Preliminary Results

Adelle Molina

Boosted Regression Trees for herring recruitment

Code for getting data, generating time series of possible predictors, plotting recruitment and predictor time series, producing correlation matrices to reduce the possible predictor list, and running boosted regression trees (preliminary results).

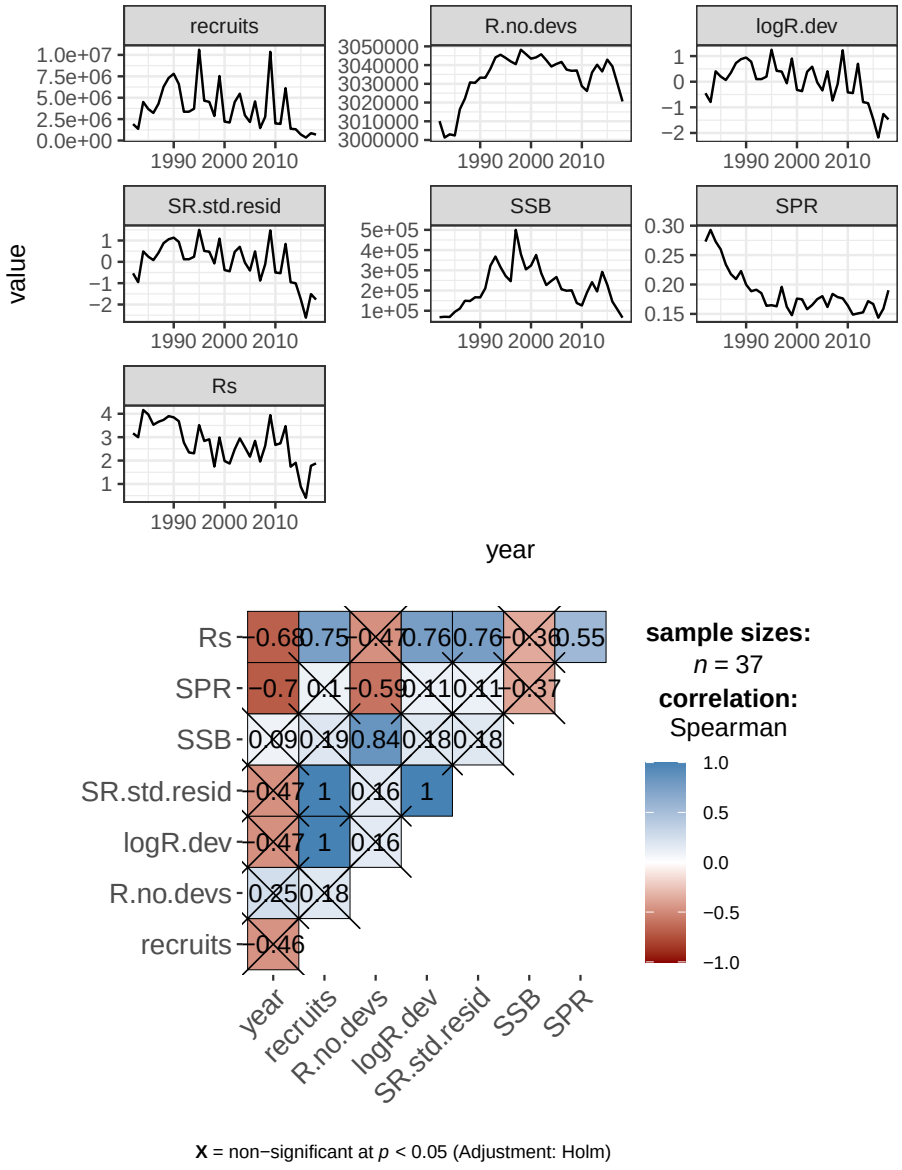
Load data

The combined dataset being loaded in here came from various sources (mostly NOAA ecodata github) and were modified, summarized, compiled and exported in a separate script.

- SST & BT raw data came from the NMFS survey and were clipped to EPU regions
  - Annual survey derived SST and BT were area weighted
- Annual and/or regional environmental indices came from ecodata
  1. Heatwave index (missing some years)
  2. Cold Pool index
  3. Gulf Stream Index
- Various zooplankton indices include abundance, density, and calanus copepod abundance
- Recruitment indices came from 2022 ASAP model
- Other indices, such as salinity and chlorophyll might require a separate analysis b/c time series are shorter, but they are included in preliminary runs here
- Haddock index is the retrospective adjusted SSB from the GoM Haddock 2019 assessment

Preliminary plots

Plot time series of possible variables and correlation matrices.



Trends in recruitment variables:

Recruits: Large fluctuations, no strong year classes since 2012 and lowest numbers on record

Recruitment no deviations: increased rapidly early in the time series and has remained high throughout (this is based on the stock recruit function, right?)

Log Recruitment Deviations: Lots of interannual variability, but was gen-

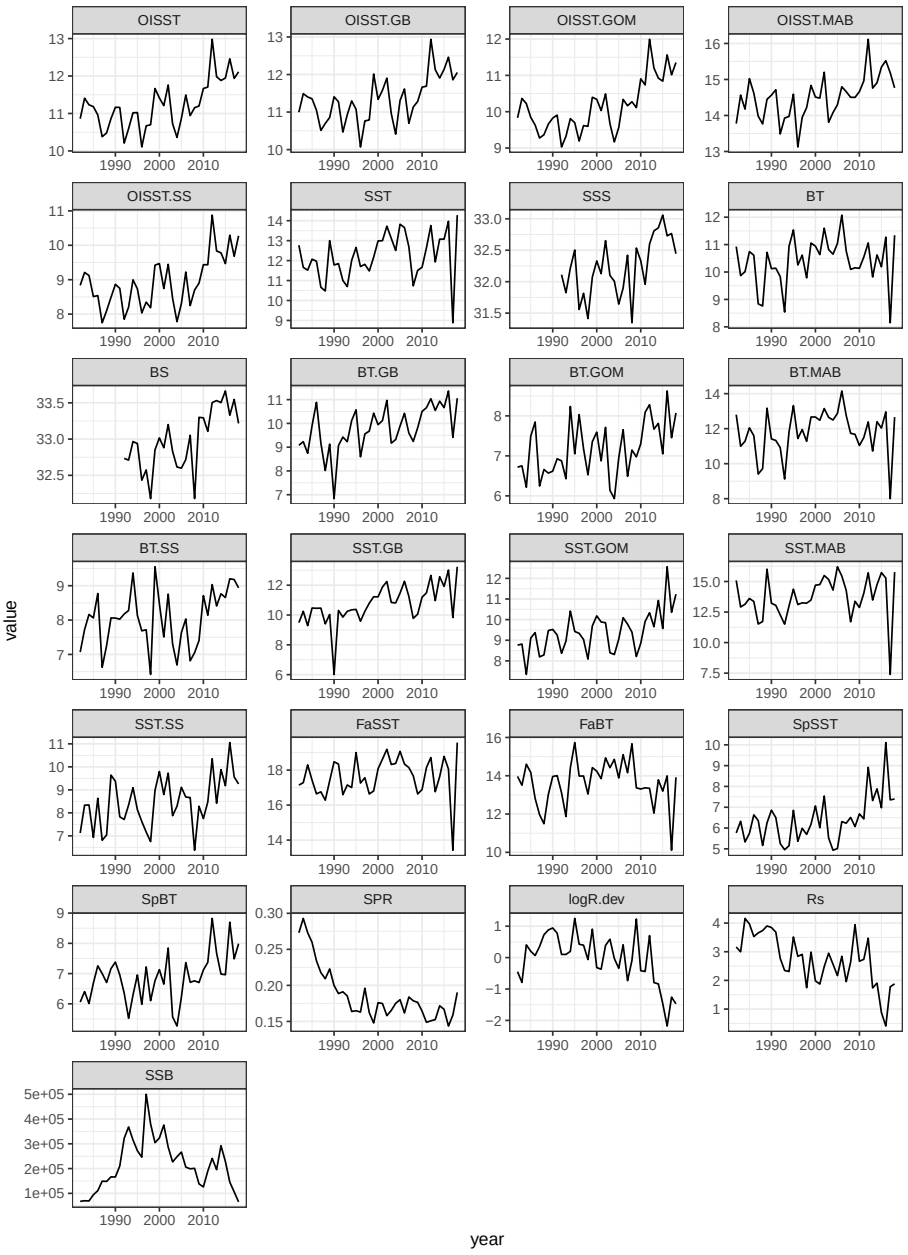
erally hovering around zero, since 2010 has been lower than time series average (no sig correlations). Deviations from stock recruit function are larger in recent years.

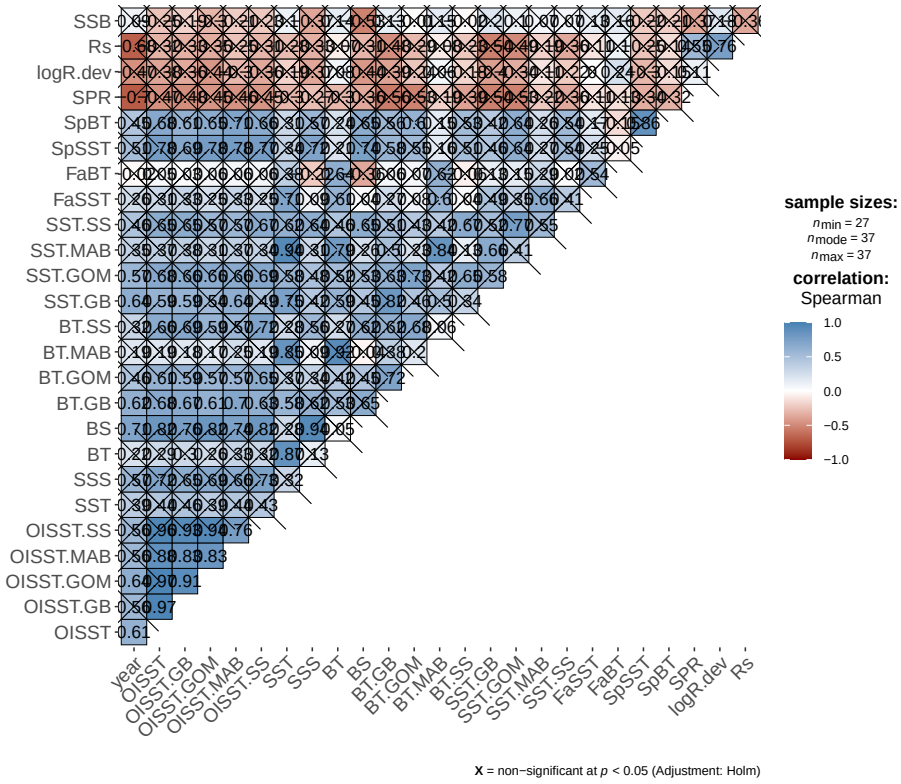
Stock Recruit Standardized Residuals: Same pattern as log deviations with shifted scale

SSB: Trimmed data misses high biomass in 70s before decline, from lows in 80s, increase into the 2000s, then steady decline to 2010 and since relatively low (no sig correlations)

Spawner per recruit vector: Was high in the 70s-early 80s, declined sharply from 1980-2000 and has since been low. This means that the proportion of fish that reach maturity has declined (sig correlation with year and GB BT)

log Recruitment success (Recruits(t)/SSB(t-1): increased throughout the 80s and 90s, declined in 1990 to intermediate levels, then declined sharply after 2010 to 2018 and has since risen (Only significantly correlated with year)


OISST
OISST.GB
OISST.GOM
OISST.MAB
OISST.SS
SST
SSS
BT
BS
BT.GB
BT.GOM
BT.MAB
BT.SS
SST.GB
SST.GOM
SST.MAB
SST.SS
FaSST
FaBT
SpSST
SpBT
SPR
logR.dev
Rs
SSB
value
year

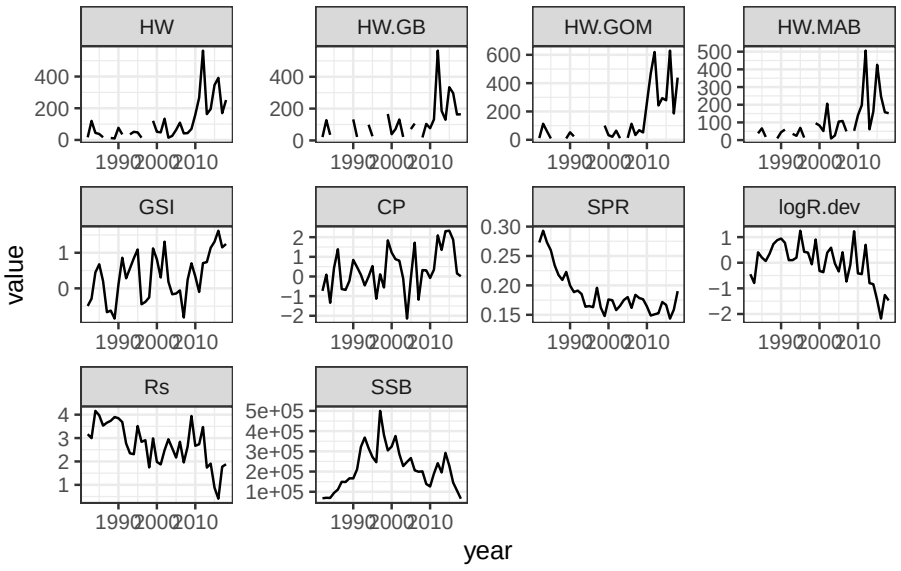


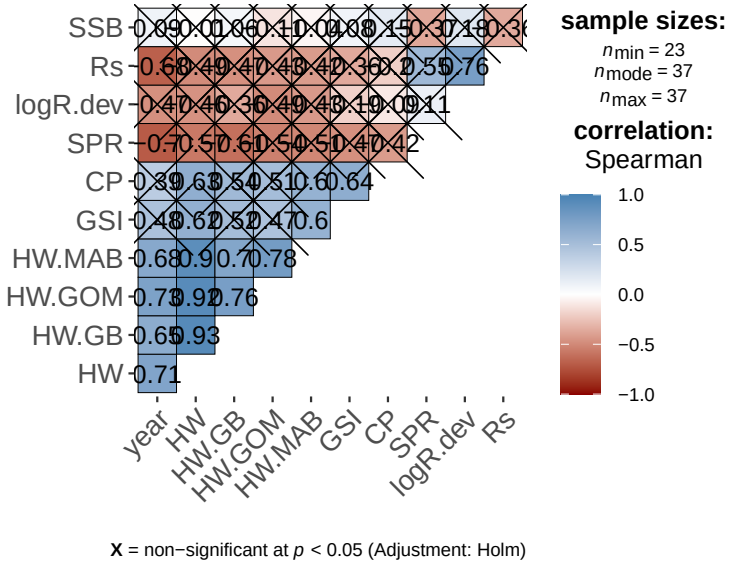
Patterns and correlations in environmental variables:

- OISST in each region is similar to the basin scale annual OISST; these were all highly correlated (drop regionals)
- OISST and survey derived SST have different patterns; not significantly

correlated (keep both)

- Bottom temp and SST patterns are different, but they are correlated (keep both)
- Bottom and surface salinity are significantly correlated and patterns are broadly similar (just keep bottom)
- Seasonal (fall and spring) SST and BT patterns are different and correlated to various of the other temperature metrics, including the regionals
- Distinct regional patterns in SST and BT that also differ from the basin wide annual averages
  - Although annual SST is correlated to the regional values, because MAB area is largest and it's a weighted average, this correlation is strongest (keep regionals)
  - Bottom temperature is not significantly correlated to all the regional bottom temps (keep regionals)





Patterns and correlations in environmental indices:

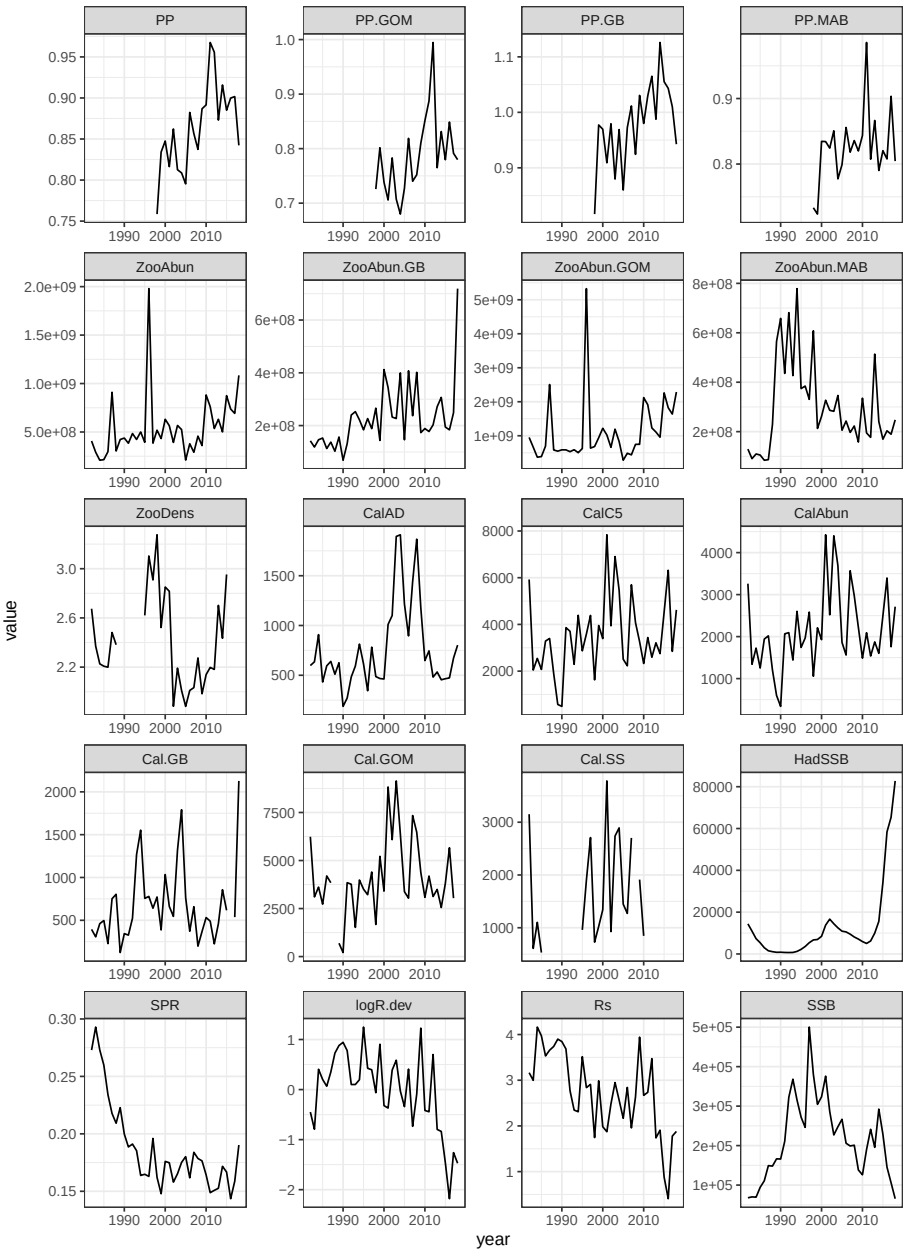
HW: Lots of missing data in heatwave index, generally low in 80s-2000s, sharp increase around 2010 and high but varying intensity in last ten years (significantly correlated with other indices)

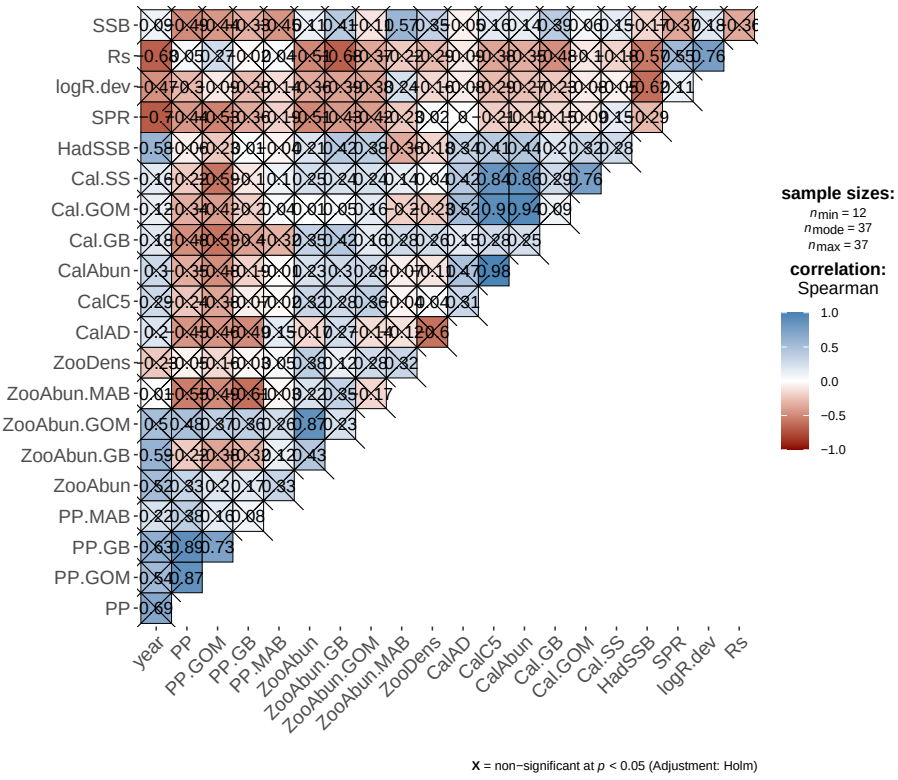
- Regional patterns broadly similar to each other and to annual (all significantly correlated)
- SPR significantly correlated to HW (& GB heatwave)

GSI: Periodic variation from 80s-2000s, increase since 2010

Cold Pool: highly variable, more positive years since 2010

```
Warning: Removed 16 row(s) containing missing values (geom_path).
```


PP
PP.GOM
PP.GB
PP.MAB
ZooAbun
ZooAbun.GB
ZooAbun.GOM
ZooAbun.MAB
ZooDens
CalAD
CalC5
CalAbun
Cal.GB
Cal.GOM
Cal.SS
HadSSB
SPR
logR.dev
Rs
SSB
value
year



Patterns and correlations in biological indices:

Primary Production: short ts, increasing from 2000-2010, slight decline since then

- Regional patterns look different from each other and annual (annual sig

corr with GOM and GB; just keep annual)

Zooplankton Abundance: Regional patterns diffrent from eacother and annual, esp MAB, all characterized by episodic spikes (only sig corr with GOM)

Zooplankton Density: Missing several years, different patterns from abundance, was high in early 2000s, rapid decline to low levels through late 2000s, steady increase since 2010

Copepod Abundance: Different patterns as total zooplankton abundance

- Stage 5 patterns similar to the total pattern (CAD diff from CC5 and Cope, but CC5 and Cope very similar; significantly correlated)
- Regional patterns differ (scotian shelf data missing many years), but GOM looks most similar to the total copepod abundance (GOM sig corr to total copepod abundance)

GB zooplankton abundance was significantly correlated to recruitment success

- Keep regional calanus abundance and regional zooplankton abundances but drop density and abundance

Haddock SSB

- Was low for most of the time series and increased rapidly around 2010

Add lags

Prepare data for BRTs –> add lags to the variables and then join back with the herring data

```
[1] 105
```

Run Boosted Regression Trees

```
# Run boosted trees for RS (recruitment success aka R(t)/SSB(t0))
ncol(brtdat) # how many columns
which(colnames(brtdat)=="Rs") # which column is Rs, the dependent variable
Rs.mod <-gbm.step(data=brtdat,
                  gbm.x=c(1, 5, 6:105), # Select variables, incl ssb & year
                  gbm.y=4, # Rs
                  family="gaussian", tree.complexity=1, # (1 = no interactions)
                  learning.rate=0.01, bag.fraction=0.7)
```

```
null.dev<-Rs.mod$self.statistics$mean.null
resid.dev<-Rs.mod$cv.statistics$deviance.mean
```

```r
dev.expl<-((null.dev-resid.dev)/null.dev)*100
dev.expl

# Plot relative influence
Rs.rel <-summary(Rs.mod)
```

```r
ggplot(data=Rs.rel,aes(x=reorder(var,rel.inf),y=rel.inf))+
  geom_bar(stat="identity")+
  labs(x="",y="relative influence")+
  coord_flip()
```

```r
# still lots of unimportant variables, zoo and copepods in GB had an extremely high relat

# Repeat w slightly smaller dataset (remove variables with 0 or near 0 influence)
brtdat.sm <- brtdat%>%
  dplyr::select(-OISST, -OISST_1,-OISST_2, -OISST_3, -SST, -SST_1,-SST_2, -SST_3, -PP, -P
ncol(brtdat.sm) # how many columns
Rs.mod1 <-gbm.step(data=brtdat.sm,
                   gbm.x=c(1, 5, 6:77),
                   gbm.y=4, # Rs
                   family="gaussian", tree.complexity=1, # (1 = no interactions)
                   learning.rate=0.01, bag.fraction=0.7)
```

```r
null.dev<-Rs.mod1$self.statistics$mean.null
resid.dev<-Rs.mod1$cv.statistics$deviance.mean
dev.expl1<-((null.dev-resid.dev)/null.dev)*100
dev.expl1

Rs.rel1 <-summary(Rs.mod1)
```

```r
relinf.fig <- ggplot(data=Rs.rel1,aes(x=reorder(var,rel.inf),y=rel.inf))+
  geom_bar(stat="identity")+
  labs(x="",y="relative influence")+
  coord_flip()
# same pattern as model above

# Can probably remove even more 0 influence sets (cal abun) and in other cases can remove
brtdat.sm2 <- brtdat.sm%>%
  dplyr::select(-CalAbun, -CalAbun_1,-CalAbun_2, -CalAbun_3, -BS_1,-BS_2, -BS_3, -SST.GOM

ncol(brtdat.sm2) # how many columns
```

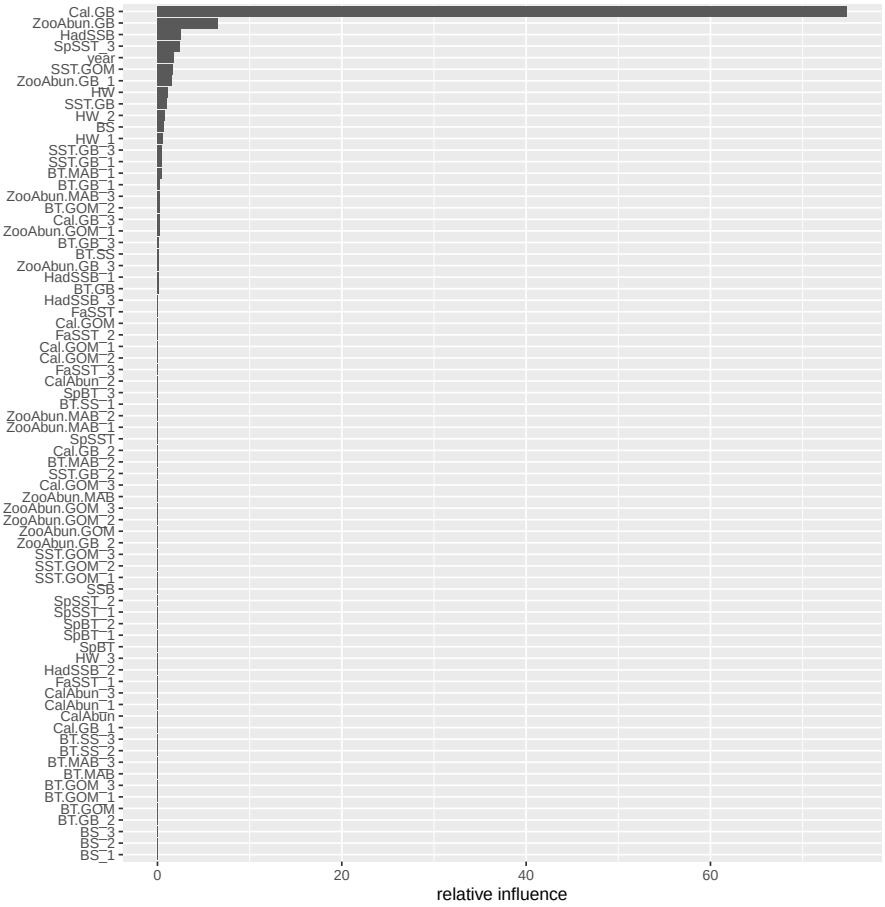
```
Rs.mod2 <-gbm.step(data=brtdat.sm2,
                   gbm.x=c(1, 5, 6:67),
                   gbm.y=4, # Rs
                   family="gaussian", tree.complexity=1, # (1 = no interactions)
                   learning.rate=0.01, bag.fraction=0.7)
```

```
null.dev<-Rs.mod2$self.statistics$mean.null
resid.dev<-Rs.mod2$cv.statistics$deviance.mean
dev.expl2<-((null.dev-resid.dev)/null.dev)*100
dev.expl2 # Dev expl decreased, perhaps removed too many

Rs.rel2 <-summary(Rs.mod2)
```

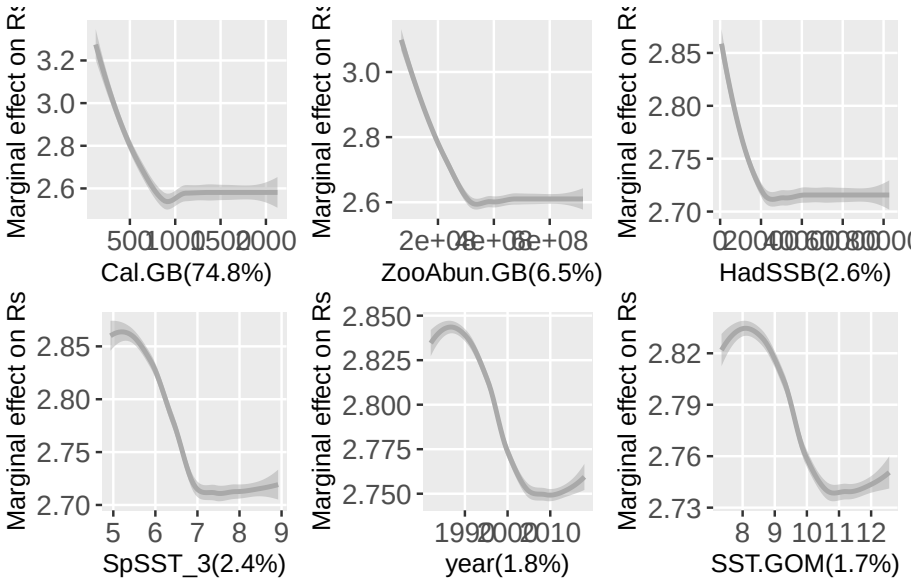
```
ggplot(data=Rs.rel2,aes(x=reorder(var,rel.inf),y=rel.inf))+
  geom_bar(stat="identity")+
  labs(x="",y="relative influence")+
  coord_flip()
```

- The two most important variables are related to food in GB followed by haddock ssb and some temperature metrics as seen in this figure of the relative influence in the model with highest deviance explained
- The deviance explained for the model shown below was 47.3039451%



- this was followed by SST in GOM and spring SST three years before
- Not shown here are models for log recruit deviations (lower deviance explained but similar relative influence plots) and for spawner/recruit (also lower deviance, but different relative influence, no GB food near the top)
- Several variables were not important at all, although a few more can still be removed

Plot partial dependence



Problematically, the partial dependence of the top two variables, which represent food abundance, are counter to expectations. Lower calanus and total zooplankton abundance was related to higher Rs. Not sure how to handle that because 80% seems really high for one variable.

Simpler BRTs

I also ran models with larger and even smaller datasets

- Models without the regional indices preformed better (higher deviance explained)
- As did models without lags
- Now haddock ssb is the most important variable followed by the year (the partial dependence plot for this model shows pretty clearly that they are now in a low recruitment regime
  - However, food variables are still among the top 6, and the direction of the partial dependence is the same: namely much higher Rs (log recruits/spawner) at low food abundances, which sharply declines at higher food
  - The fact that this same pattern keeps appearing (paired with a short conversation about this topic with my former lab mates) suggests to me that this isn't spurious and that something is there. Perhaps more food causes more competition and predation, which could end up actually reducing the larval survival