

Ecosystem Indicator Report Final Analysis

Katie Lankowicz

18 December, 2023

Contents

1	Takeaways	1
2	Continuous temperature data	3
3	“Weird” species	8
3.1	Current RShiny App	8
3.2	Next steps for RShiny app	8
4	Growth	10
4.1	The problem	10
4.2	Kernel density estimation and modes	11
4.3	Mixed distribution modeling	12
4.4	Eyeballing	13
4.5	GLMs to model age-group length bounds	14
4.6	Modeling growth	15
4.7	Results	22

1 Takeaways

- Portland Harbor tide gauge data gives us the best characterization of nearshore sea surface temperature as compared to air temperature from the airport or field-collected SST.
- The climate reference period (CRP) has been adjusted to 2002-2020 to as closely as possible match the warming report’s CRP (1991-2020). Data prior to 2002 are not available for the tide gauge. With the new CRP, we still get the same collection of years as 5 hottest and 5 warmest.
- I have collected the time series of CBASS seine catch and theoretical distribution maps for most of our encountered species into an RShiny application (https://klankowicz.shinyapps.io/CBASS_Fish_Dist/). Check it out.
- I have calculated age-group-specific growth rates for our six most-encountered teleosts. Results indicate that nearshore residents silversides and mummichogs trend toward faster growth in warmer years, though this was only statistically significant for one age group of silversides. Species that undergo latitudinal, inshore-offshore, or diadromous migrations had mixed results and no statistically significant differences in growth rates between cold and hot years. It’s possible that growth rates of these species are driven by gradients of environmental factors across areas much larger than Casco Bay alone.

- There is a possibility of including another tree plot (the plot with all the rectangles) showing relative abundance of captured species from the jigging surveys of 2014, 2015, and 2023. I have the data, but need to make the plot.

2 Continuous temperature data

One of our main goals is connecting the ecological story of Casco Bay in the last 10 years with existing temperature data. We're generally interested in quantifying annual temperature anomalies to pick up on any high-level differences in nearshore ecology that may be related to temperature.

In the last meeting, we discussed making the climate reference period (CRP) used to calculate mean annual temperature anomalies the same as what is used in the warming report. The warming report has used various CRPs to identify average daily temperatures, but currently uses 1991-2020. We will attempt to use this same reference period, as allowed by data availability.

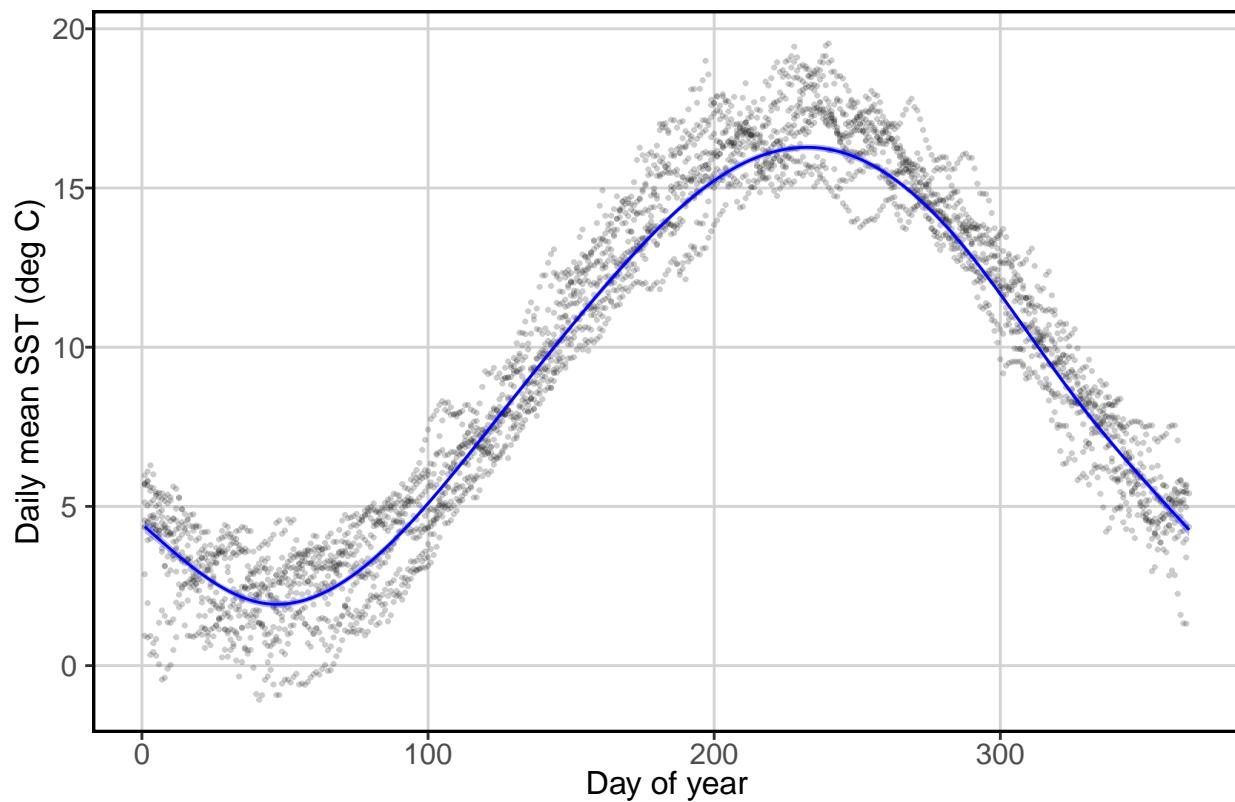
2.0.1 Portland Harbor tide gauge

The Portland Harbor tide gauge likely gives the closest approximation to the nearshore environment where CBASS seining is conducted. We were previously using 2003-2023 as the reference period and found that 2014/ 2015/ 2017-2019 had negative mean annual anomalies. The remaining years in our 2014-2023 time series had positive mean annual anomalies. It seemed intuitive to therefore categorize years with negative mean annual anomalies as “colder” and those with positive anomalies as “warmer.”

Unfortunately, the tide gauge has only collected surface water temperature data since 1997. Data are extremely spotty and unreliable until early 2002. Therefore, the closest we can get to matching the warming report’s CRP is 2002-2020. I have pulled and cleaned the raw data for this period. I also slightly improved the fit of the model of daily anomalous temperatures by more carefully cleaning out invalid temperatures (caused by equipment malfunctions). I also slightly changed the model structure of the GAM that models daily anomalous temperatures by specifying cubic spline regression smooths with penalized shrinkage (`bs='cs'`) and REML fitting.

Despite this change in reference period, data cleaning, and model structure, we get the same results. 2014/2015/2017-2019 all have negative annual anomalies and the remaining years have positive. The years even fall in the same order when ordering from most negative to most positive anomaly. This order differs slightly from the warming report, but this can be attributed to both the different reference periods and the enormously different spatial scales. These spatially-variable differences could spark some cool conversation about how the cumulative exposure of larvae/migratory fishes to detrimental warming offshore impacts our inshore communities, even if inshore conditions are about average.

All this to say: I think we keep this approach to qualifying temperature anomalies. I recognize that calling years flatly “hot” or “cold” can be misleading. We can shift the language to something more like “5 hottest/ 5 coldest years of the time series”. The warming report does a similar ranking that gets updated both annually and seasonally.

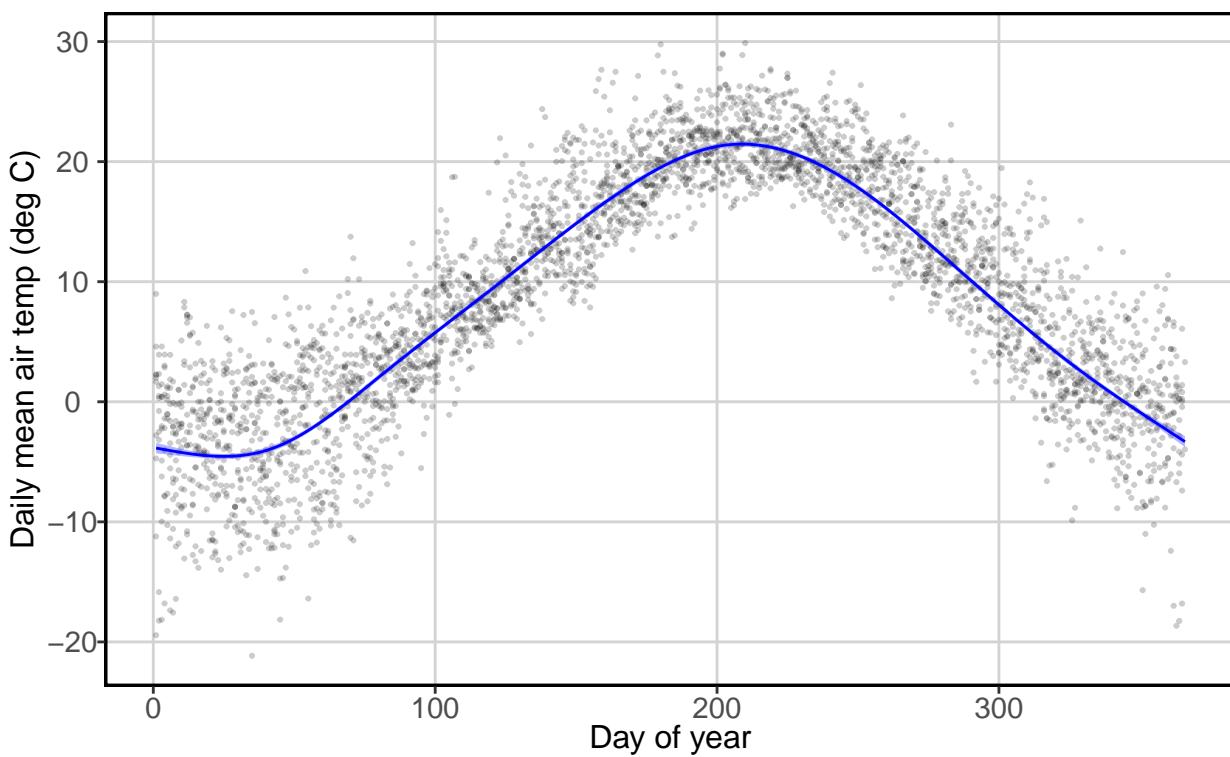


Year	Gauge anomaly (deg C)	OISST anomaly (deg C)
2014	-0.62	1.02
2015	-0.27	0.99
2016	0.58	1.52
2017	-0.44	1.02
2018	-0.12	1.36
2019	-0.26	NA
2020	0.72	1.38
2021	1.33	2.31
2022	0.92	2.07
2023	1.14	NA

2.0.2 Weather from Portland airport

I'll be brief with this one. We do have air temperature data from Portland Jetport for the same climate reference period (1991-2020) as the warming report. The minimum and maximum smoothed daily air temperatures lead smoothed daily sea surface temperatures by about 22 days, so we could consider adding a lag when calculating annual mean temperature anomaly. I have not yet done this.

Despite having the same reference period, the air temperature and OISST-derived mean annual temperature anomalies are not similar, either in magnitude or order of warmest to coldest year. We wouldn't expect them to match exactly, as the physical mechanisms linking the two are complicated. I think this highlights a larger point—we want our temperature time series to match what's going on in the nearshore environment as closely as possible. Air temperature, though useful because its data records reach further back in time, is not a close enough approximation. I think it's cool to look at, but will be dropping it from further analysis.



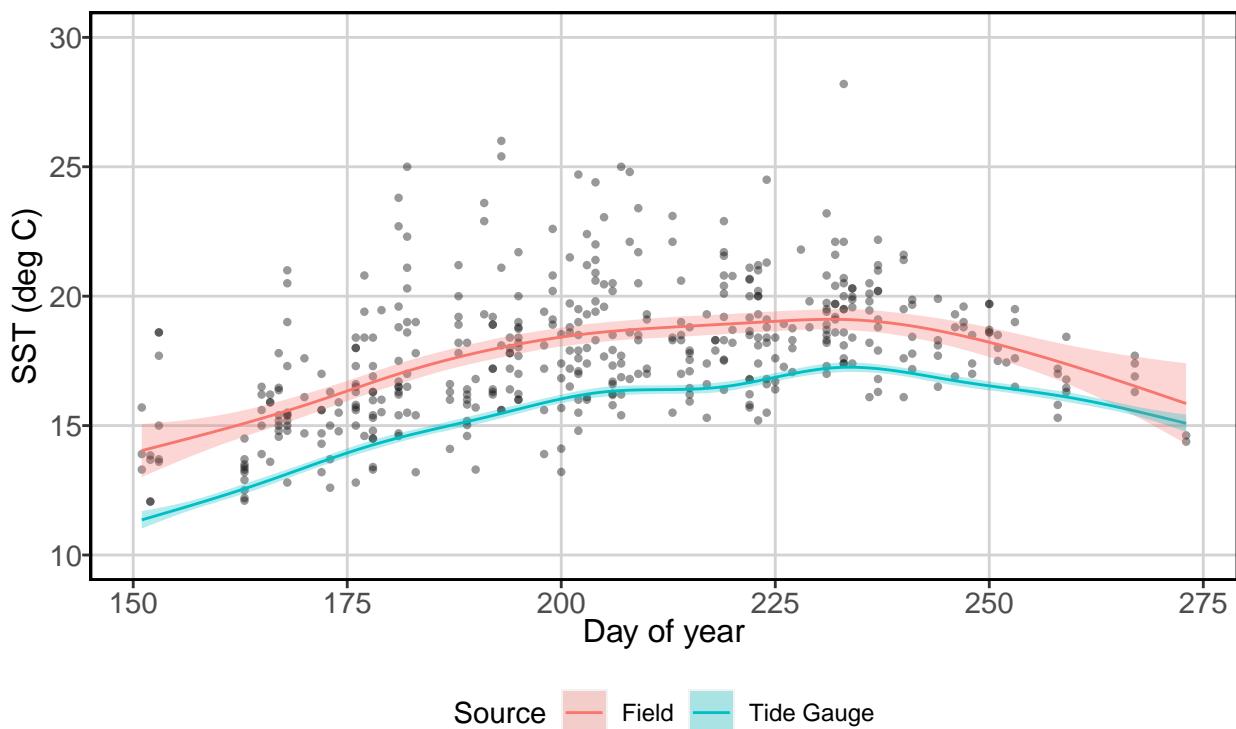
Year	Air temp anomaly (deg C)	OISST anomaly (deg C)
2014	-0.30	1.02
2015	0.16	0.99
2016	0.83	1.52
2017	0.34	1.02
2018	0.31	1.36
2019	-0.01	NA
2020	1.44	1.38
2021	1.01	2.31
2022	0.62	2.07
2023	0.80	NA

2.0.3 CBASS field data

We can do a simple sanity check on the match between the Portland Harbor tide gauge SST and our field sites SST. Keep in mind that field site SST is inconsistent in the frequency of collection (missing some data in 2015-2017 and 2020-2021), the equipment used (sometimes YSI, sometimes boat-mounted sensor), and the frequency of calibration (we think this was forgotten or incorrect in 2023). Also keep in mind that we only sample in daylight hours (7AM-4PM) to align with the working day. I filtered and re-smoothed the Portland Harbor gauge data so it only includes points collected 7AM-4PM 2014-2023. For the best match between data sources, I also excluded all field SST data collected at QBC and Presumpscot River sites. These are too dissimilar from Portland Harbor.

In the following plot, the red line represents the GAM-smoothed mean daily temperature of our field-collected data (Mid-bay and Outer-bay sites only). The blue line represents GAM-smoothed mean daily temperature of the Portland Harbor tide gauge, filtered so the temporal extent matches that of the field data. The field data varies from 0.8 to 2.7 degrees hotter than the Portland Harbor data, with a mean of 2.1 degrees hotter. Portland Harbor is about in the latitudinal center of the mid-bay and outer-bay collection of sites, so this increased heat at the field sites probably isn't due to a latitudinal gradient, especially within such a small spatial domain.

I can't pinpoint why exactly the difference in smoothed temperature is so dramatic. It could be due to both low data load in "colder" years and uncalibrated YSI/ boat temperature sensors. Bottom line, these data have some problems that I cannot overcome and therefore I will not use them past this point.



3 “Weird” species

We want to characterize how common or uncommon the various species we capture are. I made an RShiny application (https://klankowicz.shinyapps.io/CBASS_Fish_Dist/) to illustrate both the spatial distribution and the time series of catch for most teleosts we saw. *PLEASE DO NOT SHARE THIS LINK OUTSIDE THIS GROUP.* I use the free version of RShiny and have a limited quota of “active hours” per rolling month, so a bunch of people accessing it will quickly limit my ability to use the app building functions for this and other projects.

3.1 Current RShiny App

Here's a tour of functionality and plots.

You can select a species from the drop-down menu on the left, and the app will build results based on your selection. For each species, three plots will appear. The leftmost is a map of theoretical spatial distribution. The top right is expected distribution of that species along its latitudinal gradient, with the center of gravity and Casco Bay identified. The bottom right is a time series of our capture of that species. Details for each plot follow. At the very top of the page, you'll see a readout that identifies how far south or north Casco Bay is in kilometers from the calculated center of gravity.

I pulled all available theoretical spatial distributions of our captured species from FishBase. Some species have not yet been modeled. Modelers have identified environmental characteristics (eg. depth, SST, salinity) important to various fishes and developed maps of “relative probability of occurrence” based on known spatial distributions of these environmental characteristics. A few of these maps are probably far from reality (see bluefish), but most seem reasonable. I use the maps to calculate a weighted center of gravity for each species’ distribution. I also then plot a kernel density estimate of relative abundance along a latitudinal gradient, identifying both the center of gravity and the location of Casco Bay along that gradient so it's easy to identify whether that species is in the northern or southern half of its range, respective to the center of gravity.

The time series plot is a raster in which rows represent weeks of the year and columns represent years. The color of the year labels on the y axis indicates whether that year is in the hottest 5 (red) or coldest 5 (blue) year period of our Portland Harbor temperature time series. The color of each grid cell is a gradient of log-normalized abundance captured in that week. Cool colors indicate fewer fish caught, warm colors indicate more fish caught, and grays indicate no fish caught. Transparent raster cells indicate no sampling took place in that week. The color scale is stretched along all catch values for all years, so you can view relative abundance both between weeks of a single year and weeks of different years.

Check out these fish to see representations of common and uncommon species:

* Uncommon / South of center: shorthorn sculpin, ninespine stickleback * Uncommon / North of center: crevalle jack, permit * Common / At center: Atlantic silversides * Common / South of center: Atlantic herring * Common / North of center: mummichog

3.2 Next steps for RShiny app

We should have a discussion about our ability to pay for a full RShiny license as an institute. It's a great tool and multiple labs are already using it. Alternatively, is it possible to build a similar dropdown selection application on our own website? I could save the output for all the species we want to include as static images, so we wouldn't need to link to an online version of R to generate those.

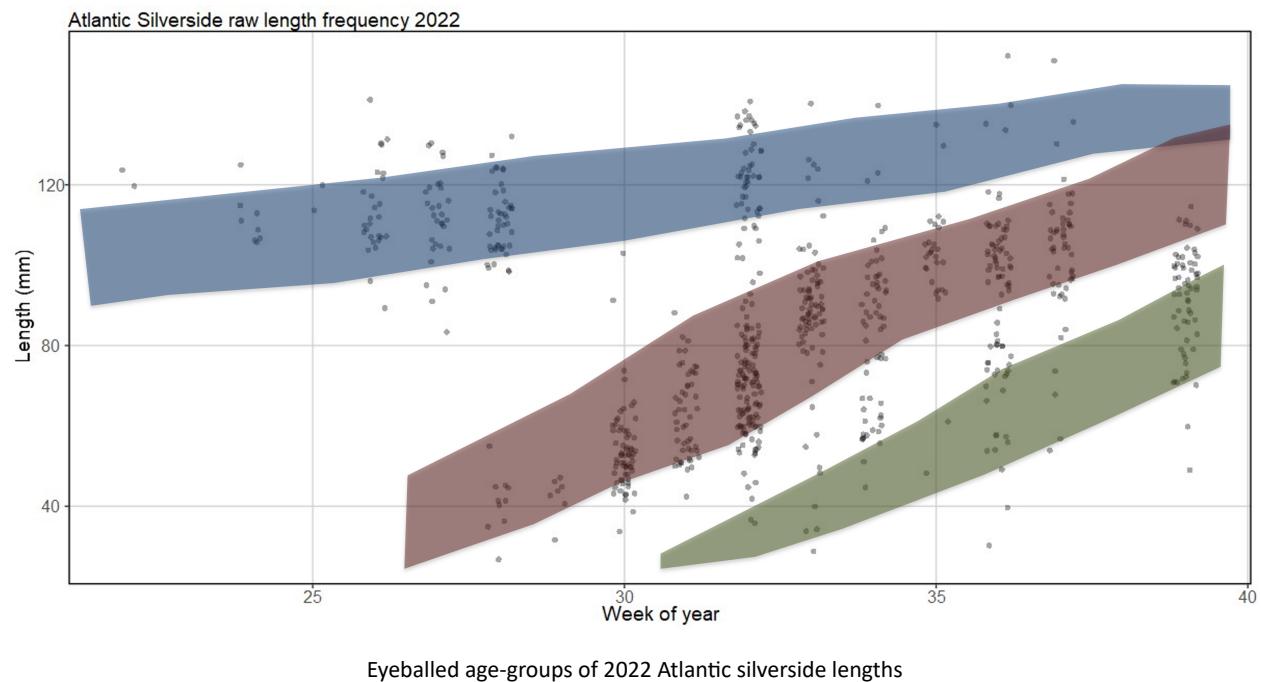
Regardless of whether we pursue RShiny licenses/ our own dropdown method, a few changes will be made. I'm going to place species into these Commonality-Distribution groups seen above, then plot relative abundance of these groups over our time period. In theory, this could illustrate any high-level emergent properties of temperature-driven distribution shifts.

4 Growth

Last time we met, figuring out how to quantify growth rates was escaping me. I have since developed a mostly statistically-robust method of accomplishing this task. There's a bit of less-robust handwaving, but not too much. Let's go through it.

4.1 The problem

We sample a population of mixed age-based groups. I won't say cohort, because I think that will incite arguments about the nature of a "cohort" and obscure my point here. Those groups, by nature of being different ages, have distinct size distributions and growth rates. We can sometimes identify these groups just by eyeballing length distributions over time, like the theoretical groups I plotted below based on Atlantic silversides 2022 length distribution.



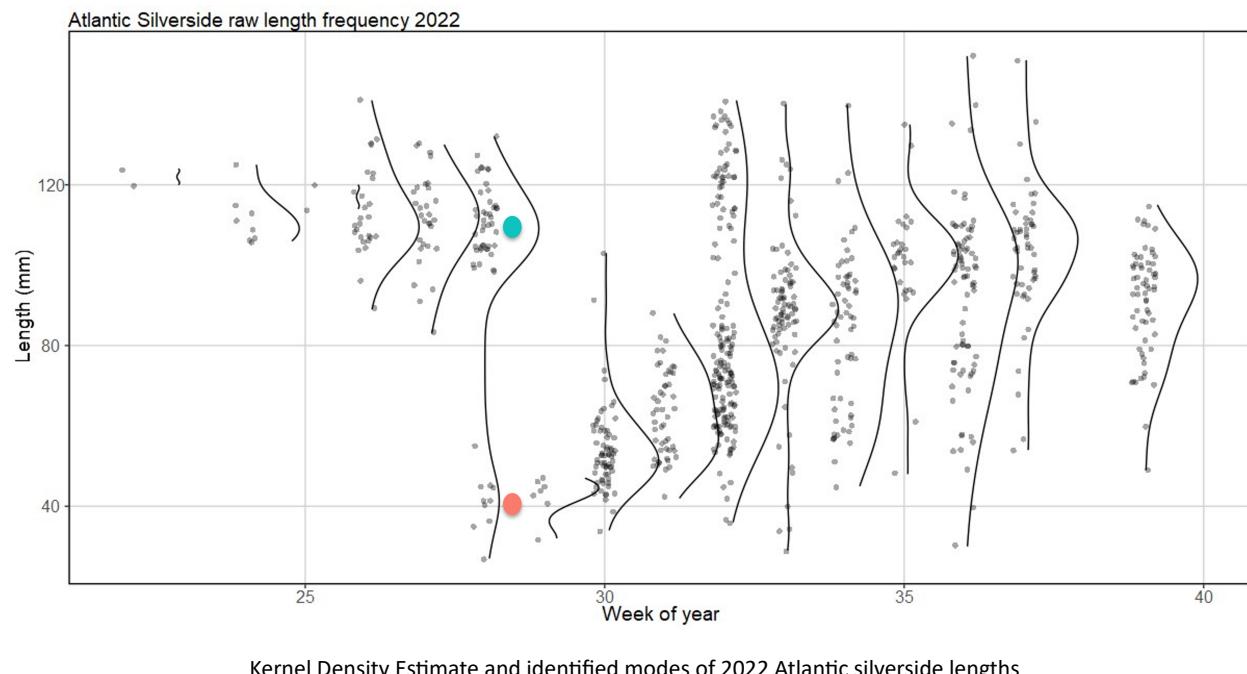
We need to separate the groups as best as possible for all species-years combinations so we can more clearly see realistic growth rates. Note that the eyeballed version presented here is from probably the most clear example of separate age groups out of all our species-year combinations, and the "eyeball" method will be neither consistent nor robust for more noisy data.

4.2 Kernel density estimation and modes

The first step of separating age-groups is to make some assumptions. Let's assume that the lengths of conspecifics in the same age group come from a normal distribution, and the range of each age-group's length distribution has to be reasonably compact (less than 40% of the maximum asymptotic body length for the species, as pulled from FishBase).

We then plot a kernel density estimate of lengths by species for each week of sampling. Considering our assumptions, we should see peaks of length-density around the mean length for each age group. Deterministic methods can be used to determine both the number of modes present in the length-density distribution and the length-values at those modes. These should theoretically be mean length for an age-group.

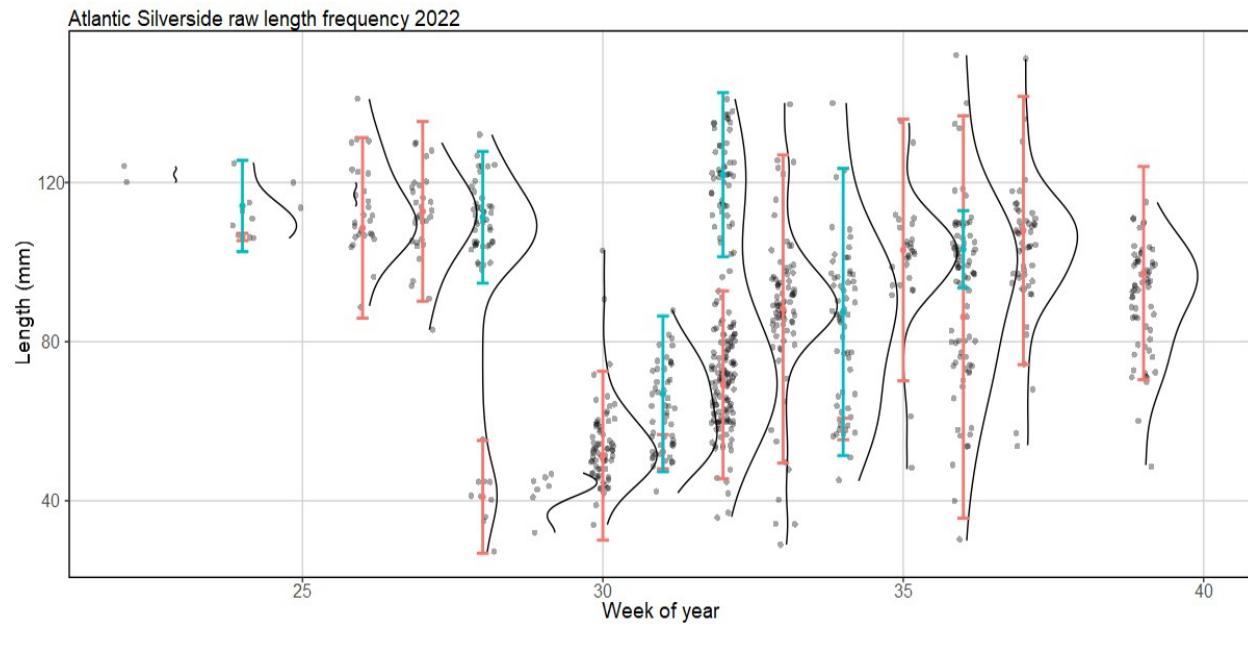
At this point, we must remove observations from weeks in which fewer than 7 fish of a species were lengthed— it isn't enough data to continue. This limits this analysis to the most frequently-encountered teleosts: alewife, atlantic herring, silverside, mummichog, sandlance, and winter flounder. The approach does not work for other commonly-encountered species (tomcod, stickleback spp, sculpin spp) because there aren't enough observations of lengths. Finally, I am hesitant to use this approach for green crabs. I don't know if my length and growth assumptions hold for crustaceans. I can try, if there is interest.



4.3 Mixed distribution modeling

From here, we use mixed distribution modeling to calculate the upper and lower bounds that create a 95% confidence interval around our identified mean lengths, effectively identifying the length range of each age-group. Again, our assumption is that distribution of lengths is normal, so we use an expectation-maximization (EM) algorithm for mixtures of univariate normals as our approach. The normal-mix EM model is given the lengths of all conspecifics per week and the number of modes identified in the previous step. It then returns mu (means of each component), sigma (standard deviations), and lambda (mixing proportions). For each component, we calculate an approximation to a 95% CI by finding 2 standard deviations around the mean.

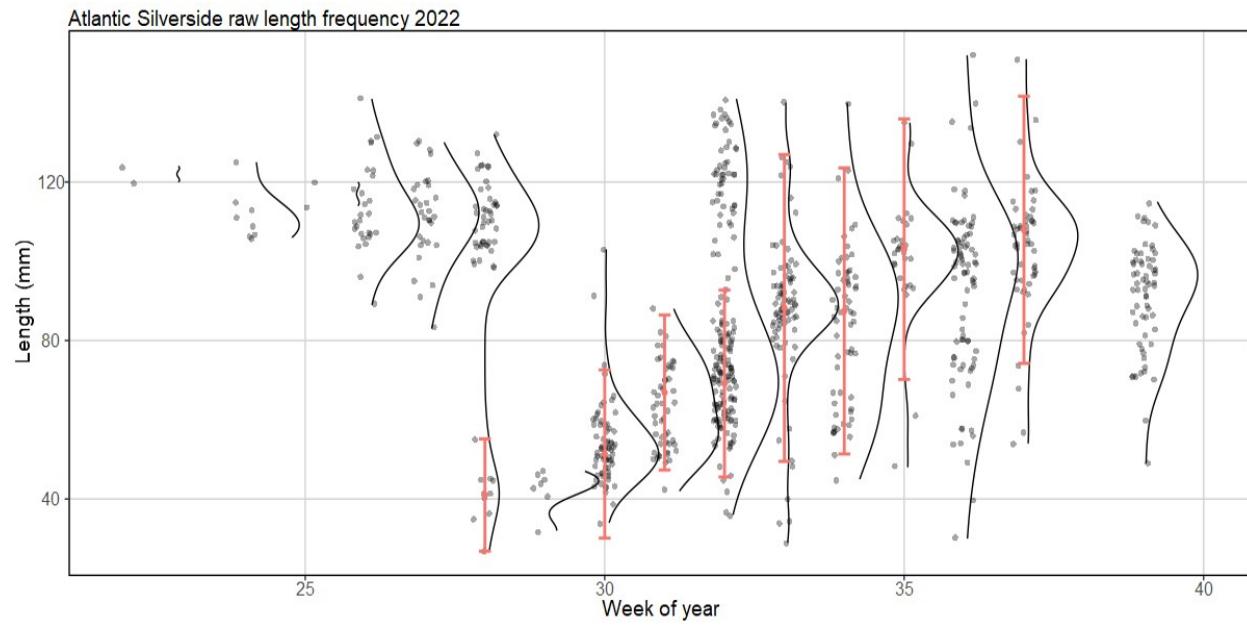
This gets us pretty close to where we want to be. We can remove distributions where the upper or lower bounds are biologically invalid (less than 0 cm or above the reported maximum asymptotic body length) or biologically improbable (range is greater than 40% of maximum asymptotic body length). We can also melt together distributions which overlap each other by more than 80%, as this isn't enough separation to clearly identify age-groups anyway.



4.4 Eyeballing

The next step is to give age-groups a distinct identifier and confirm that their ID remains the same between weeks. I haven't come up with a good method to automate this. I just go off my plots and returned mixed distribution model values and track age-groups by eye. Here's a theoretical example of what I do: in week 1, we identify one distribution of mean length 100 cm. In week 2, we identify two distributions, one of mean length 102 cm and one of mean length 20 cm. In week 3, we identify one distribution of mean length 50 cm. We would assume that the groups with mean length 100 cm in week one and mean length 102 cm in week 2 are the same age-group growing over time and make sure they have the same age-group identifier. The groups with mean length 20cm in week 2 and mean length 50 cm in week 3 are not linked to each other, nor to age-group 1. They need unique identifiers.

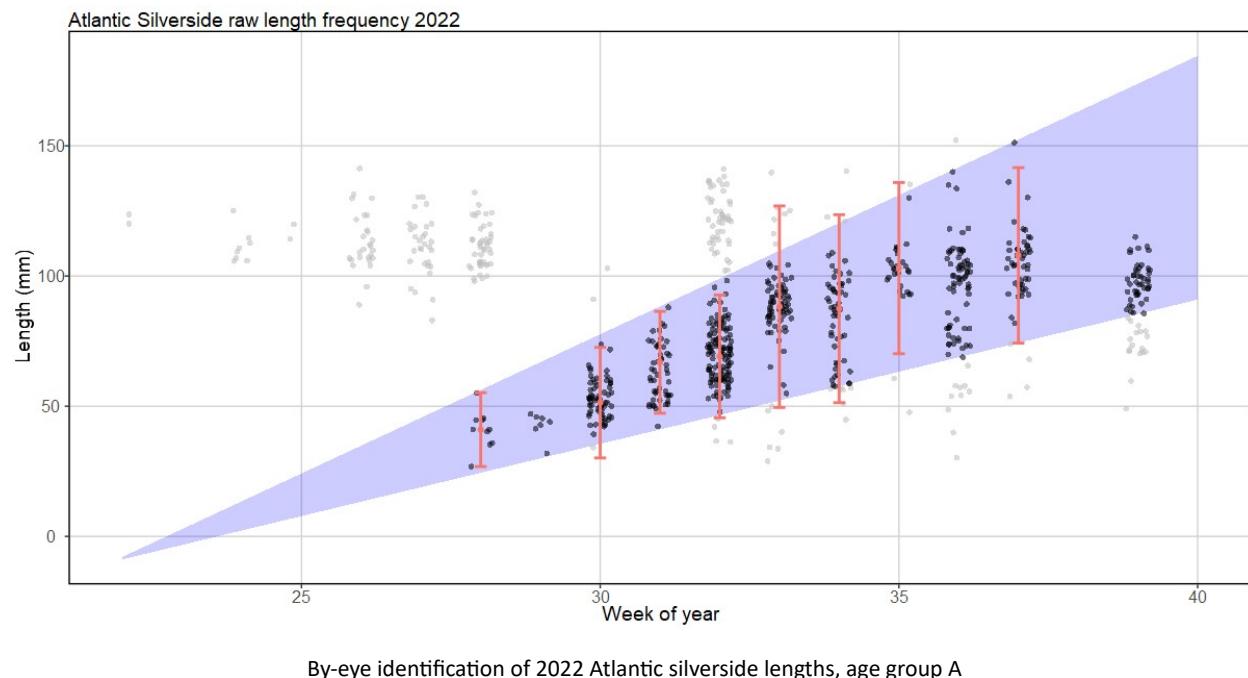
I'm sure there's some kind of way to code this out to do it automatically, but it wasn't worth my time to figure it out with all the edge cases. Importantly, I also did not want to *a priori* assume any maximum or minimum growth rates, which would have made this process much easier.



4.5 GLMs to model age-group length bounds

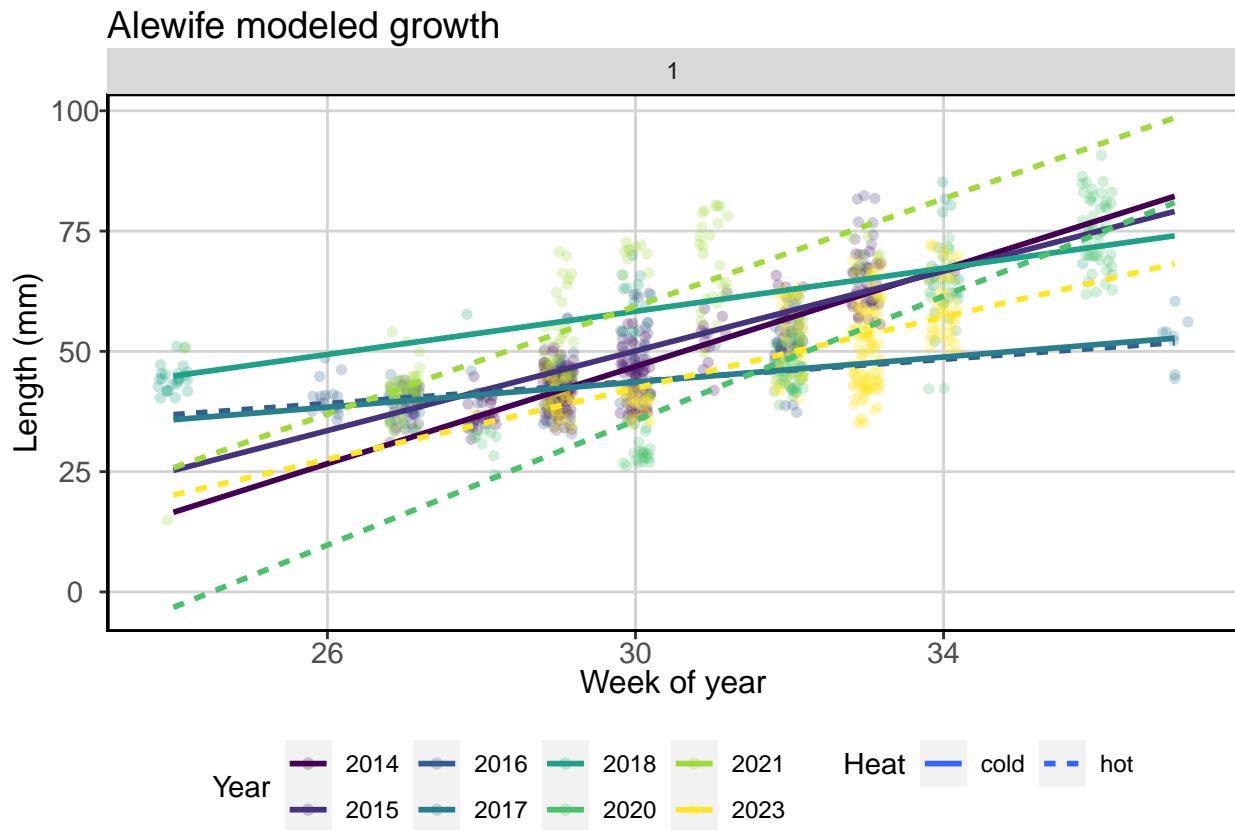
This leaves us with distributions of length for various age-groups for some weeks of a sampling year. Remember that we did remove some biologically improbable distributions, so sometimes we are left with no identified length distributions in a week. This is fine. The lower bounds, means, and upper bounds for each age group *generally* are monotonically increasing for the remaining weeks, but there is some variation. We also need to fill the gaps we made. The next step is therefore to model the upper and lower bounds of length in each week. Here, another assumption is made— that growth in this extremely short period (typically June 8 - August 28) is approximately linear. We therefore use generalized linear models to characterize length distribution bounds.

Once the GLM process has been completed, we can assign age-groups of fish accordingly.

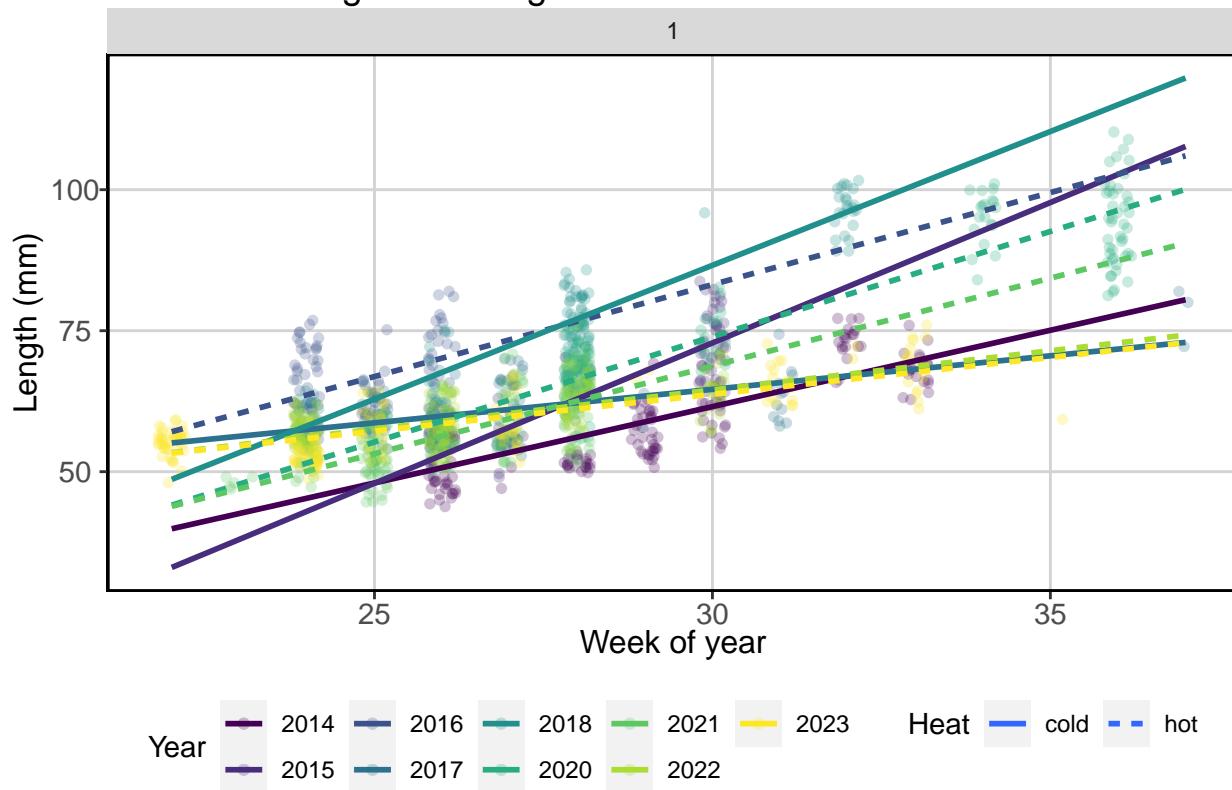


4.6 Modeling growth

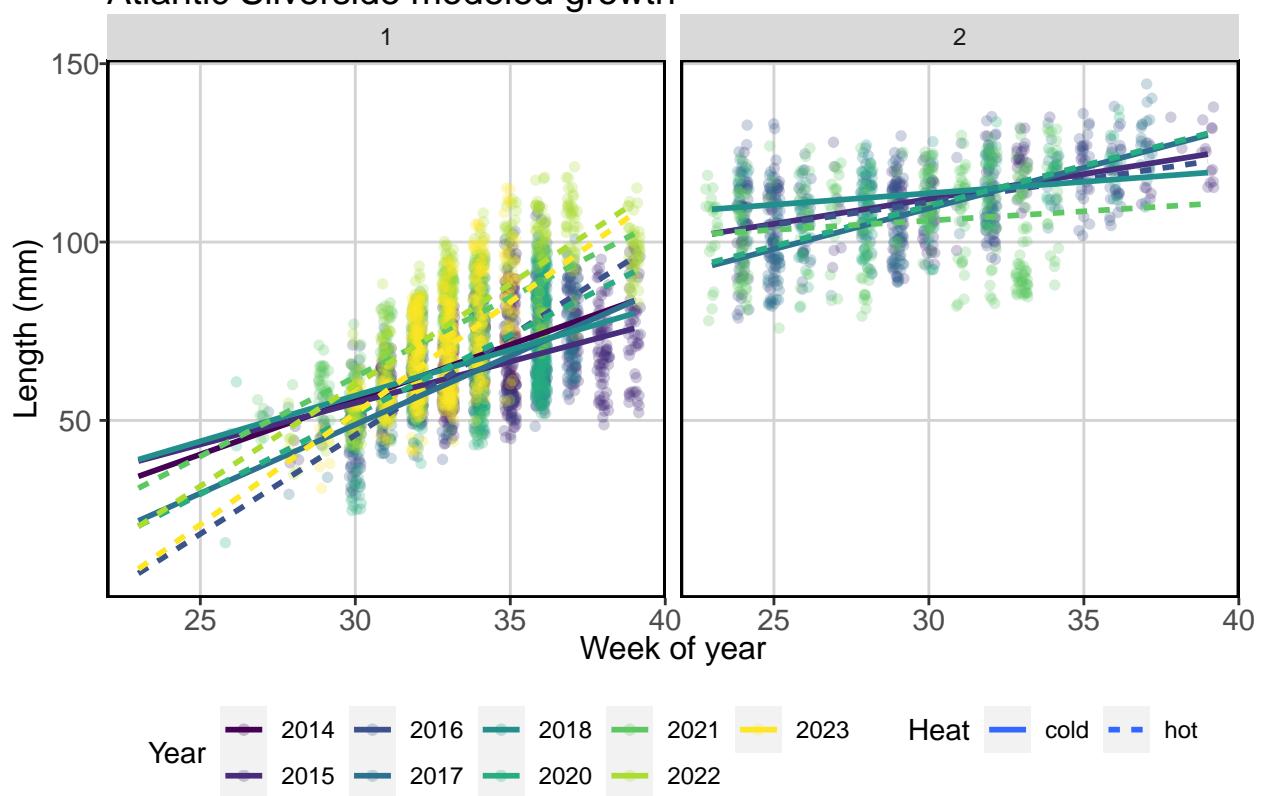
This last bit is what finally gives us the results we wanted. Now that we have assigned fish into age groups by length in each week, we can model the growth of those age groups. Again, we assume that growth is approximately linear over our short time period and use a generalized linear model. Growth should be equal to the slope of the GLM.



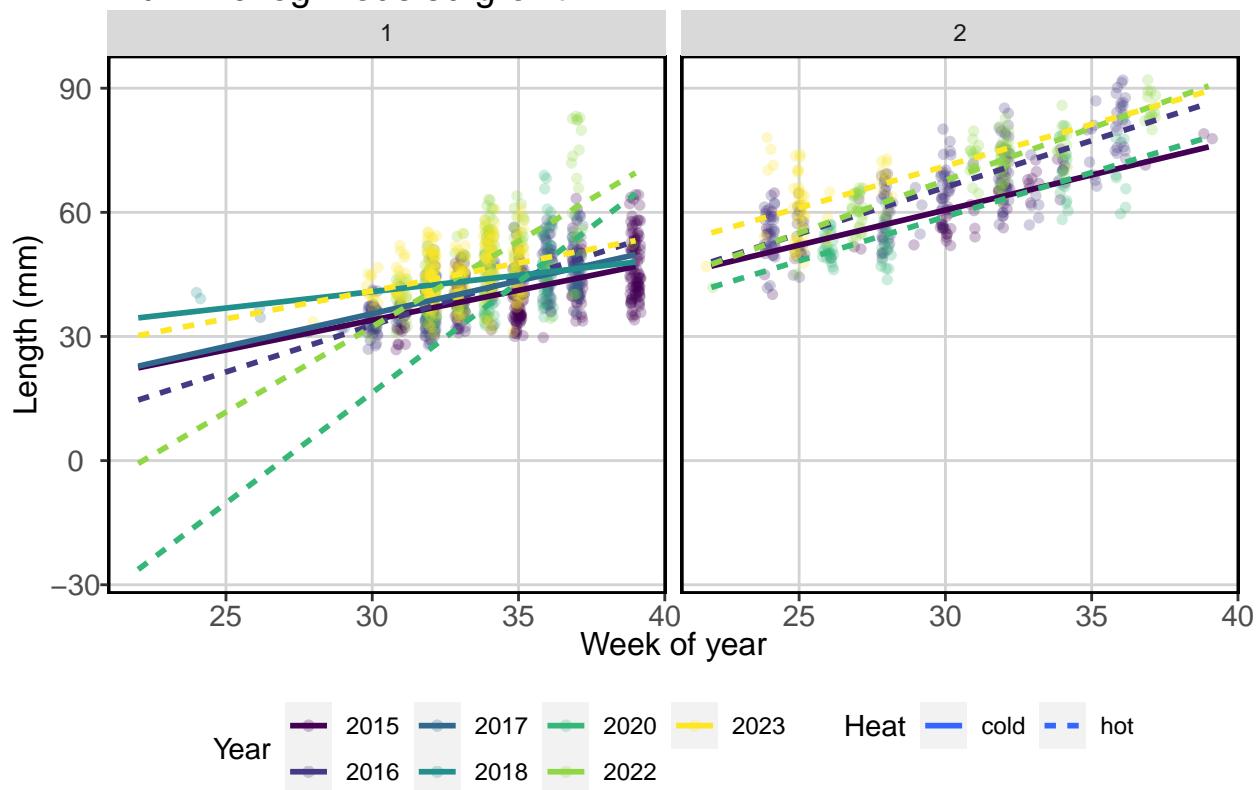
Atlantic Herring modeled growth



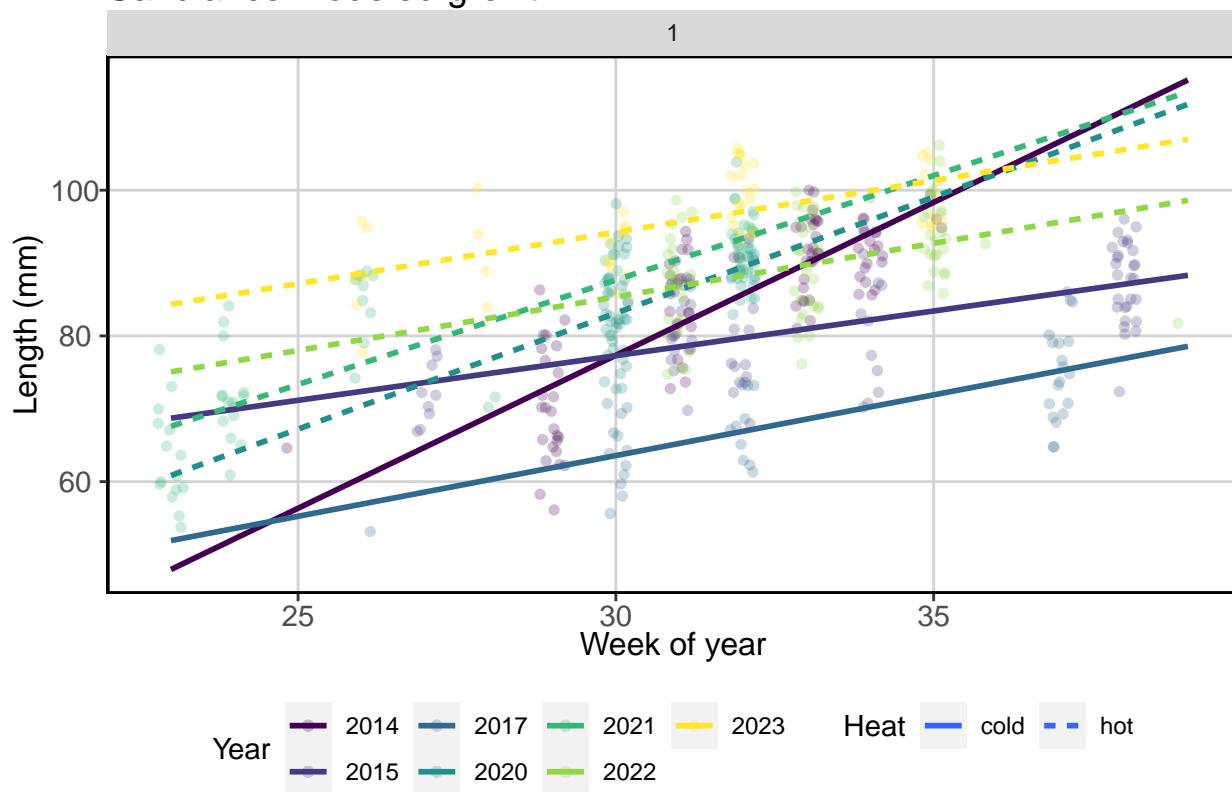
Atlantic Silverside modeled growth



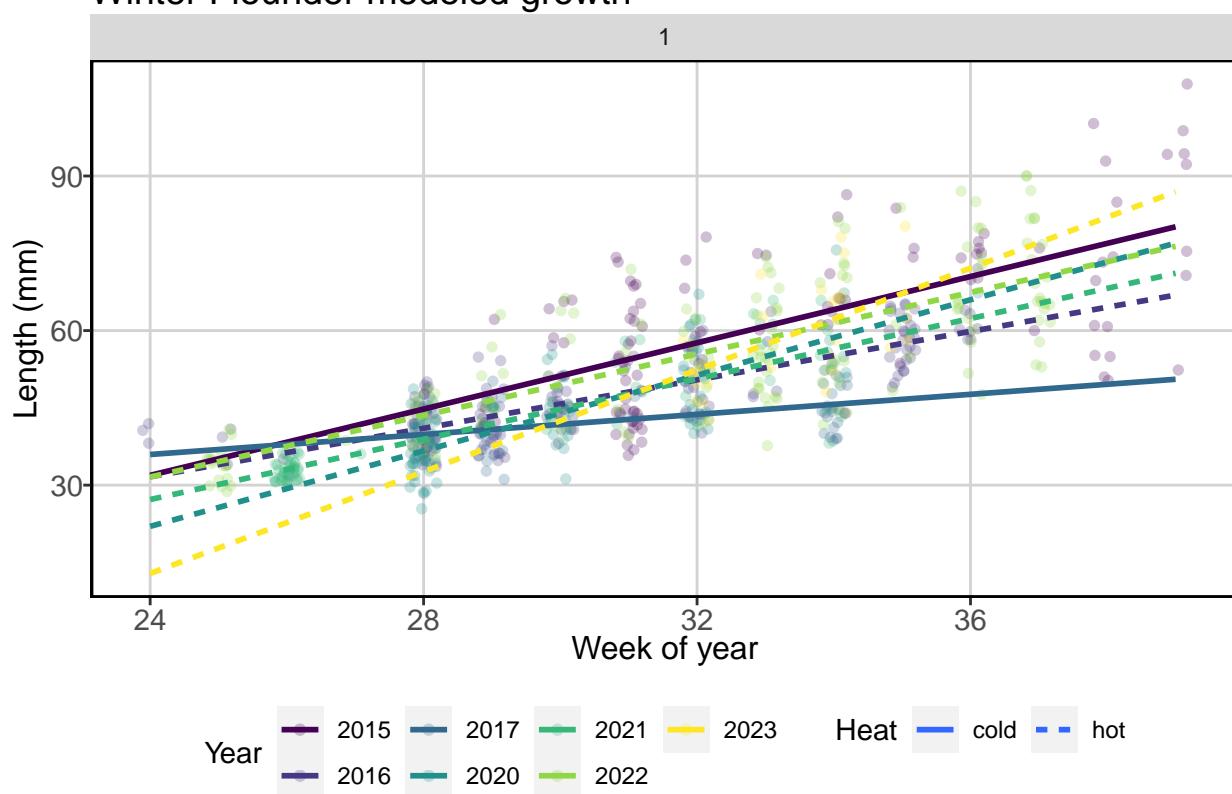
Mummichog modeled growth

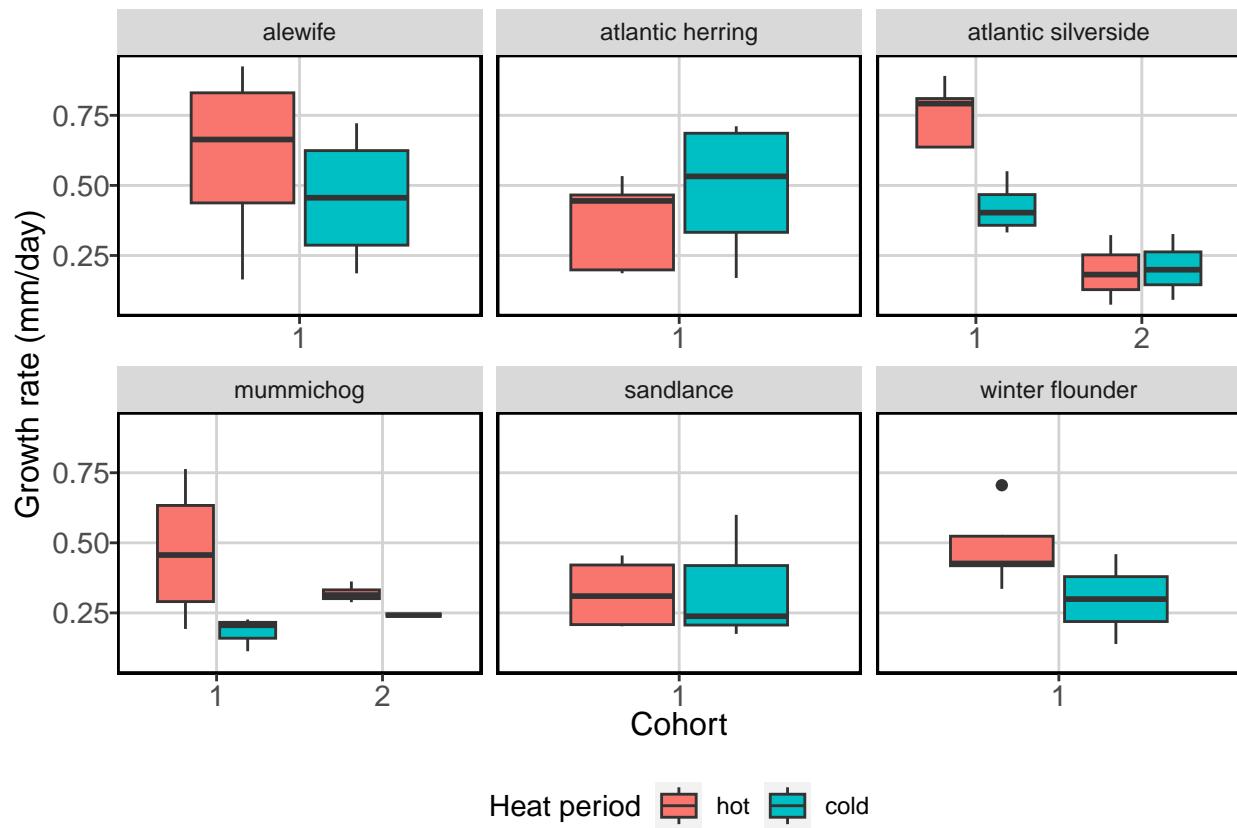


Sandlance modeled growth



Winter Flounder modeled growth





4.7 Results

Of the six modeled species, 2 are likely north of their populations' centers of gravity: atlantic silverside and mummichog. We would expect to see increased growth rates in hotter years. Because mummichog and silversides are short-lived (4 and 2 years expected lifespan, respectively) and lifelong residents of the inshore area, we were able to model growth rates for 2 age groups for both species. We found a trend of increased growth rates in hotter periods for age groups 1 and 2 for mummichogs, and age group 1 for silversides. Silverside age group 2 had a slightly increased growth rate in colder years. The only significant difference in growth rates for any species occurred in the silverside age-group 1: significantly higher growth rates in hotter years than colder years.

Three of the remaining species (alewife, Atlantic herring, and winter flounder) are likely south of their populations' centers of gravity. Increased temperatures should lead to slower growth rates. However, no significant differences were found, and only Atlantic herring exhibited a trend of faster growth in colder years. It's important to note that these species have more complicated migration strategies than the obligate inshore dwellers. It's possible that the cumulative effects of temperature in areas outside of Casco Bay are driving these differences in growth rates.

Sand lance are particularly weird; the species we are sampling is most likely to be *Ammodytes americanus*, which is typically found in coastal areas shallower than 20m and with sediment of an appropriate grain size for them to bury themselves in. Their level of seasonal migration is unknown, though they do go through a period of winter dormancy, where they bury themselves for extended periods of time. Casco Bay is not that far from the species' calculated center of gravity (about 270km north, at the northern tip of Nova Scotia). There is very little difference in growth rates we have calculated between cold and hot periods. It's possible that the population dynamics and growth of sand lance are currently being impacted more by habitat availability, food availability, and predation than by the temperature shifts we've seen.