some breadcrumbs on redwing + brms + “resims” for when you come back to this…

# resims

Resims generally refers to the approach of

1. Constructing probability distributions for the ISD using gmms
   1. Fit on multiple estimates of the ISD, to help smooth out sampling error but maintain relative abundances of different species.
   2. Currently I am using **5** estimates for the isd. This is because fitting gmms on large numbers of estimates is nonlinearly computationally/time intensive. For what it’s worth, I don’t actually believe there’s very **much** error introduced due to sampling from normals; the GMMs smooth it out and come out very close to each other even if you just use 1 estimate. 5 has been tractable.
2. Drawing individuals **for each year** from the timeperiod-level ISD probability distributions
   1. This gives us annual estimates of total biomass & energy use, assuming there is one ISD for each time period but taking into account interannual/intra-timeperiod fluctuations in abundance
   2. These estimates do seem to consistently differ a little bit from just drawing individuals from normal distributions directly. Seem to slightly underestimate relative to the direct draws. I’ve tried various tricks to get it closer (increasing hte resolution of the density functions, including more estimates, log/not log transforming) without success. So just be aware of that.
3. Doing this under 2 scenarios:
   1. Use the begin-ISD-density for the begin and the end-ISD-density for the end. This one I generally call the “actual”. It incorporates changes in total biomass/energy use due to the combination of changes in the number of individuals, and changes in the size structure between the two time period.
   2. Use the begin-ISD-density for the begin **and** for the end. This one I generally call the “sim”. This one simulates, if the begin ISD had been conserved; or alternately, the estimates for biomass and energy use we would expect for the end time period taking into account changes in abundance, and sampling error due to drawing from the probability function - but assuming *no* change in the ISD from the beginning.
4. Computing total biomass and total energy use for each year.
   1. This gives us annual estimates for total biomass and total energy under the above scenarios. When we put this in a model to compare the two time periods, this will help us account for the interannual variability in all these variables within time periods.
5. Repeating the drawing **of individuals** (i.e. starting from step 2) many times (currently using 100; 50 may be sufficient).
   1. Accounting for sampling variability in how those dynamics play out over time. You do get variability here such that one estimate can easily be spurious.
   2. There is no inherent relationship between the sim and actual dynamics within each sim. I mean, the sim iteration doesn’t matter; comparing actual from draw 1 to sim from draw 2 is equivalent to comparing actual from draw 1 to sim from draw 1.

Notes on running this process:

* GMMs are a compute limit
* You want to do this in parallel
* Currently I am not able to run parallel using my usual clustermq method so I am running things as multicore processes within one big job.
* A run with 5 ISD estimates, 100 sims, and all 530 routes took about half an hour to run as a multicore process with 20 cores of 2gb each.
  + This is fast enough that with 40 cores I can run 100 null models in about a day
  + I never successfully ran any of it (including overnight) in series using drake. It might be an impatience thing, but I kind of feel like it was also slowing down my computer in a nonlinear way.

# brm

I have then started using a gaussian linear model fit with brms to compare *observed* dynamics of change (incorporating change in the ISD as well as changes in abundance) to the dynamics expected given only changes in abundance.

Model formula: (total\_e or total\_b ~ (timeperiod \* source) / route)

I like using brm for this because it allows

* Running all the routes together in one model, rather than individually
* In theory I can incorporate year as a random factor, but I haven’t tried that yet because the compute is already slowing me down
* Using the draws from the posterior to get estimates of the begin/end sim/actual values and construct distributions comparing them
  + Helpful because we care about the route scale dynamics and “significant” differences in effects via some weird comparisons that are difficult to extract using existing packages or a frequentist framework
  + E.g. we care about if there is a significant interaction source:timeperiod *for each route*. More so than we care about if there is a significant interaction between source:timeperiod:route.
* Also, in theory the bayesian framework allows a little bit of separation from getting hung up on p-values and repeated comparisons. Rarely do I think this is actually going to matter.

I am using a gaussian brm because it runs faster and I am more comfortable messing with the estimates with an identity link situation. It makes me anxious to think about trying a different link-y situation, because scaling can be very tricky in this space.

The brm is slow on my computer so I am currently trying to run it on the hipergator.

As of 6:30 on Friday 10/8, it looks like the “short” - 50 / 100 sims - model with all routes *should* take about 2 hours to run with 4 cores. The current request is for 4 cores of 32gb each, because smaller requests (I was doubling each time, 4 > 8 > 16 > 32) all maxed out memory almost instantly. I won’t be shocked if I need to increase the number of iterations; current is the default so 4000. When I ran it with 25 routes x 50 sims locally, I got a message about that.

# Estimates from brm

I am then computing estimates for various quantities from the posterior distribution of the brm.

* Comparing sim beginning to sim end is a measure of the change expected due to changes in abundance. Because the isd has been maintained, this is going to be parallel to the trend in abundance - with some sampling error.
  + If these distributions overlap,
    - So if the distribution of (end\_estimate - beginning estimate) overlaps 0
  + “Not a significant change in abundance”
  + Declines/increases // declines/increases tracking changes in abundance
* Comparing actual beginning to actual end is a measure of the change in whichever currency you’re looking at (biomass or energy).
  + Again, if the distribution of (end\_estimate - beginning estimate) overlaps 0, not a significant change in whichever currency.
* Comparing the *change* begin-end between sims and actual is a measure of the additional change caused by changes in the size spectrum.
  + Whatever you do, they’re in the same units so a lot of comparisons are going to be reasonably intuitive.
  + I like using a scaled slope ((end - begin) / begin) here.
    - In this case - because the begin values for both sim and actual are going to be nearly equal to each other - I believe that any way you compare the ratio differences, you are essentially comparing the *slopes*
    - I like to standardize the slopes in a way that makes them more comparable across different routes that may have different actual values.
    - This metric is pretty intuitive. .1 = “end is beginning + 10% of beginning”. -.1 = “end is beginning - 10% of beginning”.
    - So I mean
      * ((end\_actual - begin\_actual) / begin\_actual) - ((end\_sim - begin\_sim) / begin\_sim)
      * Is proportional to (end\_actual - begin\_actual) - (end\_sim - begin\_sim) because begin sim nearly= begin actual
      * But if you have different begin actual (and others) across routes - like one starts at 1000 and one at 3000 - the *magnitude* of the slope will be more intuitive
      * Note that end\_actual - begin\_actual is the slope\_actual because the run is always the same across routes
  + Again, if the distributions of scaled slopes overlap each other, the currency-specific dynamics don’t deviate from that attributable to changes in abundance
* Still grappling with the best visualizations here.
  + I like some kind of raincloud-like plot to allow for both the estimates and the error around things, but might become unreadable with 500 points
  + Also a fan of a 1:1 plot with simchange(aka abundance forced change) on x and actual(akacurrency)change on y. Then you can plot the crossbars and visualize how wide they are, how many overlap 0, etc.
  + The issue with the 1:1 plot is it can be kind of hard to grasp how the densities are distributed
  + One option might be 3 currency-speicifc “rainclouds” (density of means on top, ~scatter of some kind of means + horizontal error bars below) and then the 1:1 plot
    - The rainclouds give you density and aggregate (continental-level trends), allow for showing how the different currencies are concentrated writ large
    - The 1:1 shows how they relate to each other at the route level
* And summary statistics I’m still not sure of
  + I mean, summary(estimated\_sim\_scaled\_slope) and mean(sim\_slopes\_different) is pretty informative as a start

# Null model

Unless I’m very surprised, I suspect the aggregate result from this is going to reveal too-common-to-be-random, but probably not-huge, deviations from N dynamics, particularly in biomass.

So the question for the null is no longer “is the ISD highly conserved” but “is the increase in mean body size detectably different from non-size-structured dynamics”.

Spoiler alert, I looked at this a little bit in the existing null model draws I have and it looks like….yes, but very weakly so (10% of routes relative to 5% expected > 95th percentile; stronger continental > regional > local). This might break down going into and out of the density GMMs though so check.

One question I’m not sure of is if it’s meaningful to look at whether the routes with a decoupling are the routes that pop as a statistically-significantish size increase relative to null model.

So I think I will want to look at these dynamics in the null model using the same density resampling method. I am not sure yet if it makes sense to run the brm on the null models. The brm might be computationally intensive and it might be sufficient to look at how the changes in the “true” - i.e. not model estimated - values compare between the null model and the real dynamics.

* So like for each route in each null, compute the ratio of mean body size begin:end and compare this distribution to the actual mean body size ratio begin:end for that route
* You might be OK averaging over all years at that point. Coarse but reasonable?
* But it might also seem important to get an estimate of the proportion of routes in which the brm detects a decoupling of biomass:abundance.