Acoustics-Only GAMs

Yvonne Barkley

9/26/2020

Load libraries

```
library(tidyverse)
library(mgcv)
library(corrplot)
library(geoR)
library(tidymv)
library(here)
```

Load universal variables

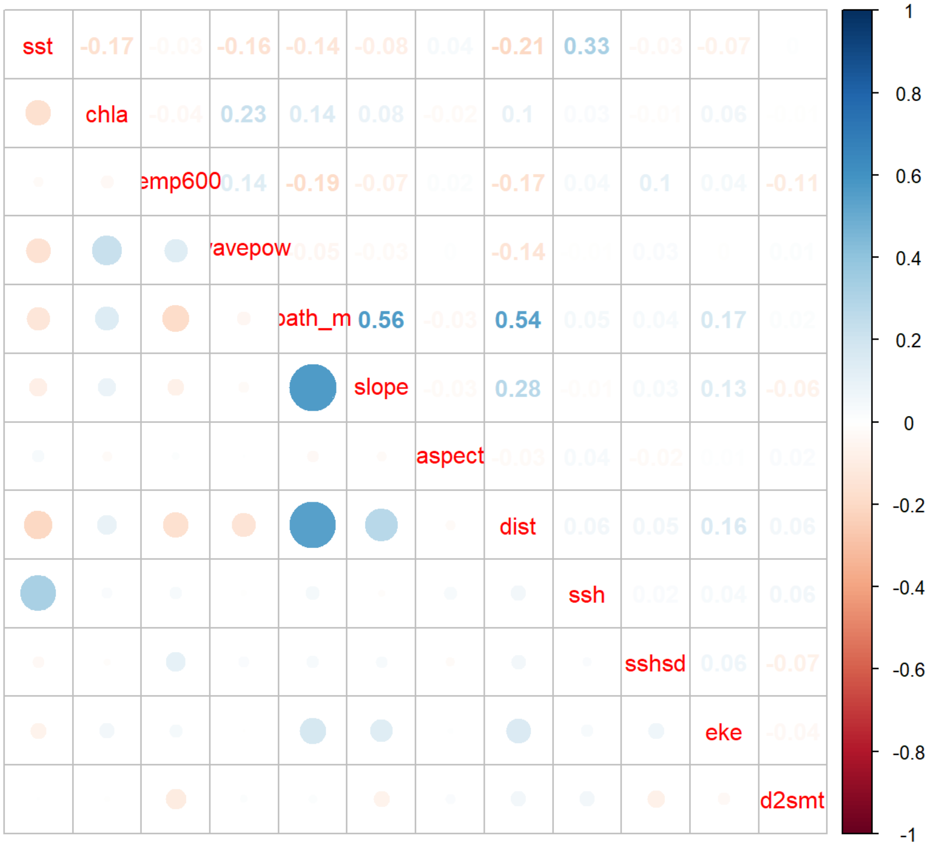
```
#Values used for file and directory names
survey = 'AllSurveys'
gridsize = 25
loctype = 'AcOnly'
loctype2 = 'Ac'
```

###Load data

```
PmScaled <- readRDS(here::here( paste0('output/models/', loctype, '/data/', 'CompletePm_', gridsize, 'km_', loctype2, '_scaled.rda') ))
# add column for log effort as offset #
PmScaled$log.effort = log(PmScaled$EffArea)
```

Check correlation of covariates

```
require(corrplot)
corrplot.mixed(cor(PmScaled[,18:29]), upper="number", lower="circle")
```



```
# Are all correlation coefficients < |0.6|?
abs(cor(PmScaled[,18:29])) <= 0.6
```

```
          sst  chla temp600 wavepow bath_m slope aspect  dist   ssh sshsd   eke
sst     FALSE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
chla     TRUE FALSE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
temp600  TRUE  TRUE   FALSE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
wavepow  TRUE  TRUE    TRUE   FALSE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
bath_m   TRUE  TRUE    TRUE    TRUE  FALSE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
slope    TRUE  TRUE    TRUE    TRUE   TRUE FALSE   TRUE  TRUE  TRUE  TRUE  TRUE
aspect   TRUE  TRUE    TRUE    TRUE   TRUE  TRUE  FALSE  TRUE  TRUE  TRUE  TRUE
dist     TRUE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE FALSE  TRUE  TRUE  TRUE
ssh      TRUE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE FALSE  TRUE  TRUE
sshsd    TRUE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE FALSE  TRUE
eke      TRUE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE FALSE
d2smt    TRUE  TRUE    TRUE    TRUE   TRUE  TRUE   TRUE  TRUE  TRUE  TRUE  TRUE
        d2smt
sst      TRUE
chla     TRUE
temp600  TRUE
wavepow  TRUE
bath_m   TRUE
slope    TRUE
aspect   TRUE
dist     TRUE
ssh      TRUE
sshsd    TRUE
eke      TRUE
d2smt   FALSE
```

KS tests

I compared the distributions of environmental data between the whales and the absences. Plots are attached in separate powerpoint. In summary, temperature at 600 m, SSH, and chlorophyll were the only variables with significantly different distributions (p-value < 0.05). However, the D statistics were close to zero (D ~ 0.1) for each, indicating that although the distributions were different, they were not that far apart. The plots also show how similar the general shape of the distributions are between where the whales were observed and where they were not observed.

Data Splitting

Split the data into train and test sets

```
require(dplyr)
splitdf <- function(dataframe, seed=NULL) {
    if (!is.null(seed)) set.seed(seed)
    index <- 1:nrow(dataframe)
    trainindex <- sample(index, trunc(length(index)*0.7))
    trainset <- dataframe[trainindex, ]
    testset <- dataframe[-trainindex, ]
    list(trainset=trainset,testset=testset)
}
trainAcOnly = NULL
testAcOnly = NULL
for (s in c(1641, 1303, 1604, 1705, 1706)){

 trSub <- filter(PmScaled, survey == s)

 #subset for presences and split 70/30
 pres1 <- filter(trSub, pa > 0 & sid == 999 )
 listPres  <- splitdf(pres1, 555) #output is list for train and test

 #subset for absences and split 70/30
 abs0  <- filter(trSub, pa == 0 )
 listAbs <- splitdf(abs0, 555)  #output is list for train and test

 #combine train data for presence and absence
 trainAll <- rbind( listPres$trainset, listAbs$trainset )

 #combine test data for presence and absence
 testAll  <- rbind( listPres$testset,  listAbs$testset  )

trainAcOnly = rbind( trainAcOnly, trainAll )
testAcOnly  = rbind( testAcOnly,  testAll)

# trainAcOnly$log.effort <- log(trainAcOnly$EffArea)
# testAcOnly$log.effort <- log(testAcOnly$EffArea)
}
saveRDS(trainAcOnly, here::here(  paste0('output/models/',loctype, '/data/Train_', gridsize, 'km
_', loctype2, '_S999.rda')  ))
saveRDS(testAcOnly, here::here(  paste0('output/models/',loctype, '/data/Test_', gridsize, 'km_'
, loctype2, '_S999.rda')  ))

# nrow(dplyr::filter(trainAcOnly, trainAcOnly$pa >0))
# nrow(dplyr::filter(testAcOnly, testAcOnly$pa >0))
```

Generalized Additive Models

The data are treated as count data, number of sperm whale encounters per cell, and we used the Tweedie distribution since it has been shown to work well when fewer positive responses exist within the data. We used thin-plate regression splines (the default basis) for the smoothers of the environmental predictors. Each smoother was limited to 3 degrees of freedom (k=3) to reduce overfitting parameters per recommendations from other studies building similar types of cetaceans distribution models.The log of the effort was included as an offset to account for the variation in effort per cell.

Tweedie - 25 km spatial scale

- Knots contrained to k=3 according to literature on cetacean distribution models.
- Automatic term selection is uses an additional penalty term when determining the smoothness of the function (‘select’ argument = TRUE)..
- We excluded all non-significant variables (alpha=0.05) and refit the models until all variables were significant.
- REML is restricted maximum likelihood used to optimize the parameter estimates.

Load training and test data

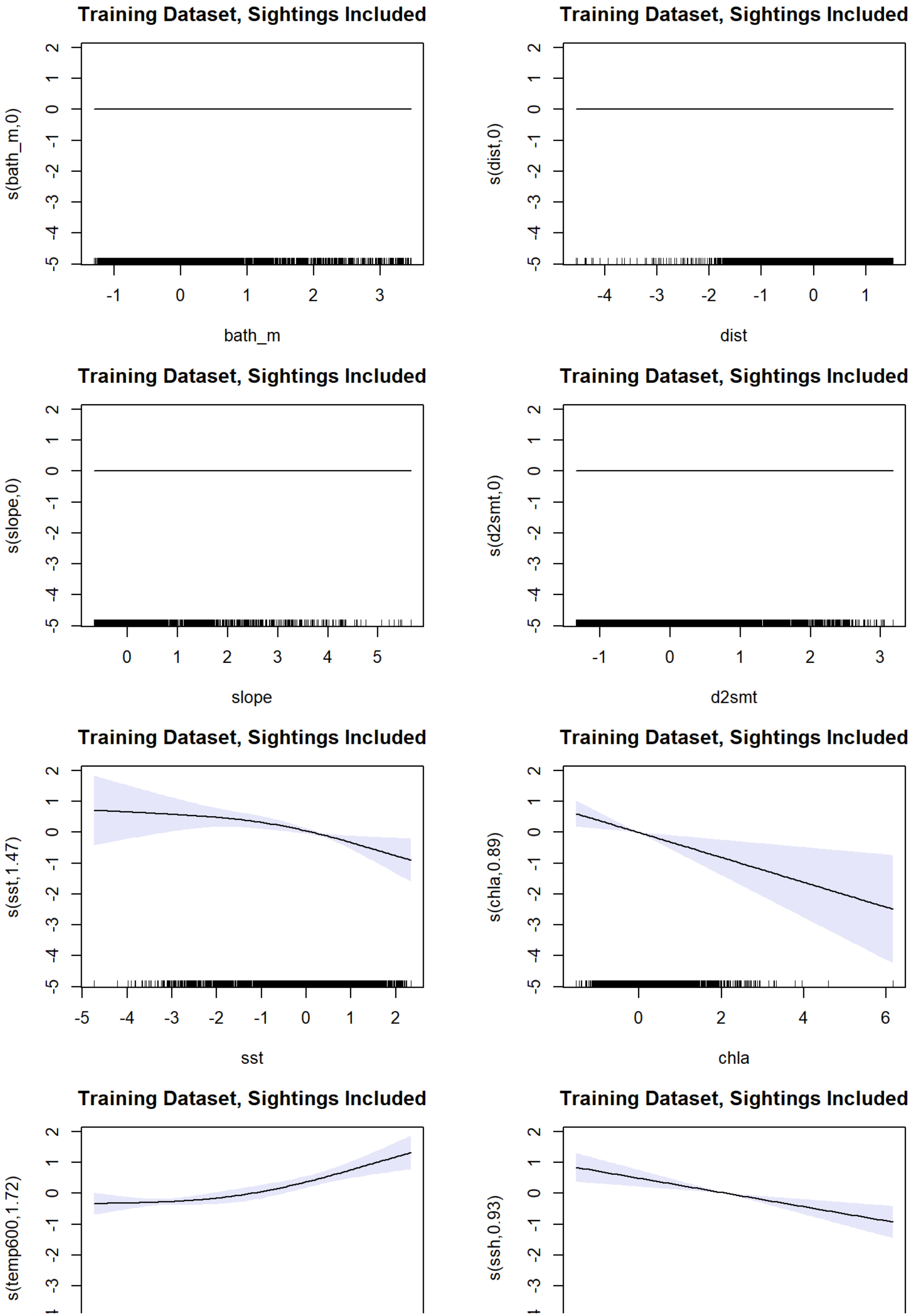
```
trainAcOnly <- readRDS(here::here(  paste0('output/models/',loctype, '/data/Train_', gridsize,
'km_', loctype2, '.rda')  ))
testAcOnly <- readRDS(here::here(  paste0('output/models/',loctype, '/data/Test_',   gridsize,
'km_', loctype2, '.rda')  ))
trainS999 <- readRDS(here::here(  paste0('output/models/',loctype, '/data/Train_',   gridsize,
'km_', loctype2, '_S999.rda')  ))
testS999 <- readRDS(here::here(  paste0('output/models/',loctype, '/data/Test_',     gridsize,
'km_', loctype2, '_S999.rda')  ))
```

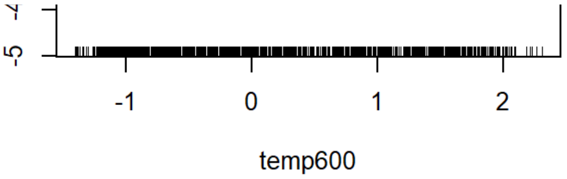
Using the training data to build the model with all parameters

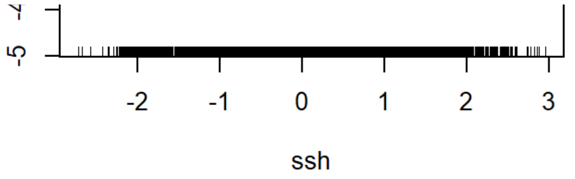
*Includes Sighted Acoustic Encounters (Filtered chla for values < 10 (scaled))

```
require(mgcv)
#with training dataset
twFull <- gam(pa ~ s(bath_m, k=3) + s(dist, k=3) + s(slope, k=3) + s(d2smt, k=3) +  s(sst, k=3)
 + s(chla, k=3) + s(temp600, k=3) + s(ssh, k=3) + s(sshsd, k=3) + s(eke, k=3) + s(wavepow, k=3)
 + offset(log.effort), data = trainAcOnly, family = tw, link = 'log', select = TRUE, method = "R
EML")
summary(twFull)
```

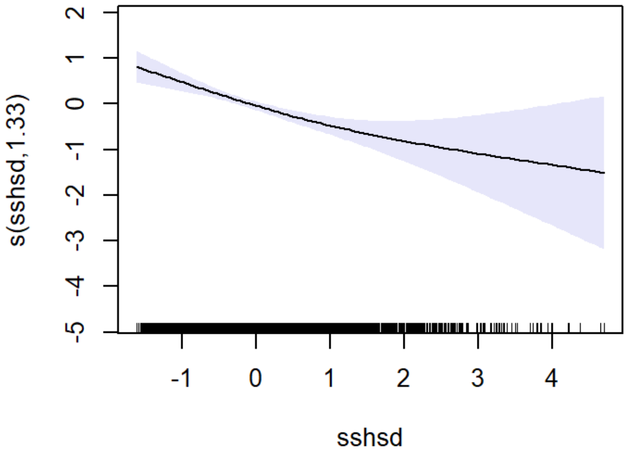
```

Family: Tweedie(p=1.01) 
Link function: log 

Formula:
pa ~ s(bath_m, k = 3) + s(dist, k = 3) + s(slope, k = 3) + s(d2smt, 
    k = 3) + s(sst, k = 3) + s(chla, k = 3) + s(temp600, k = 3) + 
    s(ssh, k = 3) + s(sshsd, k = 3) + s(eke, k = 3) + s(wavepow, 
    k = 3) + offset(log.effort)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.63378    0.09841    -230   <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(bath_m)  1.464e-04      2  0.000 0.635207    
s(dist)    1.109e-04      2  0.000 1.000000    
s(slope)   9.127e-05      2  0.000 0.998142    
s(d2smt)   1.377e-04      2  0.000 0.853560    
s(sst)     1.473e+00      2  5.661 0.000598 ***
s(chla)    8.937e-01      2  4.052 0.002188 ** 
s(temp600) 1.718e+00      2 16.396 7.34e-09 ***
s(ssh)     9.287e-01      2  6.473 0.000180 ***
s(sshsd)   1.331e+00      2 11.879 4.29e-07 ***
s(eke)     7.773e-01      2  1.495 0.049183 *  
s(wavepow) 8.453e-05      2  0.000 0.473799    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0847   Deviance explained = 11.4%
-REML = 372.93  Scale est. = 1.0362    n = 3696
```


Training Dataset, Sightings Included
s(bath_m,0)
2
1
0
-1
-2
-3
-4
-5
-1
0
1
2
3
bath_m
Training Dataset, Sightings Included
s(dist,0)
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
dist
Training Dataset, Sightings Included
s(slope,0)
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5
slope
Training Dataset, Sightings Included
s(d2smt,0)
2
1
0
-1
-2
-3
-4
-5
-1
0
1
2
3
d2smt
Training Dataset, Sightings Included
s(sst,1.47)
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
sst
Training Dataset, Sightings Included
s(chla,0.89)
2
1
0
-1
-2
-3
-4
-5
0
2
4
6
chla
Training Dataset, Sightings Included
s(temp600,1.72)
2
1
0
-1
-2
-3
Training Dataset, Sightings Included
s(ssh,0.93)
2
1
0
-1
-2
-3

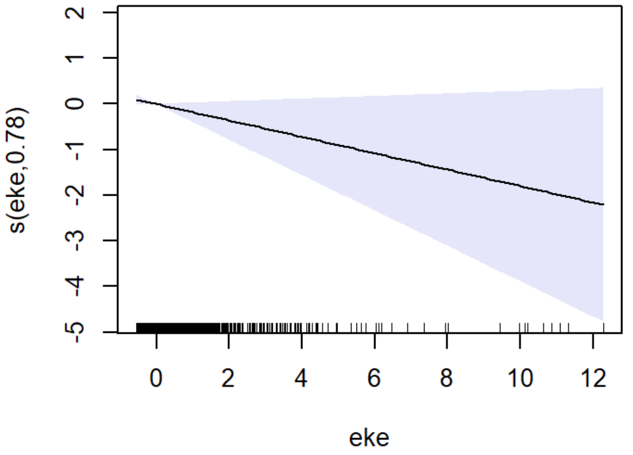




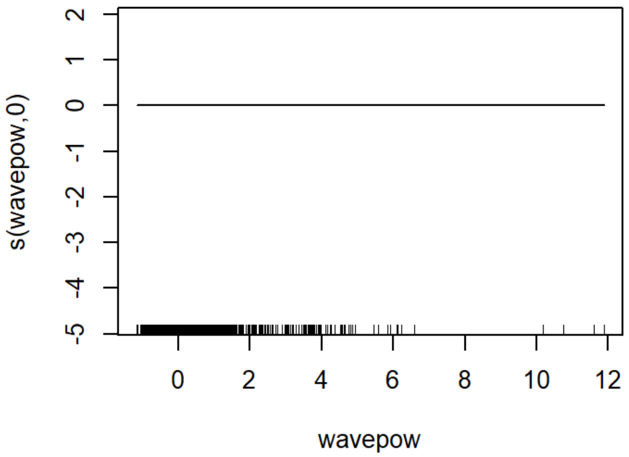
Training Dataset, Sightings Included

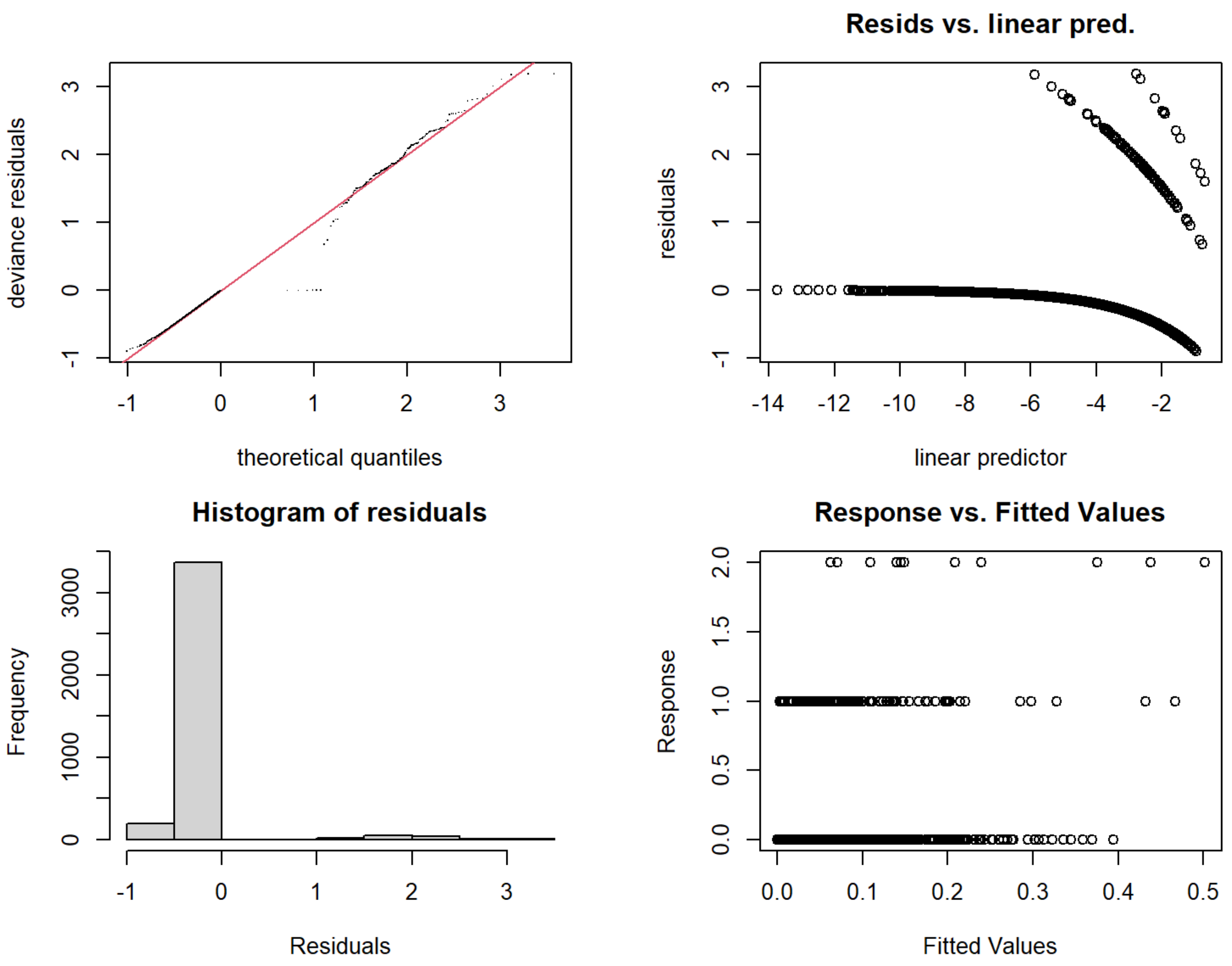


Training Dataset, Sightings Included



Training Dataset, Sightings Included




deviance residuals
theoretical quantiles
-1
0
1
2
3
Resids vs. linear pred.
residuals
linear predictor
-14
-12
-10
-8
-6
-4
-2
Histogram of residuals
Frequency
Residuals
0
1000
2000
3000
Response vs. Fitted Values
Response
Fitted Values
0.0
0.5
1.0
1.5
2.0
0.1
0.2
0.3
0.4
0.5

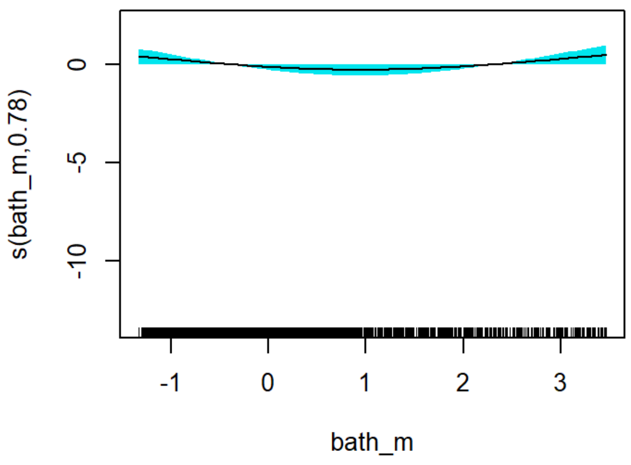
```
Method: REML   Optimizer: outer newton
full convergence after 22 iterations.
Gradient range [-0.001498619,0.0001905826]
(score 372.9256 & scale 1.036199).
Hessian positive definite, eigenvalue range [3.146237e-06,14799.46].
Model rank =  23 / 23

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                 k'      edf k-index p-value
s(bath_m)  2.00e+00 1.46e-04    0.85   0.005 **
s(dist)    2.00e+00 1.11e-04    0.85  <2e-16 ***
s(slope)   2.00e+00 9.13e-05    0.83  <2e-16 ***
s(d2smt)   2.00e+00 1.38e-04    0.86   0.025 *
s(sst)     2.00e+00 1.47e+00    0.84  <2e-16 ***
s(chla)    2.00e+00 8.94e-01    0.85  <2e-16 ***
s(temp600) 2.00e+00 1.72e+00    0.72  <2e-16 ***
s(ssh)     2.00e+00 9.29e-01    0.83   0.005 **
s(sshsd)   2.00e+00 1.33e+00    0.82  <2e-16 ***
s(eke)     2.00e+00 7.77e-01    0.84  <2e-16 ***
s(wavepow) 2.00e+00 8.45e-05    0.79  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Does NOT include sighted acoustic encounters

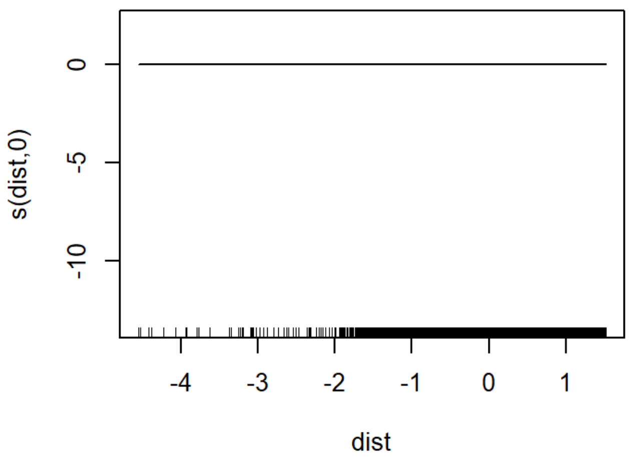
```
require(mgcv)
#with training dataset
twS99 <- gam(pa ~ s(bath_m, k=3) + s(dist, k=3) + s(slope, k=3) + s(d2smt, k=3) +  s(sst, k=3) +
s(chla, k=3) + s(temp600, k=3) + s(ssh, k=3) + s(sshsd, k=3) + s(eke, k=3) + s(wavepow, k=3) + o
ffset(log.effort), data = trainS999, family = tw, link = 'log', select = TRUE, method = "REML")
summary(twS99)
```

```

Family: Tweedie(p=1.01) 
Link function: log 

Formula:
pa ~ s(bath_m, k = 3) + s(dist, k = 3) + s(slope, k = 3) + s(d2smt, 
    k = 3) + s(sst, k = 3) + s(chla, k = 3) + s(temp600, k = 3) + 
    s(ssh, k = 3) + s(sshsd, k = 3) + s(eke, k = 3) + s(wavepow, 
    k = 3) + offset(log.effort)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.9641     0.1149  -199.9   <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(bath_m)  7.838e-01      2 1.809  0.03087 *  
s(dist)    1.152e-04      2 0.000  1.00000    
s(slope)   5.856e-05      2 0.000  0.35398    
s(d2smt)   1.590e-04      2 0.000  0.47607    
s(sst)     6.467e-01      2 0.963  0.08094 .  
s(chla)    8.417e-01      2 2.530  0.01360 *  
s(temp600) 1.762e+00      2 9.977 1.17e-05 ***
s(ssh)     8.987e-01      2 4.402  0.00155 ** 
s(sshsd)   8.384e-01      2 2.489  0.01457 *  
s(eke)     8.362e-01      2 2.171  0.02238 *  
s(wavepow) 7.465e-05      2 0.000  1.00000    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.047   Deviance explained = 9.36%
-REML = 295.31  Scale est. = 1.0404    n = 3660
```

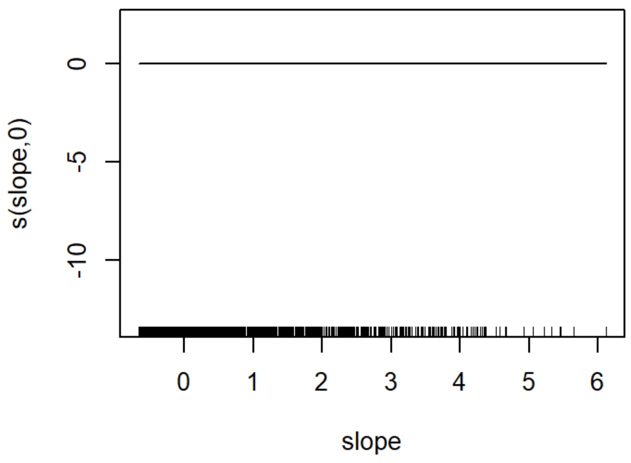
Training Dataset, No Sightings



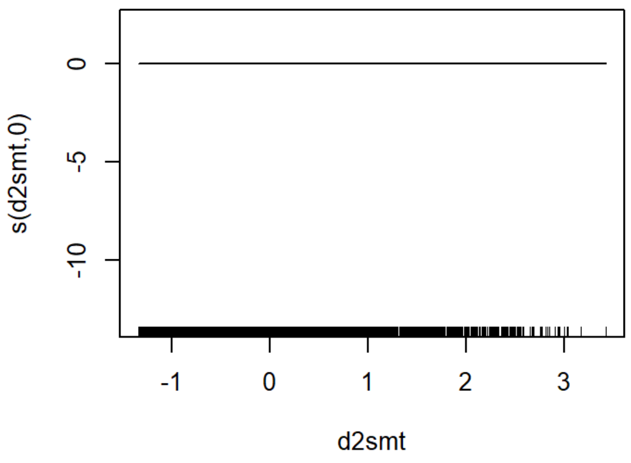
Training Dataset, No Sightings



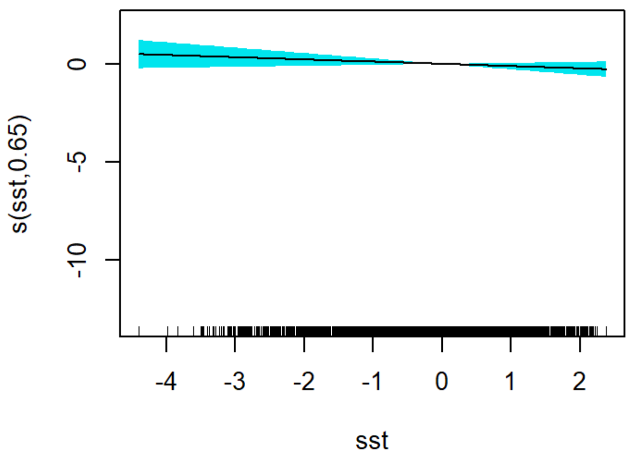
Training Dataset, No Sightings



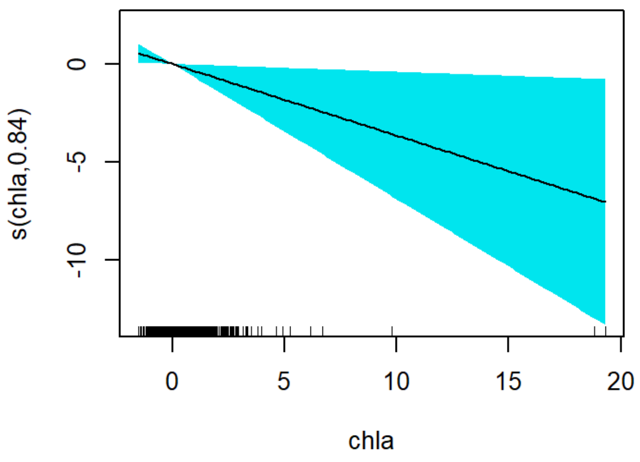
Training Dataset, No Sightings



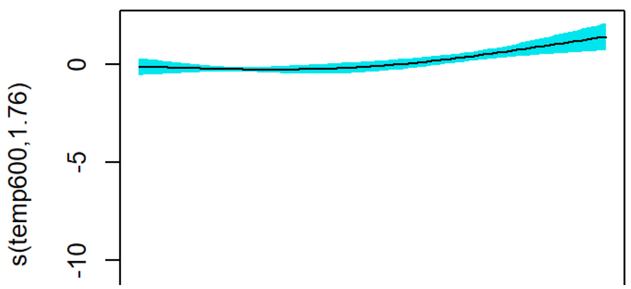
Training Dataset, No Sightings



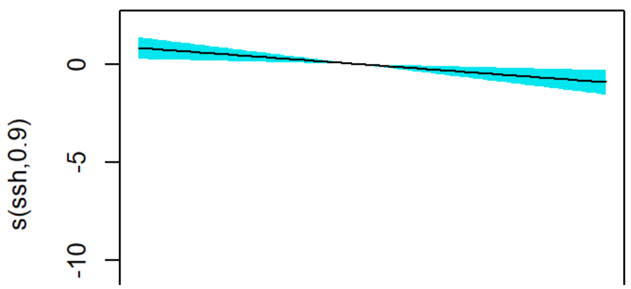
Training Dataset, No Sightings

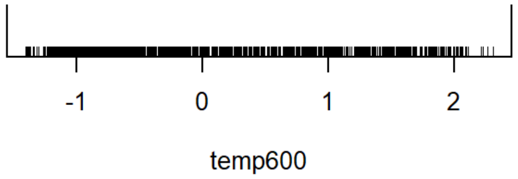


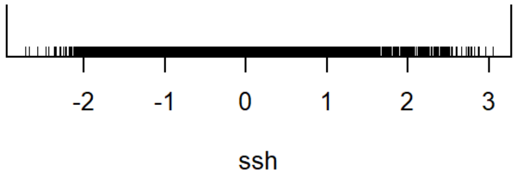
Training Dataset, No Sightings



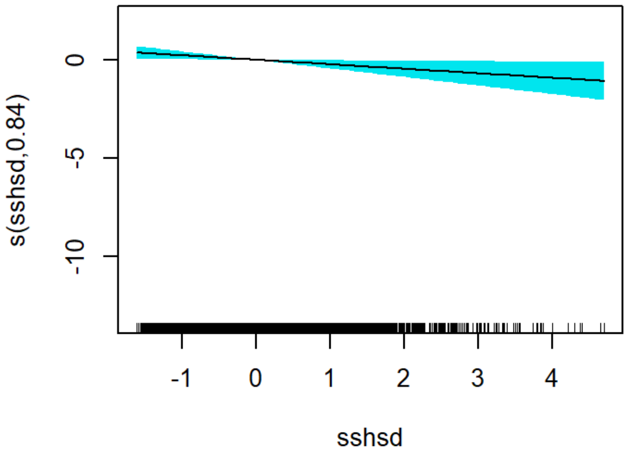
Training Dataset, No Sightings



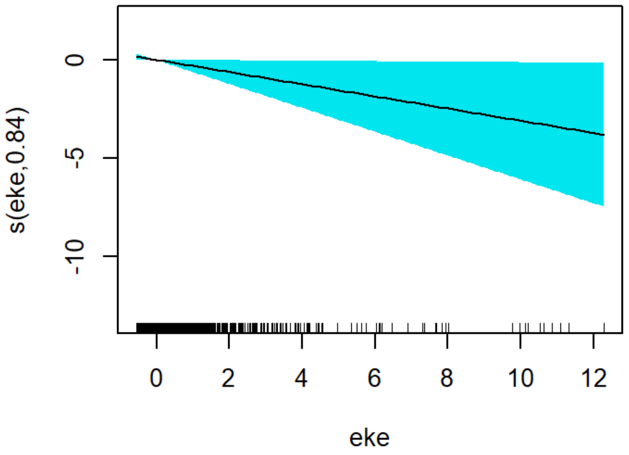




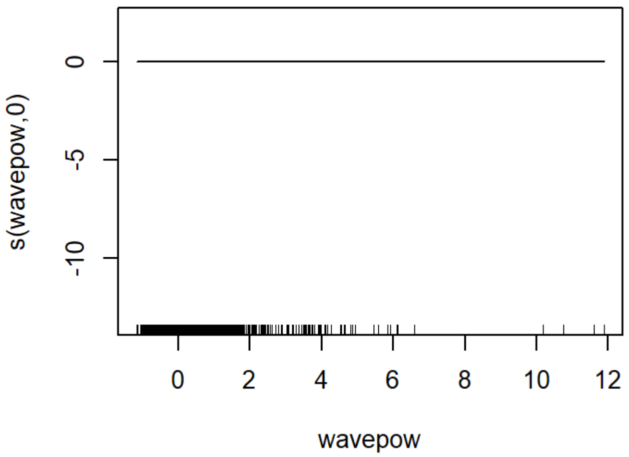
Training Dataset, No Sightings

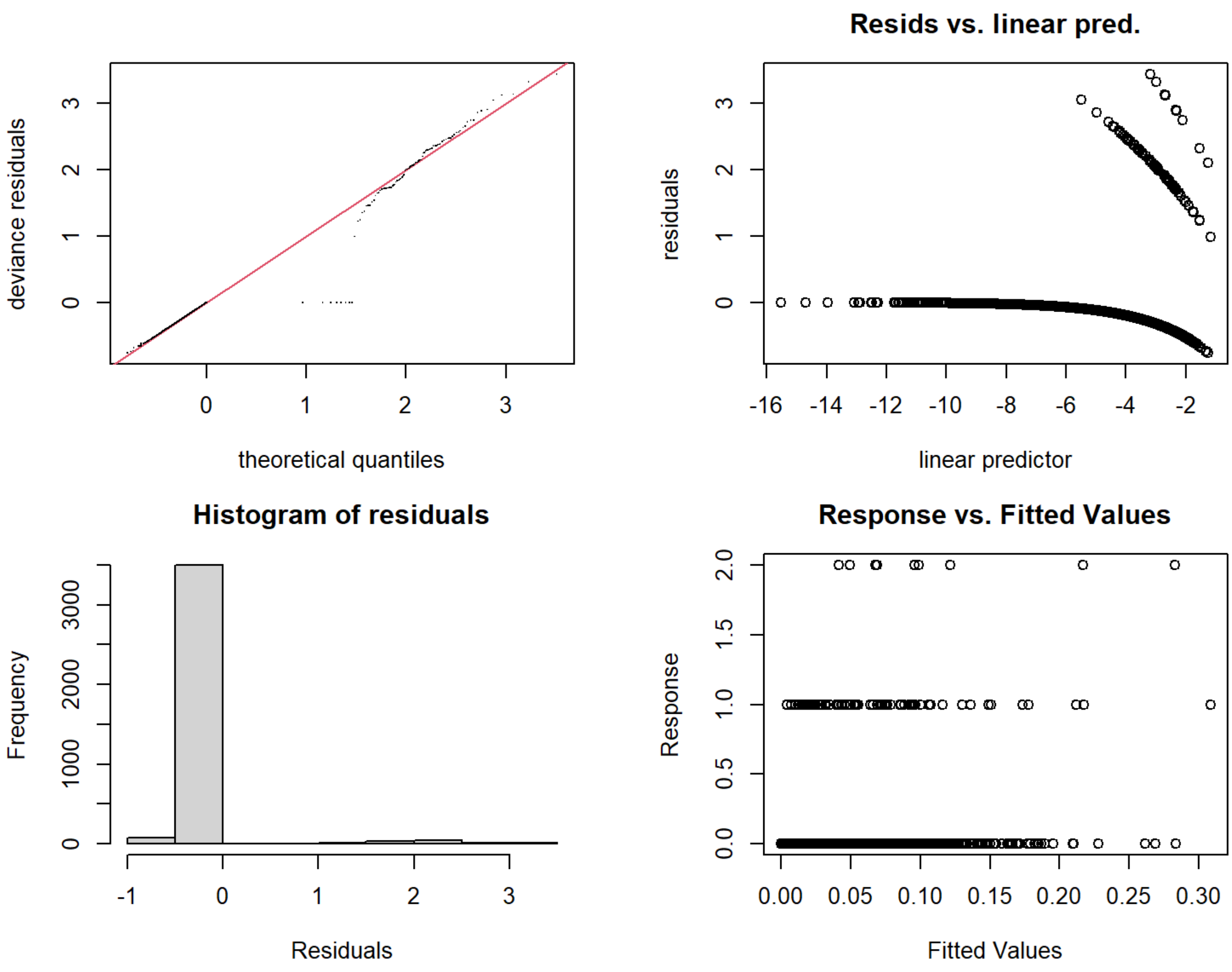


Training Dataset, No Sightings



Training Dataset, No Sightings




Resids vs. linear pred.

Histogram of residuals

Response vs. Fitted Values

```

Method: REML   Optimizer: outer newton
full convergence after 23 iterations.
Gradient range [-0.0007049199,8.188643e-05]
(score 295.3063 & scale 1.040373).
eigenvalue range [-1.443749e-05,10199.48].
Model rank =  23 / 23 

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                 k'      edf k-index p-value    
s(bath_m)  2.00e+00 7.84e-01    0.89   0.750    
s(dist)    2.00e+00 1.15e-04    0.88   0.355    
s(slope)   2.00e+00 5.86e-05    0.85   0.035 *  
s(d2smt)   2.00e+00 1.59e-04    0.85   0.015 *  
s(sst)     2.00e+00 6.47e-01    0.86   0.075 .  
s(chla)    2.00e+00 8.42e-01    0.87   0.150    
s(temp600) 2.00e+00 1.76e+00    0.75  <2e-16 ***
s(ssh)     2.00e+00 8.99e-01    0.87   0.260    
s(sshsd)   2.00e+00 8.38e-01    0.84  <2e-16 ***
s(eke)     2.00e+00 8.36e-01    0.87   0.205    
s(wavepow) 2.00e+00 7.46e-05    0.81  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*With full dataset and **unfiltered chla**

```
twFullb <- gam(pa ~ s(bath_m, k=3) + s(dist, k=3) + s(slope, k=3) + s(d2smt, k=3) +  s(sst, k=3)
+ s(chla, k=3) + s(temp600, k=3) + s(ssh, k=3) + s(sshsd, k=3) + s(eke, k=3) + s(wavepow, k=3) +
offset(log.effort), data = PmScaled, family = tw, link = 'log', select = TRUE, method = "REML")
summary(twFullb)
#```

#```{r echo=FALSE}
par(mar=c(4,4,3,3),mfrow = c(2,2))
plot(twFullb, pages = 3, residuals = FALSE, pch = 20, cex = 0.25,
scheme = 1, shade = T, shade.col = 'hotpink', all.terms = TRUE, main='Full Dataset, All Chla')

# model diagnostics
gam.check(twFullb)
```

*With full dataset and **filtered chla**

```
PmScaled2 <- subset(PmScaled, chla < 10)

twFullc <- gam(pa ~ s(bath_m, k=3) + s(dist, k=3) + s(slope, k=3) + s(d2smt, k=3) +  s(sst, k=3)
+ s(chla, k=3) + s(temp600, k=3) + s(ssh, k=3) + s(sshsd, k=3) + s(eke, k=3) + s(wavepow, k=3) +
offset(log.effort), data = PmScaled2, family = tw, link = 'log', select = TRUE, method = "REML")
summary(twFullc)
# ```
# ```{r echo=FALSE}
par(mar=c(4,4,3,3),mfrow = c(2,2))
plot(twFullc, pages = 3, residuals = FALSE, pch = 20, cex = 0.25,
scheme = 1, shade = T, shade.col = 'turquoise2', all.terms = TRUE, main='Full Dataset, Filtered
 Chla')

# model diagnostics
gam.check(twFullc)
```

Models including only static variables

What are the effects of the static/geographic variables on sperm whale occurrence?

Models only including dynamic variables to evaluate how well they explain