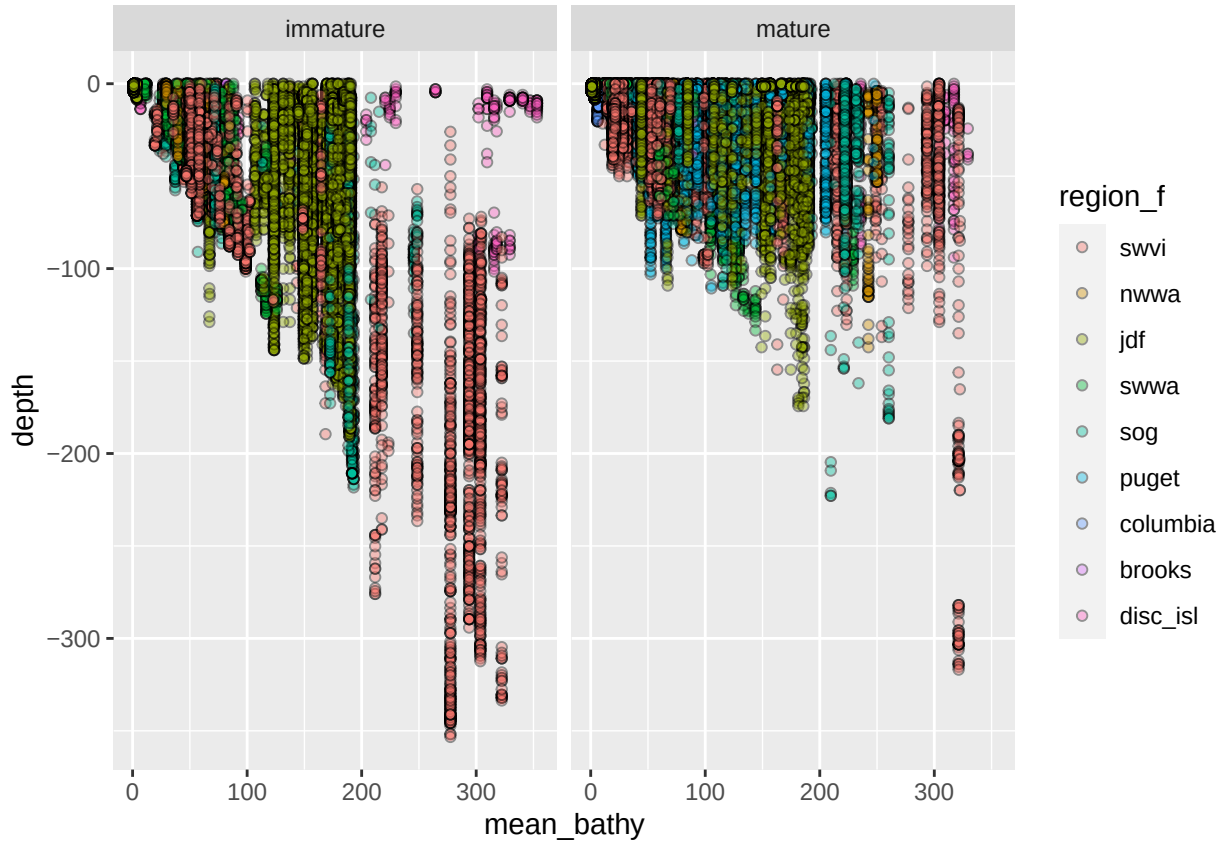
rf_model_summary

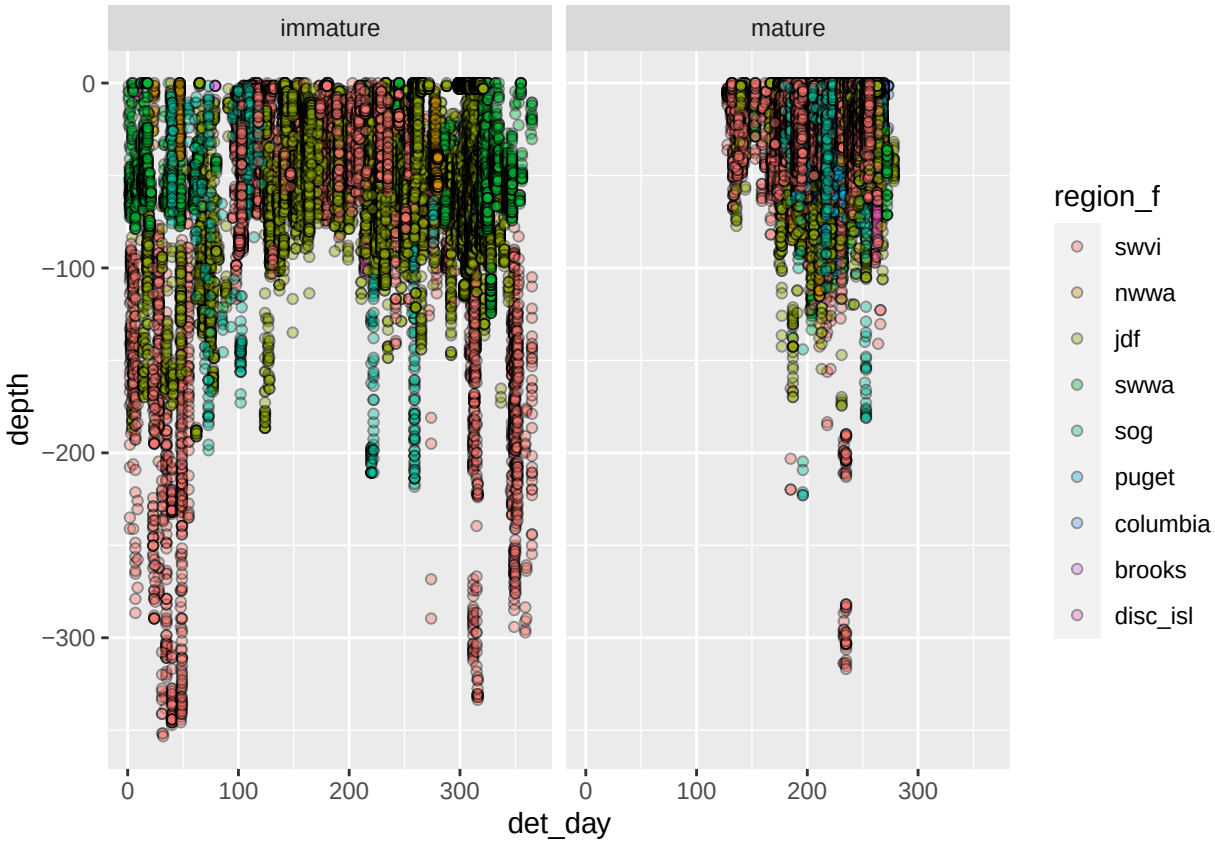
Background

Sensor data from V13P tags provide an opportunity to identify environmental variables that regulate Chinook salmon depth distributions. Intuitively depth distributions could be shaped by a range of physiological (condition, maturity stage), spatial (bathymetry, distance from shore, lat/long), temporal (diurnal and seasonal cycles), and dynamic environmental variables (temperature, salinity, prey availability). Importantly these covariates can interact with each other in counterintuitive ways.

Indeed the raw data show some relationships with each of these processes. As an example, there is (unsurprisingly) a relationship between bottom bathymetry and depth, which appears to vary among regions.



There is also a strong seasonal and spatial signal in depth distribution. Note that seasonal cycles vary among mature and immature fish. Immature fish were assumed to be mature at time of tagging based on size, but reared at sea for an additional year.



Modeling Approach

The downside of high resolution individual data and a large number of covariates is that complex models are often necessary to account for the data's underlying structure. In this case, there are a large number of repeated measurements resulting in substantial temporal autocorrelation. Accounting for this autocorrelation is particularly difficult because the lag between detections varies from seconds to weeks. The response, depth, is also not normally distributed (bounded by the surface and bottom bathymetry). Typically these issues are addressed with a combination of link functions appropriate for non-normal data, transforming the response (e.g. quantifying differences using binned or proportional depth) or binning (e.g. taking the mean within an hour interval), however there are downsides to these approaches. Additionally a spatially explicit model is preferred to account for spatial autocorrelation and to estimate how Chinook may be behaving in locations within the study area, but without receivers.

Since my initial attempts to fit traditional models were not successful, Sean suggested using machine learning techniques which have fewer distributional assumptions and are robust to interactions among covariates, as well as autocorrelation. Some preliminary analyses indicated a random forest model fit to untransformed depth data was better supported than other models (e.g. gradient boosting machines) or models fit to transformed data.

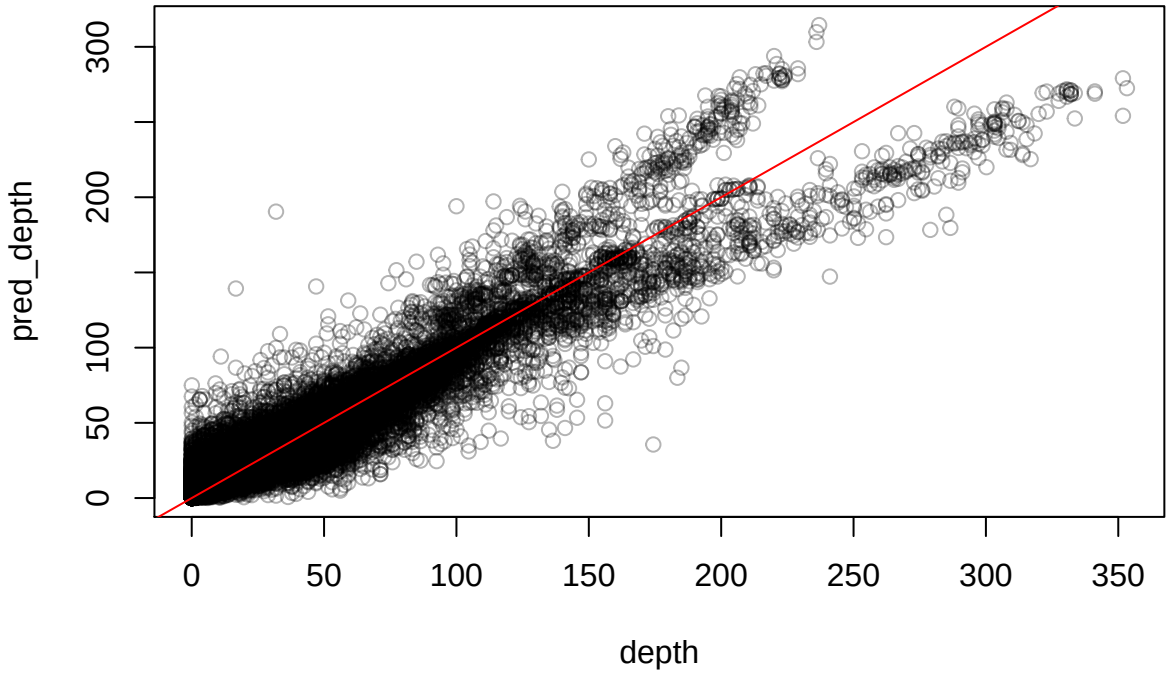
Briefly random forests generate a series of regression trees and average their predictions. Each tree begins with all the observations and partitions the data by splitting along a single covariate, choosing the covariate and node that minimizes the sums of squares at each each node. To increase stability, correlations among trees are reduced by using a bootstrap sample of the original data for each tree and at each split, randomly selecting a subset of covariates. The major downside of these models is their interpretability, but given the nature of our data I think they're the best option.

I've included the following covariates, but am open to others if you feel I'm missing something.

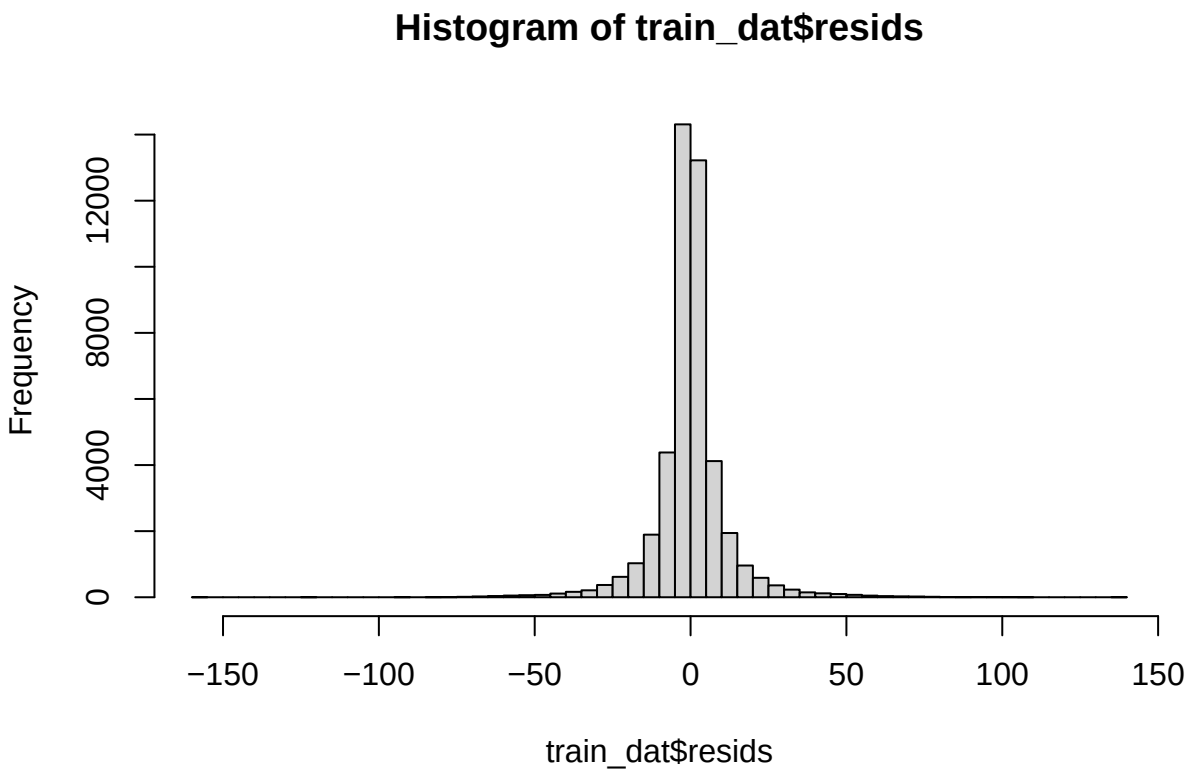
- Life stage–mature (fish will spawn that year) vs. immature (fish remains at sea)
- Hour–diurnal cycles
- Year day–seasonal cycles
- Bottom depth–mean within an 800 meter radius of the receiver (approximate detection range)
- Bathymetric slope–mean as above
- Distance to shore
- UTM coordinates–spatial variation that is not accounted for by bathymetric variables or distance to shore
- u-momentum–proxy for horizontal current strength (ROMS output)
- v-momentum–proxy for horizontal current strength (ROMS)
- w-momentum–proxy for vertical current strength (ROMS)
- Temperature (ROMS)

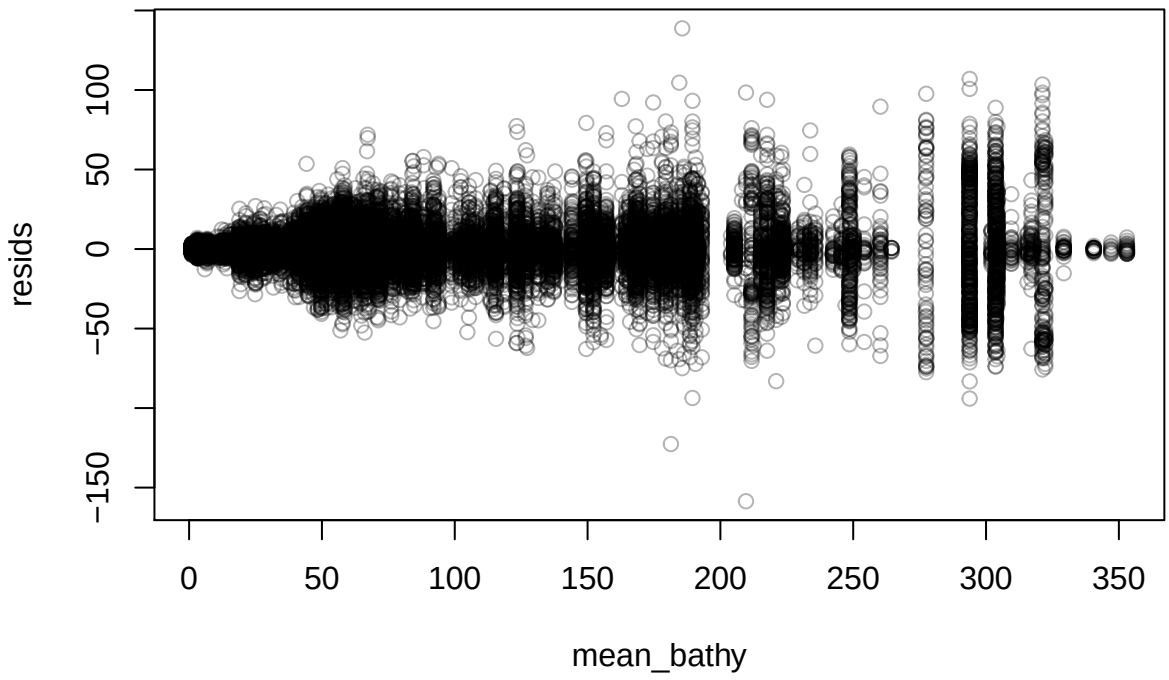
The ROMS derived variables are outputs assuming 25 m depth at the location of the receiver. I chose 25 m because this was the approximate "average" depth in the Chinook observations and because most variables were correlated at 5, 25, and 75 m depths, but we may want to use sea surface temperature instead.

Generally random forest models are evaluated relative to testing data, which are not used to fit the models. For simplicity's sake I'm showing performance relative to the training data. Ultimately the plan is to use the upcoming year's detections as our out-of-sample testing dataset to evaluate model performance.

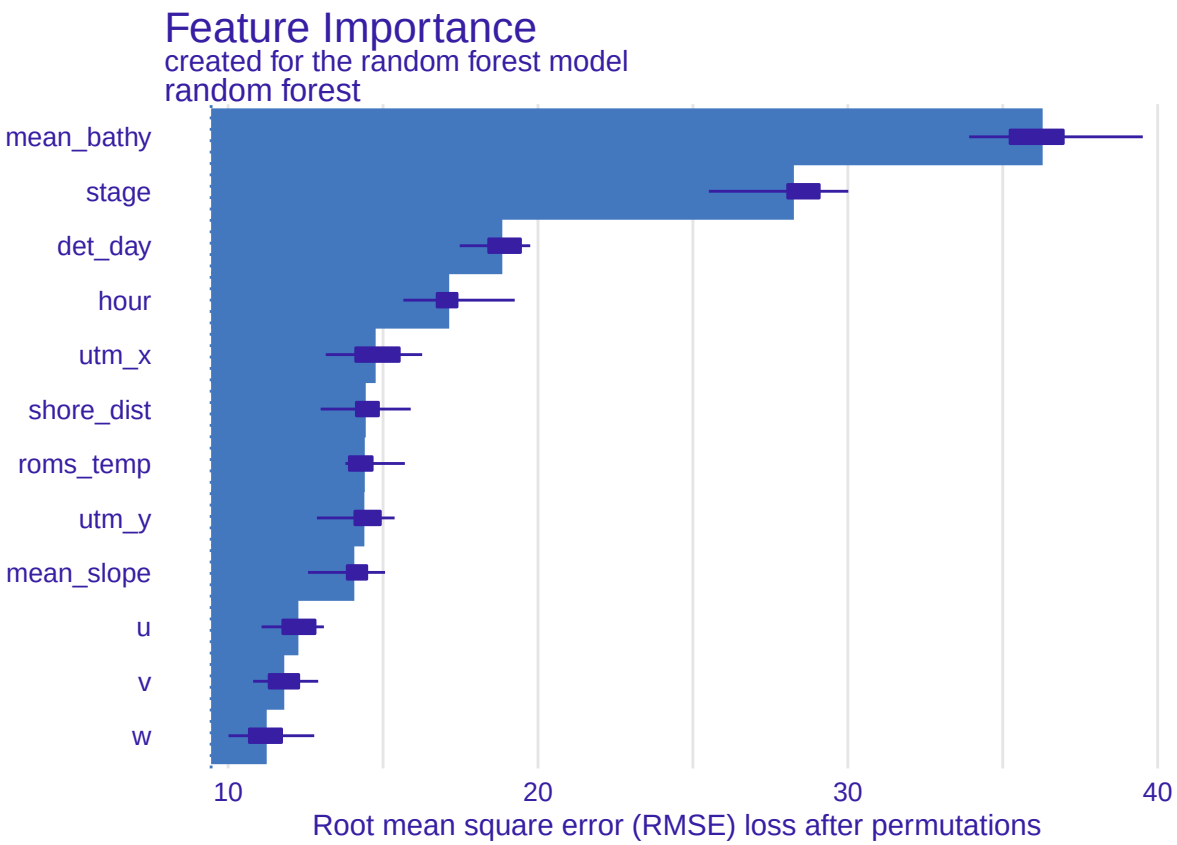


Residuals are reasonably distributed, but do show a relationship with depth. However I don't think this is an issue given the modeling framework.

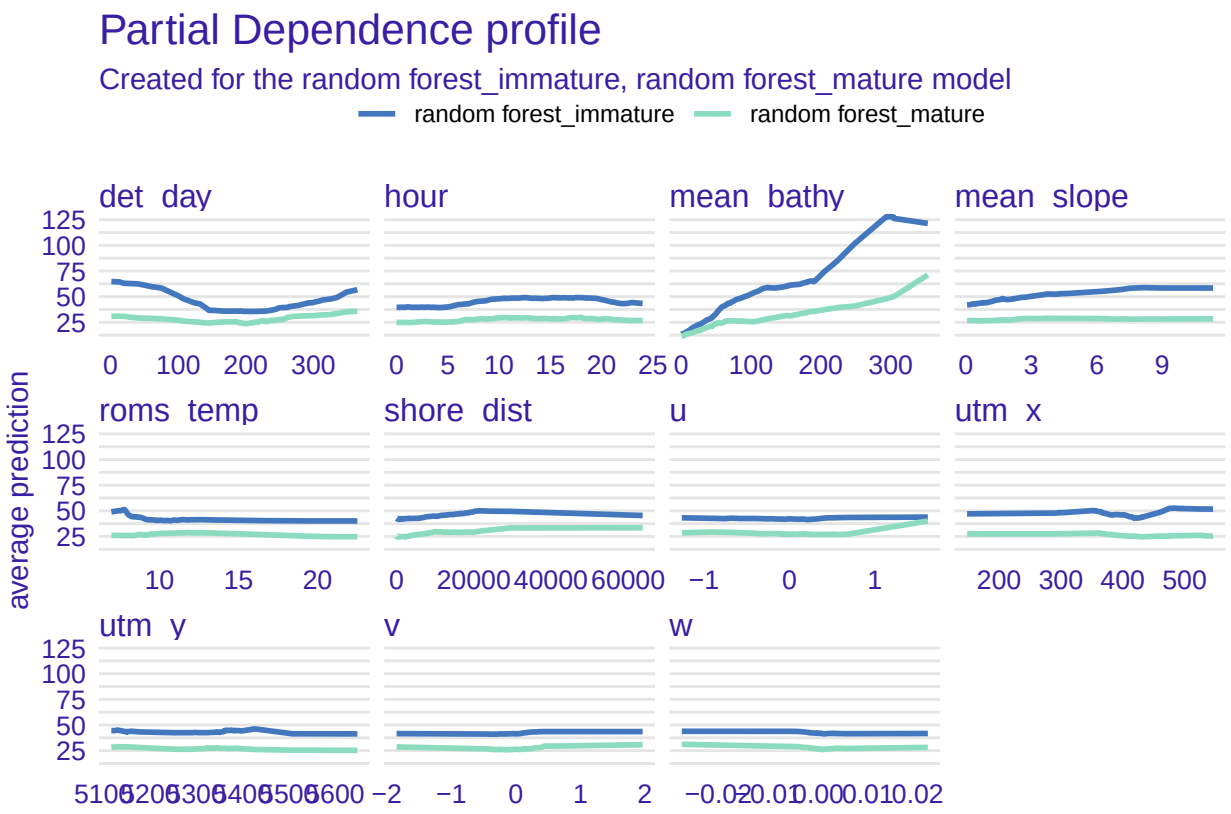

Histogram of train_dat$resids
Frequency
0
4000
8000
12000
−150
−100
−50
0
50
100
150
train_dat$resids



The following figure estimates each covariate's relative importance by permutation (a single covariate is permuted, breaking its relationship with depth, and the change in RMSE is calculated). More important variables will result in a greater increase in RMSE after permutation and the whiskers represent variation among different permutations. Here bathymetry and life stage (mature v. immature) have the greatest explanatory power, but seasonal (det_day) and diurnal (hour) effects are also present. The remaining spatial and ROMS variables have more modest influence on mean depth.



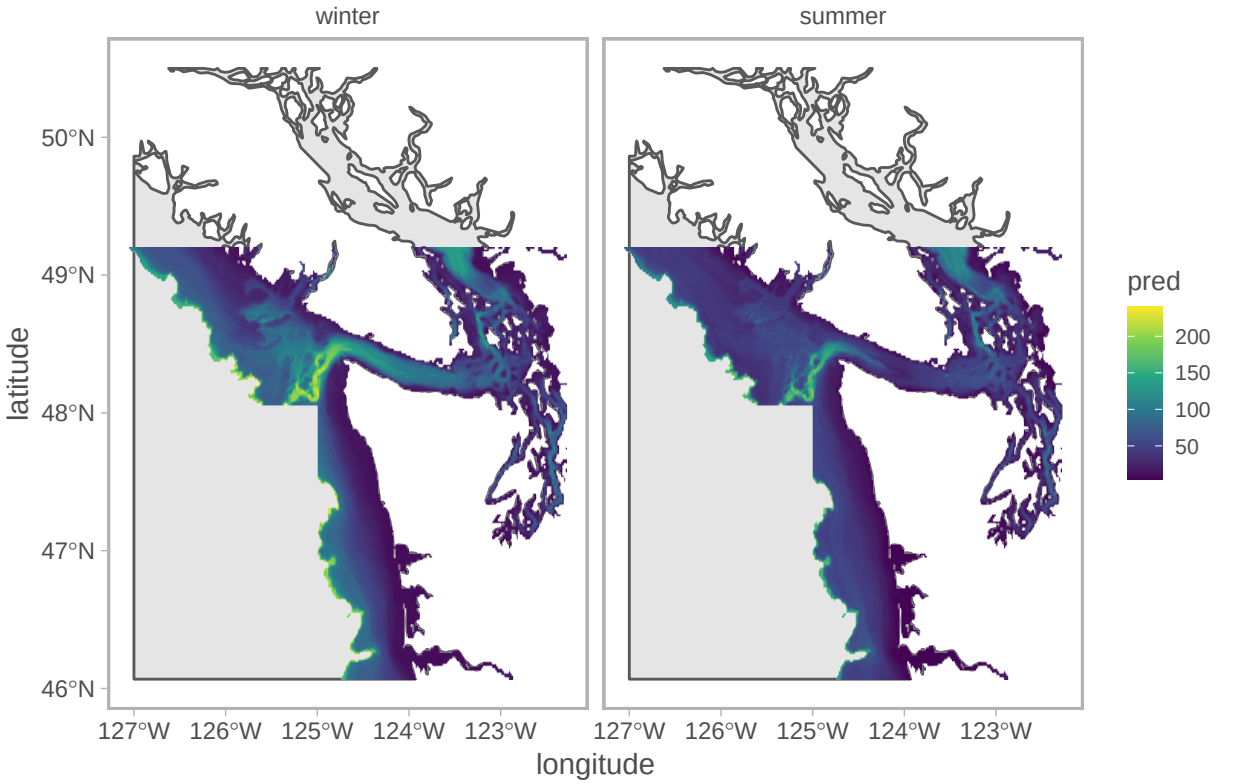
Partial dependence profiles provide an estimate of how the response changes across different values of a covariate, holding all other covariates at a specific value (here there mean). Seasonal and bathymetry effects are particularly pronounced. Generally immature fish have deeper distributions than mature fish, but there is evidence for interactive effects (e.g. bathymetry has a stronger effect for immature than mature fish).



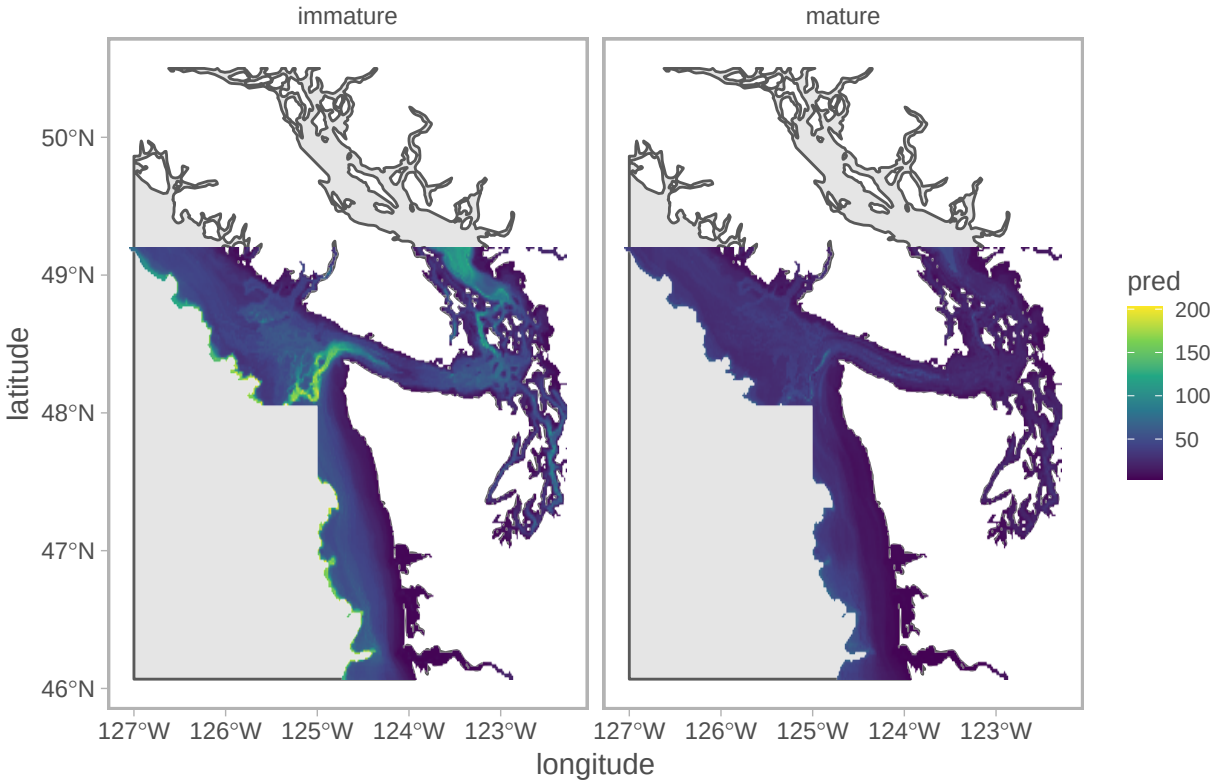
While the partial dependence profiles provide information on how individual covariates influence mean depth, many of the variables are spatial and naturally interact with one another (e.g. a location will have a single value for bathymetric depth, slope, UTM coordinates, and shore distance). We can use a grid with each of these values to visualize how predicted depth varies across space while holding temporally dynamic variables at reference values.

Each map is constrained to the core study area (i.e. continental shelf and southern areas with majority of receiver coverage). Warmer colors always represent deeper depths, but the scale is relative and varies from figure to figure.

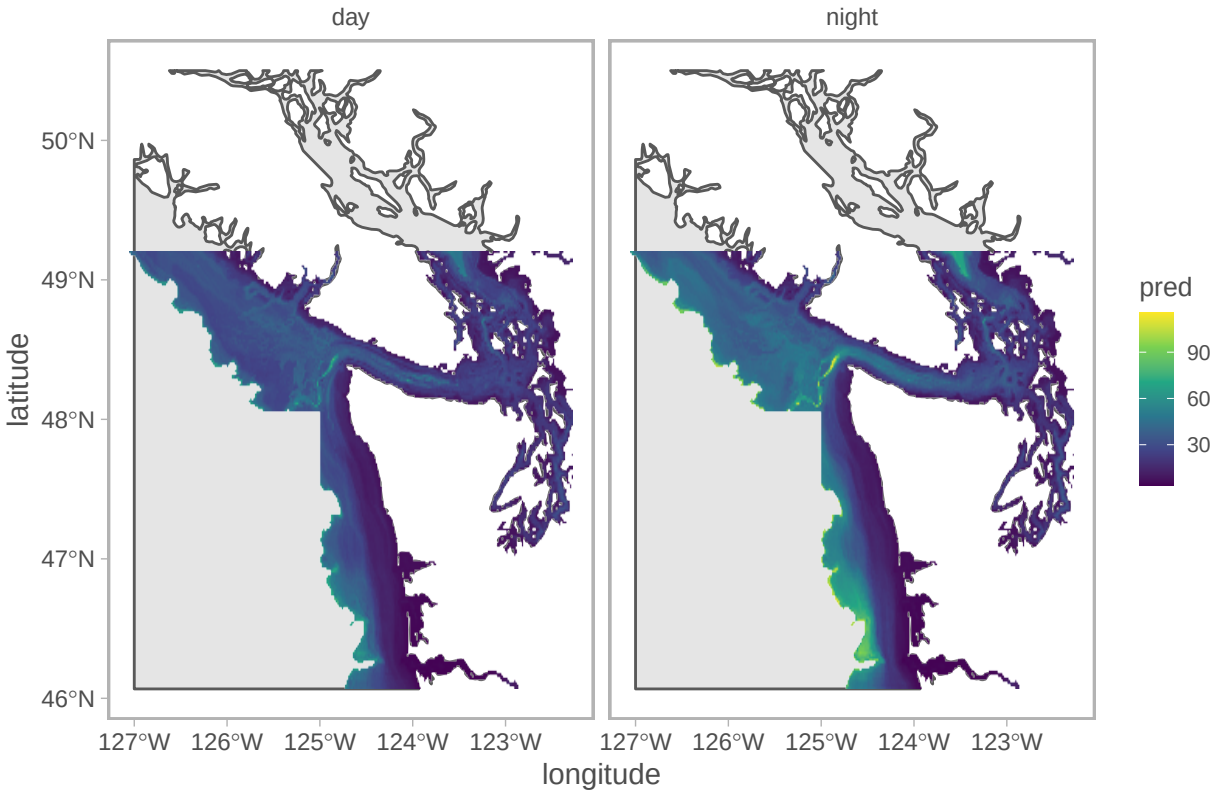
First are seasonal patterns. These are focused only on immature fish, since winter observations are unavailable for mature fish, and are daytime predictions. Note that maximum depth increases slightly in winter, but the major change seems to be more widespread deep observations in winter months.



Immature fish have deeper distributions during summer/daylight and are responding more strongly to bathymetric features than mature fish.



Counterintuitively, there is evidence for slightly deeper distributions at night. Note that because there's a spatial component, it may be more appropriate to visualize the difference between day/night predictions in a given location.



While I think these figures characterize the main take homes from the depth data, this analysis is still preliminary. In particular, I need to generate estimates of uncertainty for the partial dependency profiles via bootstrapping. As previously mentioned I'd also like to use 2022 detections as an out-of-sample test for the model to evaluate its performance with new data. I also need better bathymetric data for American waters. If anyone has digital elevation maps for the shelf off Washington please let me know. If folks have any suggestions on how it should be tweaked (e.g. alternative covariates), please let me know.