Meeting with Krkosek (08-15-18)

Admin:

* Get form signed
* Figure out the logistical format for project – am I meeting with him? A grad student? How often? As needed or regularly?
* Should I do email follow-up’s after meetings? Should I do weekly check-in progress reports?

My updates:

* Been reading lots of papers, I think I have a decent handle on the ecology of the system
* A few questions:
  + Koch’s postulates: has that actually been shown in this system at all? Is it a good way to go about determining causality of one agent or another?
  + Can they infect as nauplii or just once they reach copepodid stage?
  + Are the smaller salmonids (i.e. pink salmon) show higher mortality than other species of bigger salmon?

\*marty emailing me papers rn

Notes here:

Some management changes were effective, and corresponded with signs of recovery, components of wild fish dynamics

Lots of fishery data from DFO and stock assessment data

One paper uses atlanti salmon data,

One of the main issues is that it’s controversial – remains controversial because the scale is effects on individual fish – lethal or sublethal effects on fish, and then taking that and making inference on a population level is hard, the epidemics are quite large, the fish naturally have very high mortality rates so if they’re going to die anyways, who cares if they get infected, but are the effects additive or synergistic – complicated analyses, all analyses are correlative in nature, and so ti remains hard to show definitive stuff, there’s a paper on atlantic salmon where they tag and release salmon and look at the differential return rates (the paper he sent is a meta analysis) – they found here an overall effect – differential survival confers a 35+% difference in survival, major correlate was baseline survival rate, if it was a bad year for salmon, the lice have huge effects, in good years it makes less of a difference

Predation and infection can be synergistic or compensatory depending on different factors

Pink and chum enter the ocean at much smaller sizes (don’t stay in freshwater for a year) – decline in pink salmon during years where epidemics have occurred, but the chum salmon don’t show a decline like the pink salmon – main predator of both is coho salmon smolts, there is an obvious pink salmon preference even when rare, because parasites are elevating the attack rate, the predation becomes more elevated on pink

Infection affects swimming speed, position within schools etc, and it definitely has an effect on attack rate

Dataset:

* Different species
* We think we see more of certain lice species on certain species
  + Even though they’re comigrating through the same environments
* First thing is to look at this quantitatively and do exploratory data analysis, comparative
  + Sockeye, pink and chum
* For some samples there will be correlates that have to do with stomach contents or potentially (followup) odolith analysis or coinfections (microparasites)
* First cut: look at differences among juveniles in ectoparasite communities
  + Comparative, field-based

Meeting with MK & Sean

* Hand lens in the boat, killing them and looking under a scope, and then under the scope and then just looking at the motiles
* Motiles are always consistent among the fish – so this can be accounted for across all species, but then in a subset, theres some juvenile ones
* Notes: summarizes the dataset
* pulled fish from collections that had a reasonable number of th three fish species we're interested
* at least five of each of the three species, (some years have more than others)
* D for site means discovery islands and then J means johnstone strait

Notes about dataset:

* each row is a fish, two columns (survey columns) aren't important
* siene ID is going to be important to go into the summary dataset
* GH and are the two copepidite for the two species and then an unkown one
* calimus life stage is grouped into younger and older (clemensi A is first two and B is second 2, lep A is first stage and B is second one)
* if I end up wanting to look at species differences between the two lice chalimus stages, and you could look at this for some of the fish (just the ones that have the fine scale lice protocol)
* this means looking primarily at pre-adult stages and then some you can look at the copepodite stages
* motile stages: bunch of columns cor caligus motiles and the lep motiles, male claigus, any preadult or adult caligus male, and then same with female, then a column with caligus motile that's preidentified but its not a gravid one - column I'll work with the most is the çlaigus motile that's not a gravid female, this is the sum of the previous four columns - column Q and R are not gravid and gravid female
* for leps, the data is the same for every single fish so that's self explanatory, two groupings of caligus and five for leps for all fish

(10/04/18)

Updates:

* Currently just looking at the data, starting to subset (with regards to site is what I’m paying attention to right now)
* Plotted a map of the sites to take a look at that, am going to make a few choropleth maps that will allow us to see if there are any interesting spatial patterns in terms of gradients or not
* Took a look at a couple plots that will be interesting once I get a bit further into it
* There’s pretty big differences between the different sites in terms of how many fish as well as lice per fish
* In terms of fish per lice, there seems to be a pretty normal distribution in terms of number of lice on a fish
* Notes:
  + Early collections weren’t really balanced between species but
  + You have to look at comparisons collections by collections (is one species more infected with one species of sea lice etc)
  + eventually:
    - species id as a response variable and to pool the data, we need hierarchical structure (mixed effects models) random effect (identity of sample collection) so that any variation among collections is represented through a random variable (random normal variable) and then variation among species that’s systematic is estimated at the within collection level
  + Collection: individual space time combination in the dataset – when the boat arrives at a particular site at a particular date (often a single employment of the net, sometimes a few)
    - So aggregate so a collection is all the fish from a particularl location on a particular date
    - If theres a lot of times when there’s multiple deployments of the net we might be able to pool data to that level but it’ll make it quite unbalanced (you’ll end up having a bunch of pinks in the first deployment and sockeye in the second) so treat it as an aggregation of fish collected on a particular day at a particular day
  + Write a loop that goes through the dataframe collection by collection and calculates summary stats:
    - Relative abundance of parasites by fish species (total number of paratsites on average, per fish, per species) – store this in an array
  + and then start looking at patterns in the collections in that array
  + think of a way to standardize the data by collection – so average number of parasites per fish per species is important, then also calculate the total number of parasites in the collection and then calculate the proportion of parasites on each of the species (that’ll be six quantities per collection to calculate) – therefore that’ll put the stats on the same scale for each collection – for the proportion quantities calculate a mean and a variance for that and see where we’re at from there

10/11/18

Updates:

* created collections based on unique time/location pairings – there are 52 collections in total
* then, according to last weeks instructions, I calculated the average number of parasites per fish per species, as well as the total number of parasites in that collection, as well as calculated the proportion of lice for each of the species
  + then I calculated the mean and variance for the proportions
  + then plotted the proportions and the mean number of lice per fish per species
* it appears that Sockeye are the worst affected, and this is controlled for (it’s not just because there’s more sockeye) because of the way it was calculated

Questions:

* Proposal
  + What is it you’re looking for? An intro, more of a methods proposal?

Partition lice for leps and caligus and do the same meanlice per fish plot (exclude the ones that don’t have species ids for the parasites)

For proposal, put in some proposed methods, and also have an introduction showing that I know what I’m talking about and places the project within the literature (specific literature on salmon and sealice and broader literature on host parasite communities and host specificity and generalist vs specialist parasites etc.) and so talk about the biological background for both the parasite and host species – then there’s conceptual stuff about parasite life histories and tradeoffs between specialists vs generalists.

Generalized linear mixed effects model – negative binomial distribution likely (count data is the parasites per fish) and count data (which can be low integer values), simplest distribution is poisson, but probably negative binomial – parasite individuals will aggregate on individual fish – more fish who have zero just by chance alone 0 could use a zero inflated poisson distribution, and negative binomial, and zero-inflated negative binomial. Random effects for collection (normal distribution) and that will allow us to account for variation among samples, which we’re not very interested in, and account for variation among parasite or host species which will be fixed effects – might fit one model for caligus, one for leps, and if there’s differene in host specificity, then the fixed effects in each of the models for the host species will differ. So that’s what we’d be looking at, we could even have parasite abundance as response variable, NOT carved out by the species it is, fixed effects for the host (general specificity) and then have general interaction term between fish species id and parasite species id, which would be a more direct test of whether parasites have host specificity

10/18/18

Updates:

* Created the new graphs
  + Essentially just repeated the same process but for each species of louse
* Working on proposal currently, presenting some of my ideas for it at lab meeting on the 29th

Model will look similar, but since the distribution won’t be gaussian, but function will be glmer(), and you can specify what the error distribution was, and if it was a binary response variable, you choose a binomial distribution, and you use this to specify the family. The syntax is the same.

As a first step, fit some models using poisson family in glmer, and that’s gonna be good practice, but it’s not actually correct because the data are probably overdispersed and the poisson doesn’t account for this (overdispersed means more zeros than you would have by chanve and way more parasites than you would have just by chance).

As step 2, in order to deal with overdispersion, so we’re going to use a different package which is called ‘glmmadmb’ (it might not be available yet as a download in r, but you can still download and install manually in R) – marty to send a link – info about the package will be there, as well as how to install. Jags and Rjags are also really good for fitting these types of models

When you use the family, put in poisson and zeroInlfation = FALSE to start with

Start with analyzing the caligus and leps separately and we’ll wanna have essentially just mean abundance by host species (whatever the header label is for fish species)

All.cal ~ spp -1 + (1|Collection)

For glmer …. , family = ‘poisson’

For glmmadmb …., family = ‘nbinom’, zeroinflation = FALSE

The minus 1, if you don’t have it, it takes the first level in the categorical variable and treats it as an intercept, a then your estimates for the other levels of the categorical variable are additive, so the minus 1 estimates it directly for every level of the categorical variable

For the output, we want to see what we’re doing, and then we might see the results as being simple or not. Interested in coefficient for species level effects

Make a plot with parameter estimates and two standard errors – for two species of the parasite (two sd is 95% confidence interval) and you can see if they overlap the point estimates and that’ll make it really easy to see if there’s differences – for the plot, take the estimates, exponentiate them, along with the two standard errors

Take the estimate values, plus or minus two standard errors and raise each of them e^(x) and that’ll be our estimate of parasite abundance per species per fish

Yaxis = average abundance, on the x, have pink, chum, and sockeye (whatever order) and for each of these, there’s going to be an estimate (two points, one symbol for cal, one for leps) and

Could potentially in the future look at body size covariates, etc.

Tell marty what other interesting stuff is in the data, i.e. is some of the variation also due to environmental and body size data – so see if there’s environmental data, sean might have body size data

So we need to find the estimate of average abundance of parasite sp per fish species

10/25/18

Not a ton to go over today – showed Marty results and stuff yesterday

* Put some plots and model outputs in a document and sent it to Sean and Dylan
* Have to wait for feedback from Sean before anything else really moves forwards

Spend some time reading up on the field-work methods that have gone into generating the dataset – Look at the Hakai institute website (Salmon Program) locations, frequency of sampling, how sampling is done etec.

For lab meeting: include the collection stuff as a part of the methods, the dataset, part of the program etc. – Sean has a video of them doing the purseineing – and then the preliminary results

11/01/18

* Not a ton of updates
  + Speaking with sean tomorrow
  + Working on getting the environmental stuff all shifted over, but still not quite there yet

One dataset has measurements for fish length stuff – some are measurements from lab some from field, some fish have both, and they’re often not the same number (dead fish who go through the freeze thaw, they change a bit) – for each fish that has the body size info, if it has both, take an average, and there will be a handful of fish that have no body size data, and those will be excluded from the next set of models

When I put together a data frame, because it will be two columns that I’ll have to put it into one.

Pull out the year as well as the Julian day and make those two separate columns in the dataset (try and convert julian day by a function) – this will allow us to look at using the year as a fixed effect in the model

Another thing, would be to run the models again with just the copepodite data

The infectious stage of caligus has much higher preference for sockeye, the leps can respond to chemical cues and respond preferentially to the hosts they want

Alternative explanation: initial infection is random among the fish stages, and the ones in the older life stages would actually be choosing.

There could be a lot of movement between fish in the later stages of the lice life stages – the parasites might sort preferentially on one fish species vs the other, but this requires an ability to do this.

So fix it so that column is only pre-adult and adult – and if there’s NAs in the adult and preadult stages should be removed from that group

Redo plots first, then work on merging the data, then run the models using year and body size as fixed effects

11/02/18 Skype with Sean

* Questions about data:
  + What is the cal.mot.unid.notgf and cal.mot.notgf columns – what do they mean, which ones should I be using, there’s some rows that have neither of these columns, so I can’t use any fish that have NAs there?
    - Use the sum column and just fix the sum
  + Those rows coincide with lep.mot.unid.notgf being NA, so I guess these are all ones I need to calculate separately or no?
    - So this one is not a sum column and I can treat the NAs as zeros
* Questions about models:
  + So our next step is to add in some covariates obviously and see where that gets us, definitely will use the year as well
  + Another thing was to rerun the analyses but exclude fish with no infection
    - Doing this could tell you about looking within one lice species and allow you to control for time of exposure
    - To do this properly you could do a gamma hurdle model but doesn’t work for my data
  + Do you think there’s a way to use this data to actually make some conclusions about what is going on in terms of the abundance of these different species? – since the population abundance is probably having the most effect on what lice species are most abundant
    - This is probably a no go – there’s so many other reservoir populations for the salmon – ask Dylan about this further
  + Anything else that you really hope to see from this data?
  + For more covariates
    - You want to create a model set – each model is essentially a hypothesis, you fit them all and then you use your AIC to see what is the best way to go forwards
    - Sit down with marty, think about the models you would make and then do that,
    - There are some things you have to include: year HAS to be in every single model to prevent pseudo-replication, and then site as well, then collection itself separately – every model will have a random effects structure with site and collection number to account for psuedoreplication and then have a fixed effect for year (you want to minimize the number of parameters that you use as fixed effects because you get more power) – you can’t have a random effect with less than six levels, and it assumes that you can fit one parameter to the normal distribution, and the random fits one effect based on the standard deviation)
      * Every model will have crossed (or nested) random effects structure with site and collection – the situation I have is either crossed or nested which is site doesn’t have every collection number
    - You’ll have salmon species, fish length, temp and salinity (temperature should probably be in every one), every other variable will be with and without, and then region (i.e. Johnstone strait vs. discovery islands, with and without)
      * The idea with the above is you have all the above in one massive model, you start combining the effects in subsets with combinations of different things – there are some things that will go together and some that will not go together

Generally:

* The real idea is exactly what I think it is
* General questions:
  + What do you think are the most important skills that helped you from a quantitative perspective during graduate school? Stats, programming, etc.
    - R – modern stats – mixed effects models and model selection – learn how to do GIS stuff in R -
  + What would be the biggest thing you would advise in terms of preparing for grad school and trying to get ahead of things – reading a lot, learning a new language, etc.
  + Do you have any tips or thoughts on developing a good relationship with your supervisor?

11/15/18

Aic table with null model with no fixed effects and then all the combinations of species, year, and body length (including ones where there’s just each one of those) – so there should be six or seven models and there should be the AIC values and then with a relative look at the lowest AIC value

If there’s one that’s clearly better than the others then we can just go with that one, but if there’s uncertainty, we can calculate the model-averaged estimates, so you calculate a wighted sum of each of them and the weights are got by the AIC sum

Also add for each sample collection, I should be able to (from the metadata) find whether the collection is from the discovery islands or johnstone strait and then add that in as a fixed effect

Set up the AIC table as the increasing complexity of the model in descending order

11/22/18

Okay so now we’re going to do the whole thing again without the fork length – redo the model selection tables with the smaller set, fix the delta AICs so its relative to the smallest AIC, for the model with the lowest AICs do a table with the summary and then do a table with the reverse transformed values as well (just put the estimate and the two standard errors right beside the estimate in brackets to kind of show the whole thing).

11/29/18

So moving forwards – part of it is to come up with some figures to graphically show what’s going on – and then fit an additional model that has the interaction between years and species (make sure year is categorical) – for chum pink and sockeye, those are the estimates for the discovery islands, in 2015 – and then the estimates in the subsequent years are additive, same with Johnstone straight, so you can just add the coefficients together before you do any reverse transformation – also run a model that region\*species and region\*year, and then a full one species\*year + species\*region + year\*region – make sure that year is a factor, check online to see if the others are all factors as well

12/07/18

Create two plots, one for cals one for leps, chum pink and sockeye and then facet (group) by the year – everything is just a sum (prior to reverse transformation), for the standard deviations figure out how to plot these (is it an average etc.) “adding standard errors of coefficients generated in a generalized linear model” – potentially look in Ben Bolker’s book as well or the R book – for caligus break it into the two regions, and on the x axis, you’d have chum pink sockeye grouped within year, and then for each one, have two mean and whiskers for each region – look up the rule for adding three standard errors, the coeffiicients are still just additive

01/16/19

Break up methods into two sections: one that is fieldwork and how the data were generated (by analyzing the fish) and then a description of the statistical analysis

Start new doc with just methods and results and really dig in on that

For the map, have a double map, the main part of the map should be a zoom in of the study area, and that’s where the sites should be identified, also have arrows to the two clusters of sites that indicates Johnstone Strait and Discovery Islands

Make the plots all greyscale – pick shades of grey that are more easily distinguishable

For figure 3, standardize the y-axis and eliminate the legend on the lefthand side, and move the legend to the white space so I can spread out the x axis for both – put text in the white space on each one, put the species name of the lice

For the big plot, try and put everything on one plot and see what that looks like – so have the leps and the cals on one plot

Change the tense of everything to past tense not future/present

Don’t need to include a formula for the model description, words is okay

When you refer to the parameters that you’re estimating (species, year, region etc.), the ones that are categorical are referred to as ‘fixed effects’ so describe the FULL set of models, including the fork length ones, even though we didn’t end up using those – fork length is a continuous variable that we used as a random (?) effect

For intro, have a paragraph preceeding the start that I have here that lays out more the conceptual context for parasitism and its role in marine ecosystems and fish recruitment. Start with some of those ideas and citations that aren’t just pacific salmon, and how parasite species that have more than one host lead to multi-host disease systems where you have parent competition, reservoir hosts even when the host species is small in abundance (leading to extinction), reservoir hosts are either domestic or other wild fish that are sympatric at least for parts of the life history and are abundant and provide an anlternative host population for the parasite to be maintained and spread to wild salmon. Because the parasites here are marine only, the abundances observed in the data are produced by other fish species in the environment, it’s not a maintenance in the juvenile salmon themselves – not enough time has elapsed for the population of lice to actually establish a full population on the salmon, so the primary transmission is from atlantic salmon for l. salmonis, because there aren’t any other salmon species around in the wild that time of year with that level of abudnace. For caligus, the farm salmon are important, but what might be important is herring – so overall, apparent competition, maintenance of diseases in hosts, etc. etc. ---- all this is like two more paragraphs, narrow it down as: broad theory, the theory in aquatic systems, in fish, then in salmon

GET TO MARTY BY SUNDAY, have the error propagation thing ready for marty next week

01/23/19

Show marty the confidence interval thing from Ben Bolker’s book – pg. 410, last paragraph

Marty sending me a thing rn to take a look at.

So we’re going to take out the region component from the models – so we’re going to fit some more models to a bigger dataset from sean, we’ll have a much more informative model for region based on this other way

We’ll have to control for the week within the new dataset as a factor level random effect, so samples that are collected in the same week would share the same random effect

The model is too complicated for biological interpretation – that’s why we got rid of the crossed effects ones

So send marty an updated version of the doc (with the methods cited using the document I found), and with the region removed from the current set of models, and get the AIC table down to only the ones we actually used, don’t need to put the crossed effects in there

02/06/19

Change table 4 and put in proper reverse transformed estimates

Try with year as a fixed effect as well as a random effect (year needs to be factor both times here)

* This might help with convergence issue

Do the year thing fish species by fish species as well as just by the lice species

Do the plot the same way as last time with all the variables in there (also separate out the plots for leps and cal)

02/13/19

We can also plot mean and distribution of the data themselves – use bootstrapping. The idea is that you do a simulation where you sample with replacement from the database. Ex if you’re doing caligus on sockeye, you first take a subset of the data that reduces it to the year and the region and species combination that you’re looking at (ex. 2017, sockeye, caligus, johnstone strait), and then you run a simulation where you sample the reduced dataset with replacement, calculate the average number of caligus on sockeye in one collection, save that value, and then repeat. And because it’s got a hierarchical structure, you want to do your random sampling at both levels (first take a random sample from the collection)-suppose you do it 1000 times, each time, grab a collection, see if it has your target fish in it, if not repeat until it does it, then if you have your collection, it might have 8 fish in it, so sample with replacement, 8 times. Calculate the mean, then store that. Then repeat that process 1000 times. That’lll be 1000 calculations of the mean, where the fish is randomly sampled. That gives you a distribution of means, and then you can look at that distribution, and calculate your estimate fo the mean, which will be the median of that distribution, and then for the confidence intervals, you get the 2.5% quantil and 97.5% quantile. And if you have 1000 estimates, order them, and then find you 25th, your 500th, and your 975th And do this for 10,000 replicates, not 1000.

03/15/19

* ditch site region in the first set of models
* two datasets, two sets of models, and then for each error distribution for the region level one, first column element being error distribution, then what the fixed effects are, then aic and delta aic (keep them limited to within the error distribution)
* So I’ll end up with two tables for the first, smaller, dataset, with just the leps and caligus models
* For the second set, I’ll have 6 models for each of the 4 versions of the model (null, 1 fixed, other fixed, both fixed), so for each species combination (i.e. leps and chum), I’ll have 4 versions of the model in each of the three error distributions, which is three tables for each species combination and six species combinations so 24 tables for that one
* ASK ABOUT LAB MEETING DATES

03/27/19

* Collapse table one and two and have one more column that has the species of lice
* Combine table 3 and 4
* Combine tables 5-10
  + Organize by parasite species and then nest fish species and that would add two columns
* Order them in the same sequence by model every time, and don’t worry about ordering by delta AIC
* Do the same thing with table 6, add two new columns for lice and salmon species
* Double check that I’ve ordered Johnstone Strait and Discovery Island correctly in the plot
* Separate out the leps and the cal for the bootstraps

04/03/19

* Poster all good?
* Intro – mostly look at the first 2 paragraphs, tell me if I’m on the right track
* Discussion – main points okay? Am I going in the right direction?
* What is the ‘take-home’ that you want these results to tell?
* Other questions: I feel like I get very much caught up in what I should be spending my time on, and it never seems to be a good option – I feel like I need to get domain knowledge (obviously), but also that I need to learn new methods and new software etc.
  + How much effort would you say should go into things like getting a better grounding in math, versus getting much more domain knowledge, versus spending time learning to code in new languages etc etc?
  + Essentially, how do you tackle the seemingly impossible problem of how the hell do I acquire all this knowledge it seems I need to know, and how much of it do I REALLY need to know? Because I’m kind of unclear on the whole shebang

Notes taken on the manuscript itself today