

# Lecture 3: Multivariate smoothing & model selection



University of  
St Andrews

# The story so far...

- How GAMs work
- How to include detection info
- Simple spatial-only models

# Life isn't that simple

- Which environmental covariates?
- Which response distribution?
- Which response?

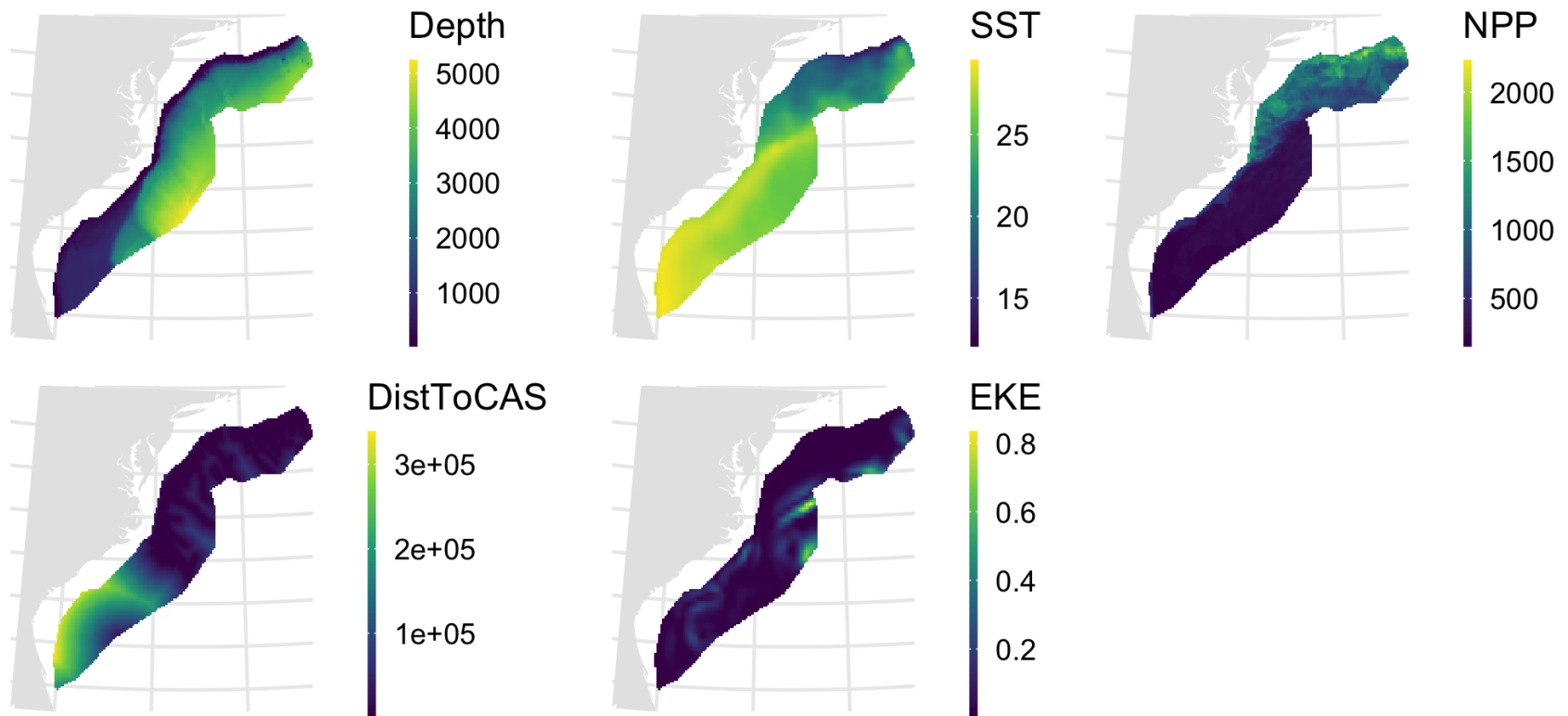
**How to select between possible models?**

# Adding covariates

# Model formulation

- Pure spatial, pure environmental, mixed?
- Prior knowledge of biology/ecology of species
- What are drivers of distribution?
- What data is available?

# Sperm whale covariates

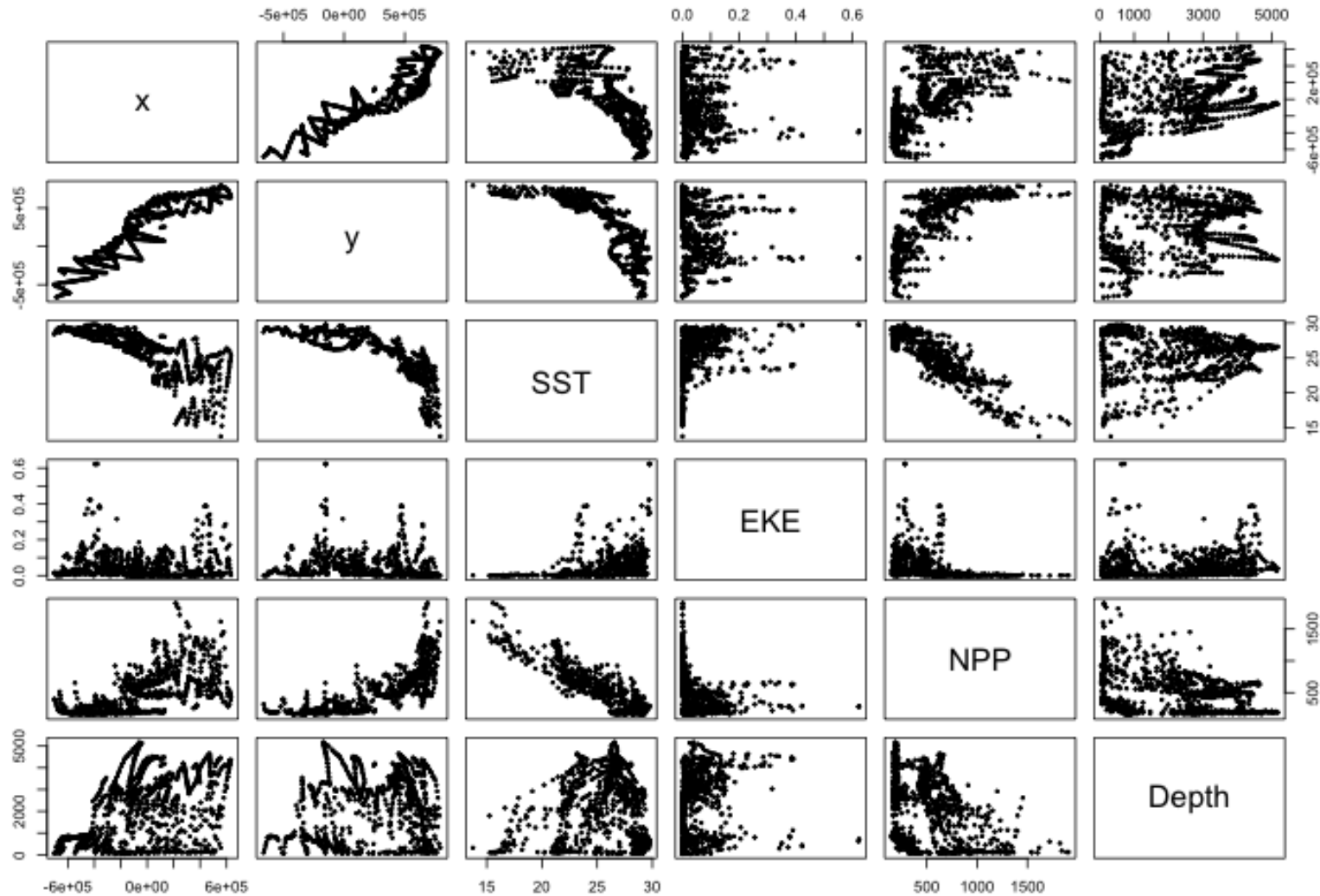


# Tobler's first law of geography

*"Everything is related to everything else, but near things are more related than distant things"*

Tobler (1970)

# Implications of Tobler's law





# Adding smooths

- Already know that + is our friend
- Can build a big model...

```
dsm_all <- dsm(count~s(x, y) +  
               s(Depth) +  
               s(DistToCAS) +  
               s(SST) +  
               s(EKE) +  
               s(NPP),  
               ddf.obj=df_hr,  
               segment.data=segs, observation.data=obs,  
               family=tw())
```

# Each `s ( )` has its own options

- `s ( . . . , k = . . . )` to adjust basis size
- `s ( . . . , bs = " . . . " )` for basis type
- lots more options (we'll see a few here)

Now we have a huge model, what do we do?

# Term selection

Two popular approaches  
(using  $p$ -values)

**Stepwise selection** - path  
dependence

**All possible subsets** -  
computationally expensive  
(fishing?)



# p-values

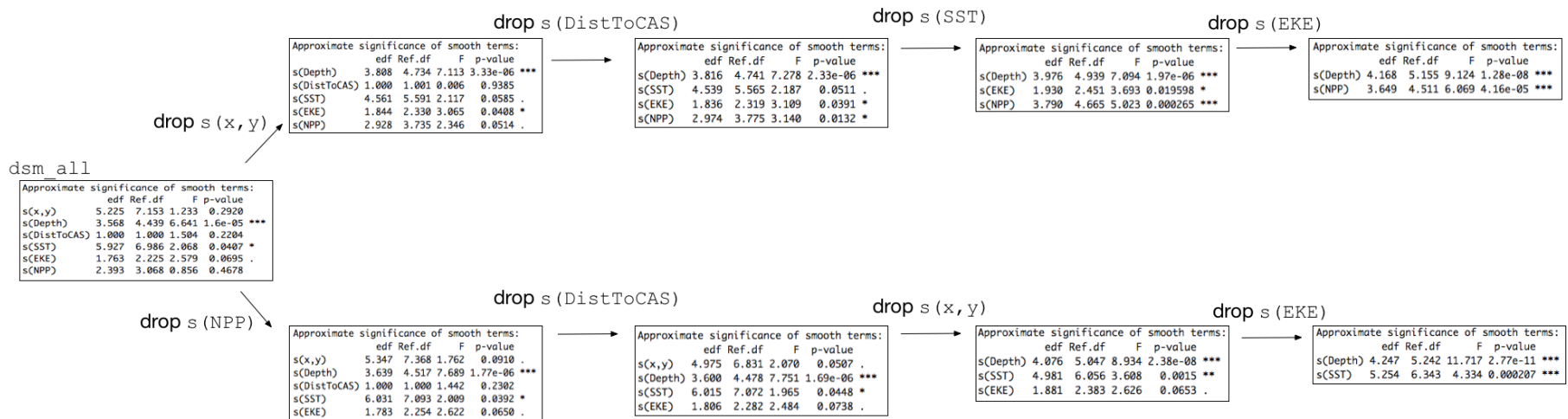
- Test for *zero effect* of a smooth
- They are **approximate** for GAMs (but useful)
- Reported in summary

# summary(dsm\_all)

```
##
## Family: Tweedie(p=1.25)
## Link function: log
##
## Formula:
## count ~ s(x, y) + s(Depth) + s(DistToCAS) + s(SST) + s(EKE) +
##       s(NPP) + offset(off.set)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.6368      0.2751    -75    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F  p-value
## s(x,y)         5.225   7.153  1.233   0.2920
## s(Depth)        3.568   4.439  6.641 1.82e-05 ***
## s(DistToCAS)    1.000   1.000  1.504   0.2204
## s(SST)          5.927   6.986  2.068   0.0407 *
## s(EKE)          1.763   2.225  2.579   0.0693 .
## s(NPP)          2.393   3.068  0.856   0.4678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Path dependence is an issue here

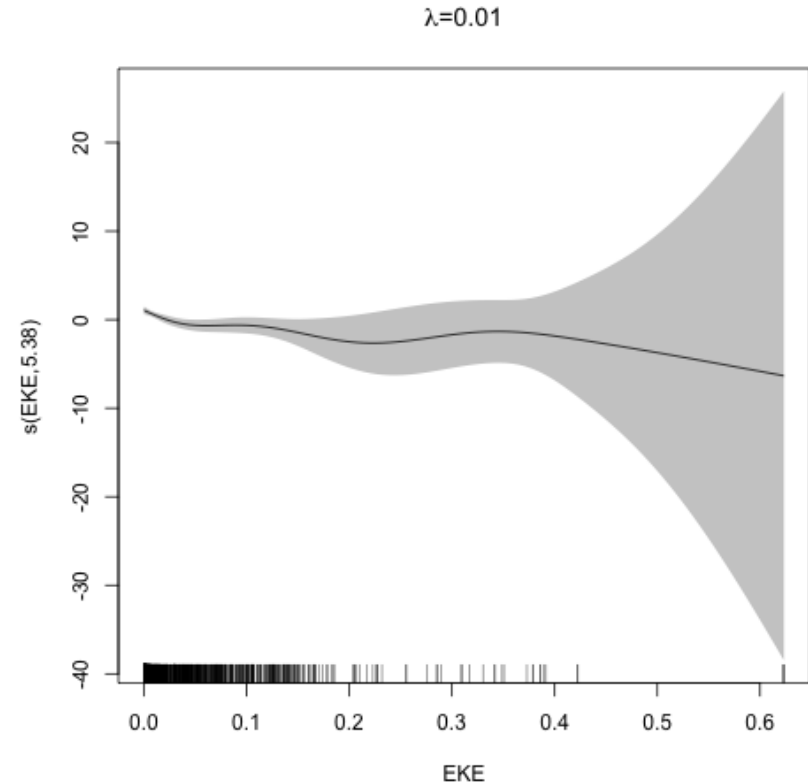
- (silly) Strategy: want all  $p \approx 0$  (\*\*\*), remove terms 1-by-1
- Two different universes appear:



This isn't very satisfactory!

# Term selection during fitting

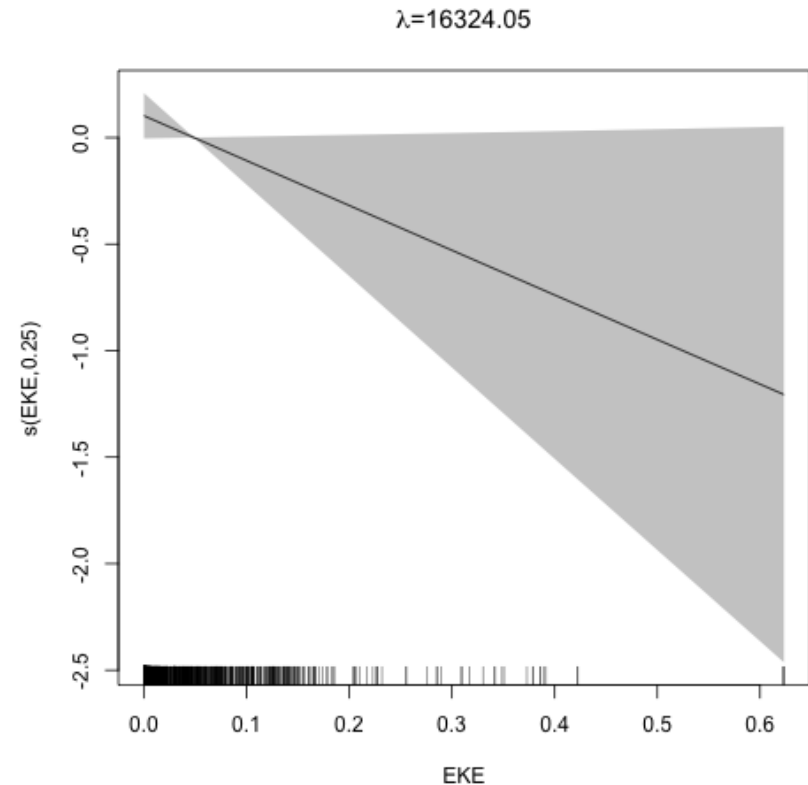
- Already selecting wigglyness of terms
- (via a penalty)
- What about using it to remove the whole term?





# Shrinkage approach

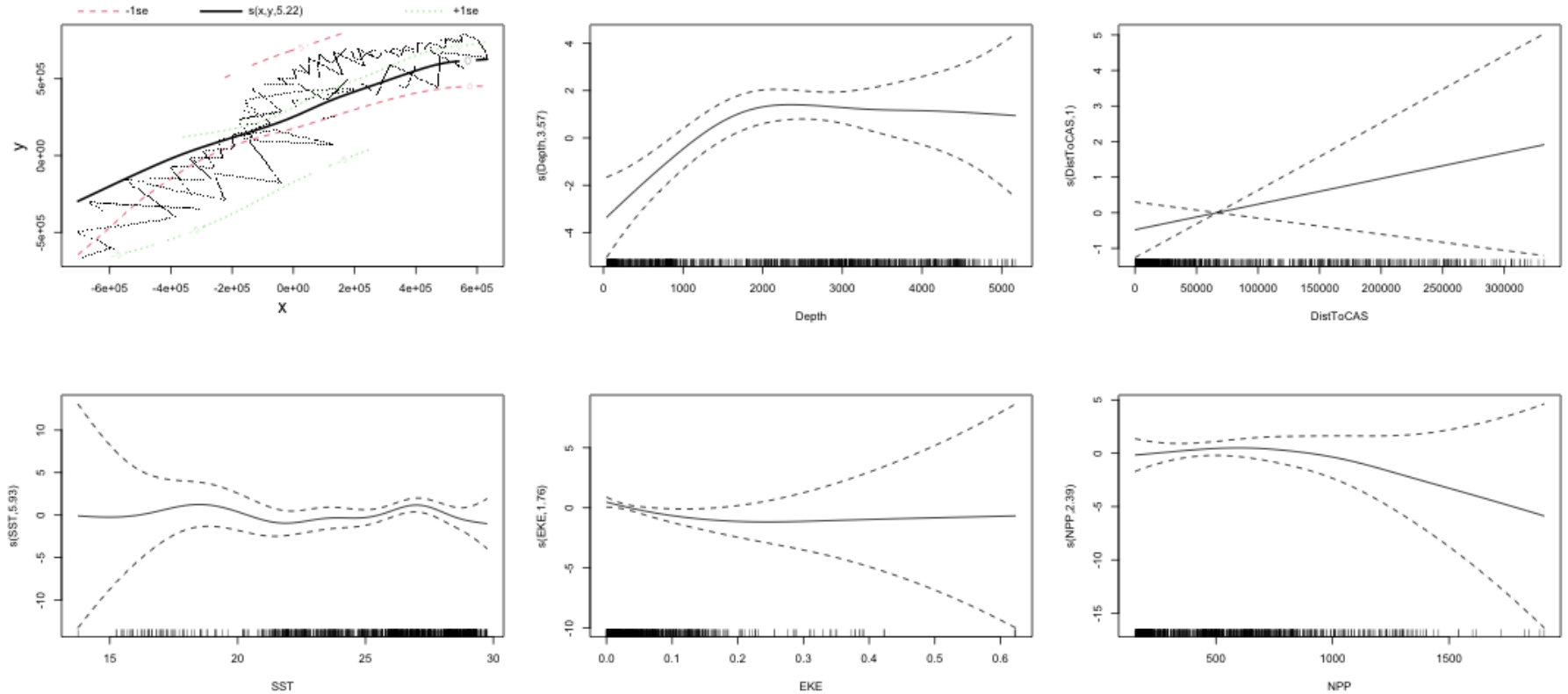
- Basis  $s(\dots, bs="ts")$  - thin plate splines *with shrinkage*
- remove the wiggles **then** remove the "linear" bits
- nullspace should be shrunk less than the wiggly part



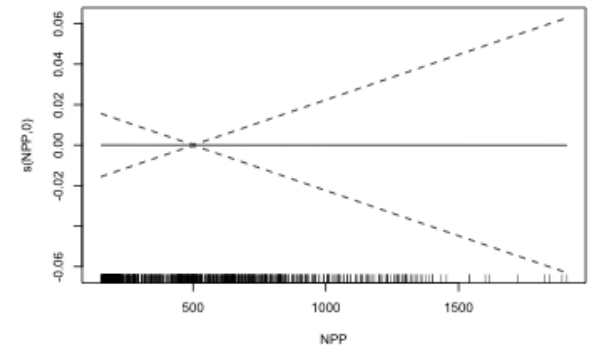
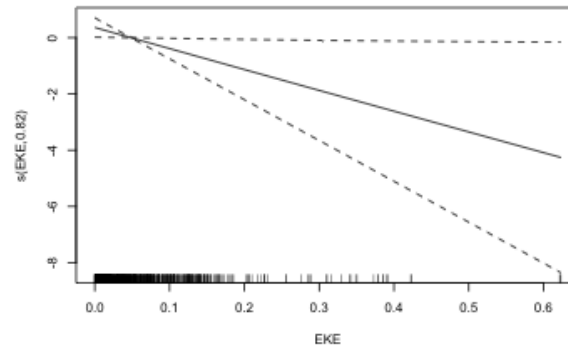
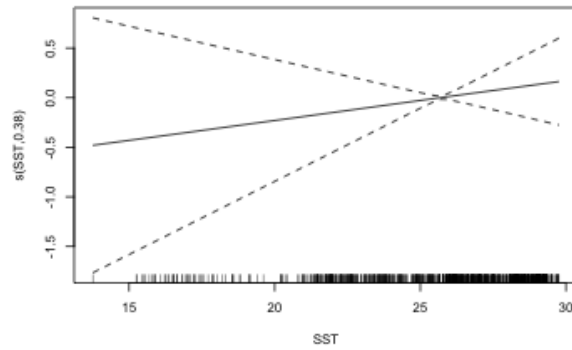
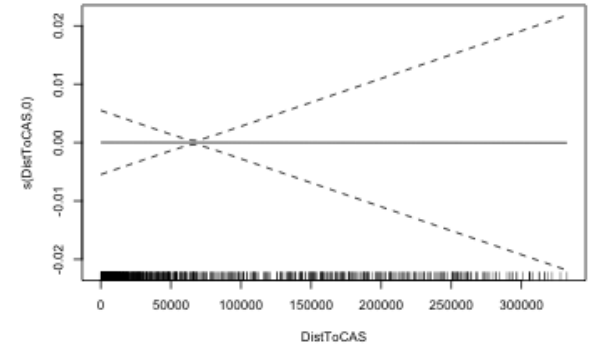
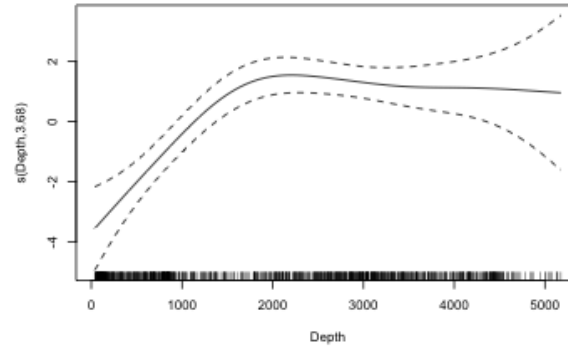
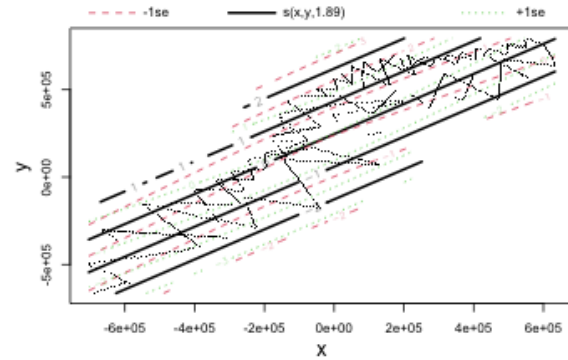
# Shrinkage example

```
dsm_ts_all <- dsm(count~s(x, y, bs="ts") +  
                  s(Depth, bs="ts") +  
                  s(DistToCAS, bs="ts") +  
                  s(SST, bs="ts") +  
                  s(EKE, bs="ts") +  
                  s(NPP, bs="ts"),  
                  ddf.obj=df_hr,  
                  segment.data=segs, observation.data=obs,  
                  family=tw())
```

# Model with no shrinkage



# ... with shrinkage



# summary(dsm\_ts\_all)

```
##
## Family: Tweedie(p=1.277)
## Link function: log
##
## Formula:
## count ~ s(x, y, bs = "ts") + s(Depth, bs = "ts") + s(DistToCAS,
##      bs = "ts") + s(SST, bs = "ts") + s(EKE, bs = "ts") + s(NPP,
##      bs = "ts") + offset(off.set)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.260      0.234   -86.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F  p-value
## s(x,y)         1.8875209     29 0.705 4.33e-06 ***
## s(Depth)        3.6794182      9 4.811 < 2e-16 ***
## s(DistToCAS)    0.0000934      9 0.000  0.6797
## s(SST)          0.3826654      9 0.063  0.2160
## s(EKE)          0.8196256      9 0.499  0.0178 *
## s(NPP)          0.0003570      9 0.000  0.8372
## ---
```

# EDF comparison

	tp	ts
s(x,y)	5.2245	1.8875
s(Depth)	3.5679	3.6794
s(DistToCAS)	1.0001	0.0001
s(SST)	5.9267	0.3827
s(EKE)	1.7631	0.8196
s(NPP)	2.3931	0.0004

# Removing terms?

## 1. EDF

- Terms with  $\text{EDF} < 1$  may not be useful (can we remove?)

## 2. non-significant $p$ -value

- Decide on a significance level and use that as a rule

(In some sense leaving "shrunk" terms in is more "consistent" in terms of variance estimation, but can be computationally annoying)

# Comparing models



# Comparing models

- Usually have  $>1$  option
- How can we pick?
- Even if we have 1 model, is it any good?

(This can be subtle, more in model checking tomorrow!)

# Akaike's "An Information Criterion"

- As for many other models, we can get an AIC from our model
- Comparison of AIC fine **but**:
  - can't compare Tweedie (continuous) and negative binomial (discrete) distributions!
  - (*within* distribution is fine)

```
AIC(dsm_all)
```

```
## [1] 1238.288
```

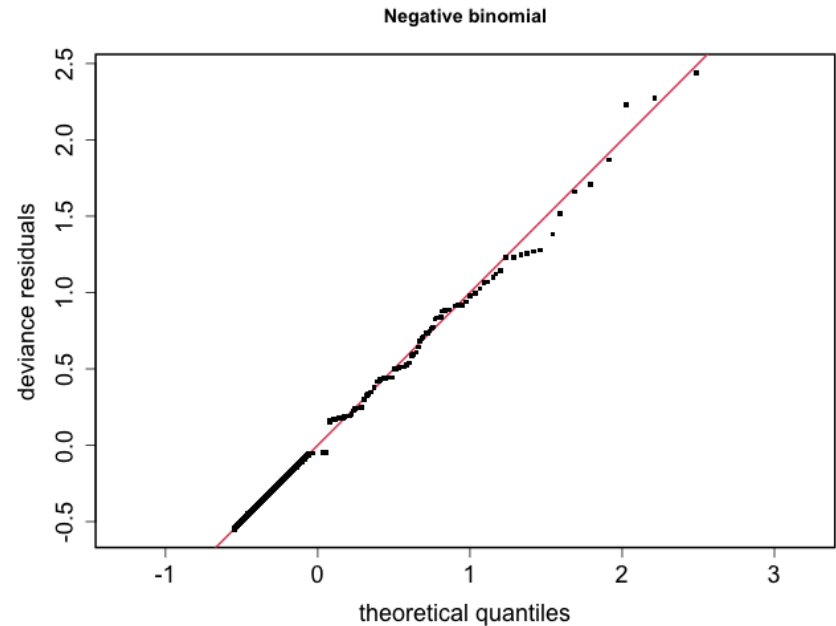
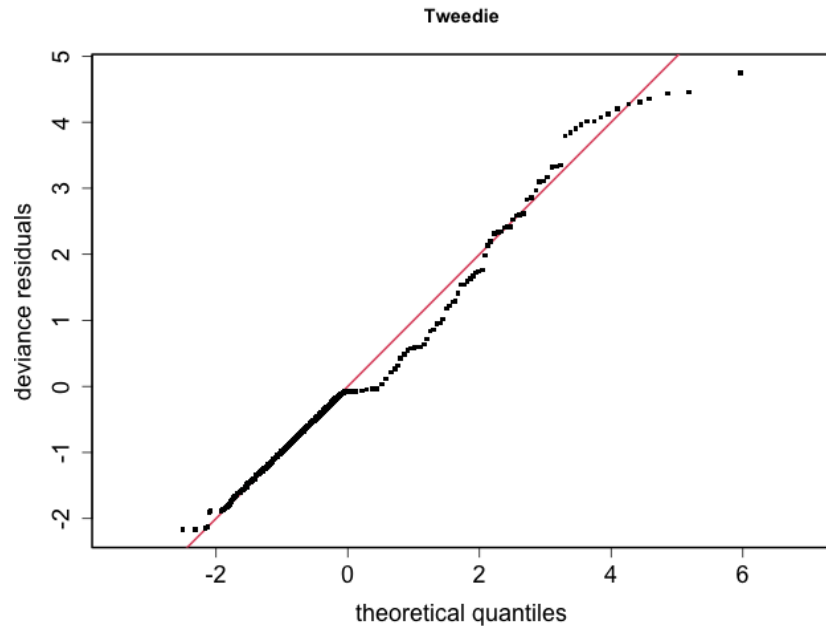
```
AIC(dsm_ts_all)
```

```
## [1] 1225.822
```

# Selecting between response distributions

# Goodness of fit

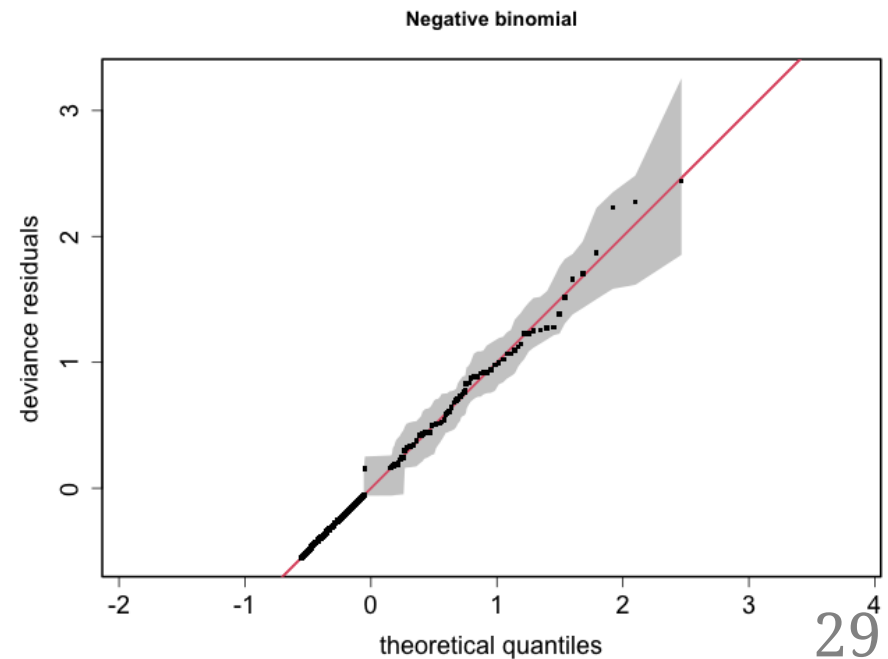
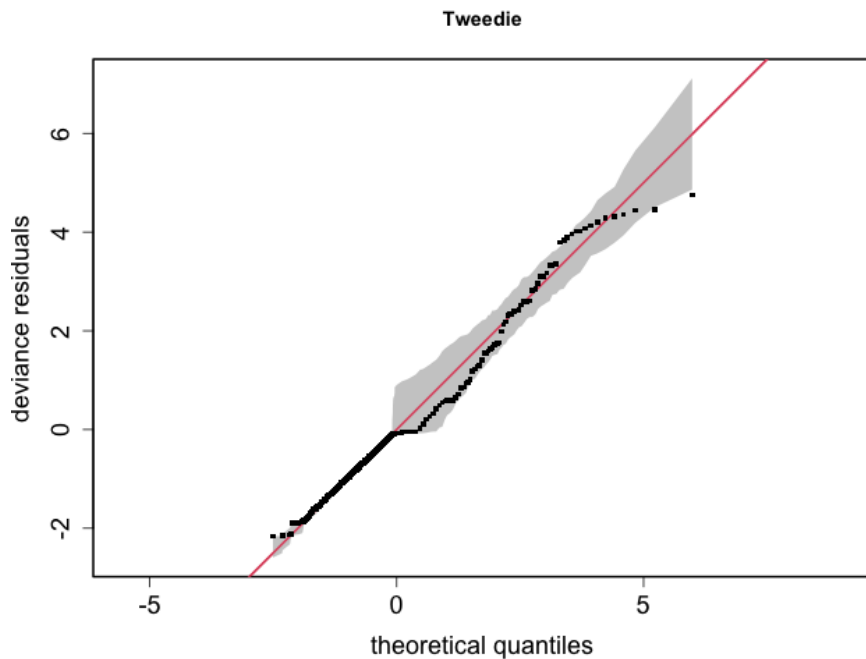
- Q-Q plots
- Closer to the line is better
- But what does "close" mean?



# Using reference bands

- What is down to random variation?
- Where does the model actually fail?
- Resampling the response, generate bands

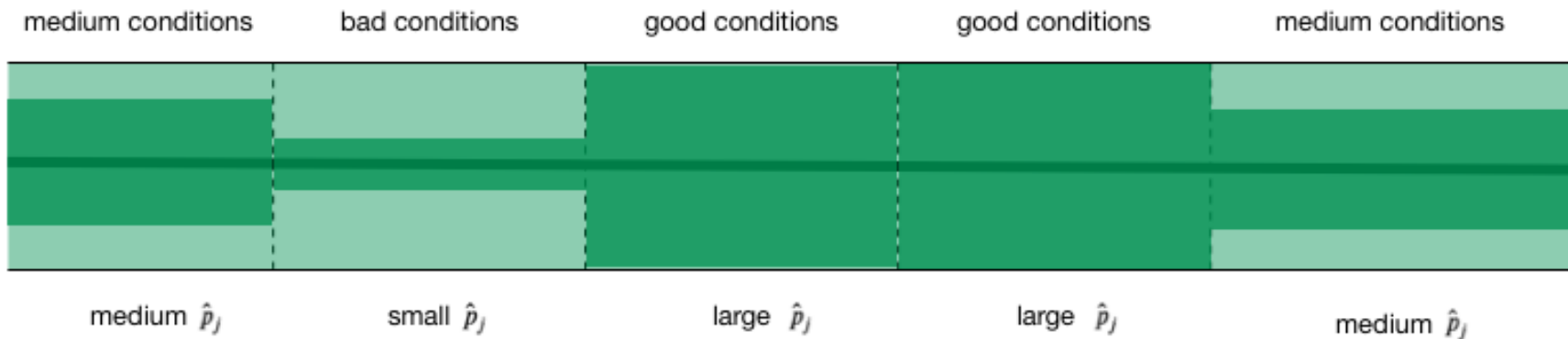
```
qq.gam(dsm_all, asp=1, main="Tweedie",  
       cex=5, rep=100)
```



Which response type?

# Count model $\text{count} \sim \dots$

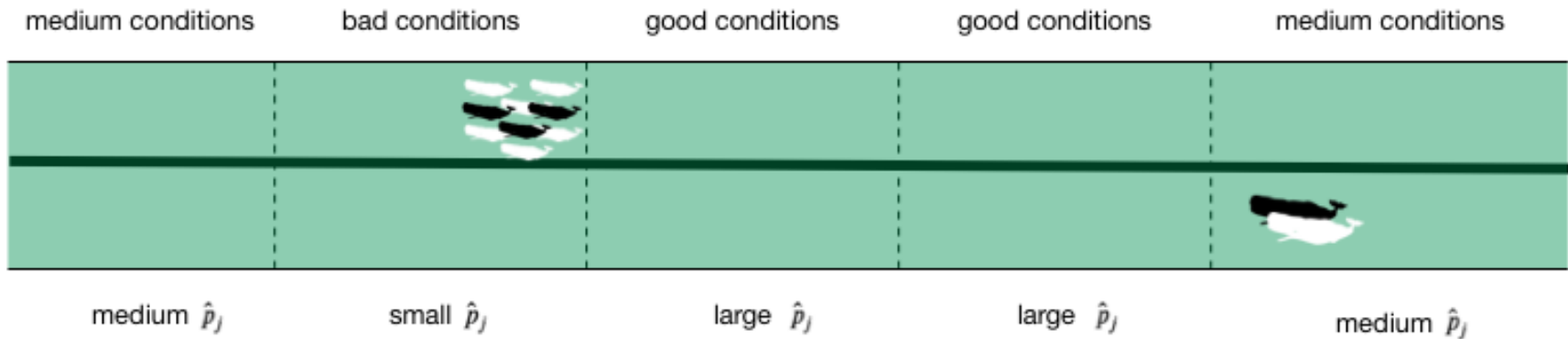
- Effort is effective effort
- Response is count per segment



# Estimated abundance

## `abundance.est ~ ...`

- Effort is area of each segment
- Response is estimated abundance per segment



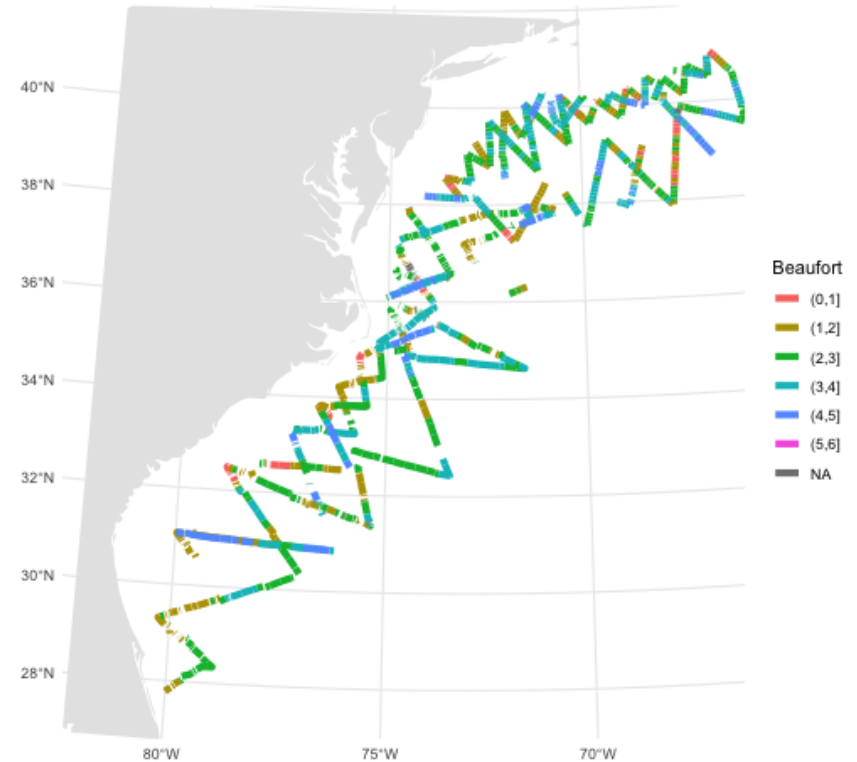


# When to use each approach?

- *Practical choice*
- 2 detection function covariate "levels"
  - "Observer"/"observation" -- change **within** segment
  - "Segment" -- change **between** segments
- "Count model" only lets us use segment-level covariates
- "Estimated abundance" lets us use either

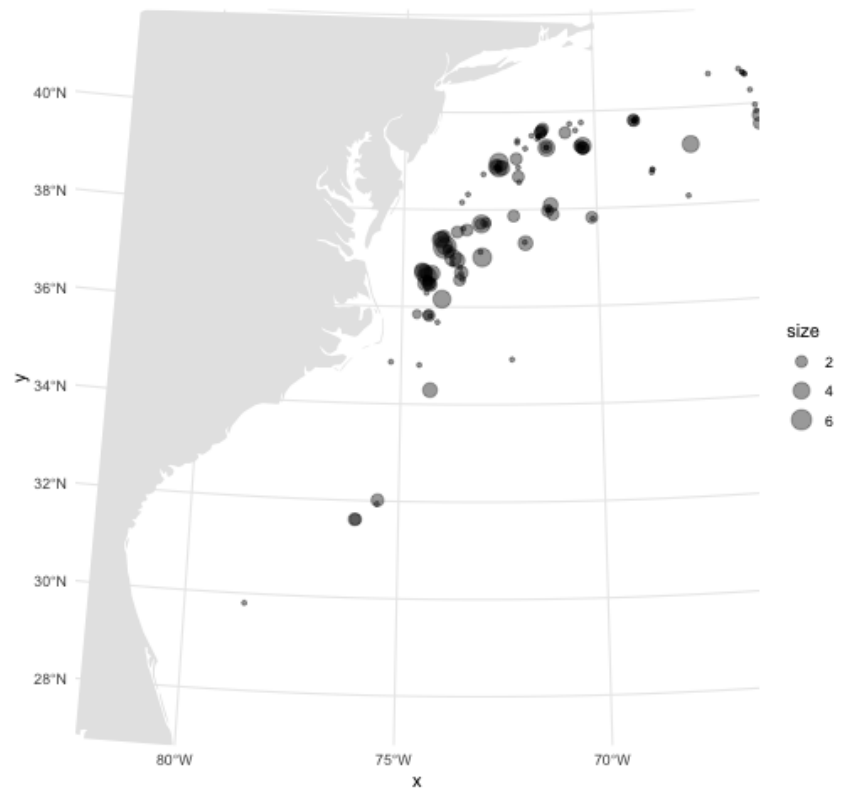
# Sperm whale response example (either)

- Detection covariate:  
Beaufort
- Changes at segment level
- `count` or `abundance.est`



# Sperm whale response example (abundance.est)

- Detection covariate:  
group size (size)
- Changes at observation  
level
- abundance.est only



# Recap

# Recap

- Adding smooths
- Path dependence
- Removing smooths
  - $p$ -values
  - shrinkage
- Comparing models
- Comparing response distributions