~ doy,
```

```
                   test_fit = FALSE,
                  run_model = TRUE,
                  getsd = TRUE,
                  optimize_args = list("lower" = -Inf,
                                      "upper" = Inf),
                  newtonsteps = 5
)

results <- plot(fit,
                working_dir = paste0(here::here(name), "/")) %>%
  try()

saveRDS(results, here::here(name, "results_100.RDS"))
saveRDS(fit, here::here(name, "fit_100.RDS"))

cog <- results$Range$COG_Table
write.csv(cog, here::here(name, "cog_100.csv"))
```